Fewer Humans, Not Zero: LLMs and Measurement Error

Sergio Blanco Piñeiro* Francesco Capozza† Juan Pascual Torres‡

October 2026

Abstract

LLMs make it cheap to measure a variable twice, which is what Obviously Related Instrumental Variables (ORIV) needs to correct measurement error. We show LLMs cannot provide the second measure but can reduce how many humans must provide it. LLM errors are shared across models and load on outcomes, and no number of LLM measures identifies the effect. LLM twins overstate effects by 25 to 160 percent, and an LLM label doubles the effect of toxicity on moderation. Retesting part of a random human sample and using LLM measures as a correction cuts the human answers needed by 16 to 34 percent.

JEL codes: C26, C36, C81, C90, D91
Keywords: measurement error, ORIV, instrumental variables, large language models, digital twins, survey methods

*Institute for Economic Analysis (IAE-CSIC), Universitat de Barcelona, and Institut d'Economia de Barcelona (IEB).

†Universitat de Barcelona, Institut d'Economia de Barcelona (IEB), CESifo, and IZA.

‡Universitat de Barcelona and Institut d'Economia de Barcelona (IEB). Code and data for every result are in the replication package that accompanies this paper.

1 Introduction

Economists increasingly use large language models (LLMs) to create variables. Models classify news headlines, score the tasks in an occupation, and answer survey questions in place of people.[1] Every such variable is measured with error, and error in a regressor biases its coefficient. Gillen, Snowberg, and Yariv (2019) show how to remove this bias. Elicit the variable twice and run ORIV, in which each measure instruments the other in a stacked regression.[2] The cost of the method is the second elicitation, and LLMs seem to make it free. A researcher can ask a second model, rephrase the prompt, or sample the same model again.

This paper shows that the shortcut fails but that LLMs remain useful in a different role. They cannot provide the second measure, but they can reduce how many humans are needed to provide it. The failure is large. In Twin-2K-500 (Toubia et al., 2025), ORIV with two LLM "digital twins" as the duplicate measures overstates the effect of a trait on a human outcome by 25 to 160 percent. In 28,188 online comments rated by crowds, an LLM toxicity label doubles the estimated effect of toxicity on moderation, from 0.14 to 0.29.

The failure turns on one assumption. ORIV requires the errors in the two measures to be independent of each other and of the outcome. Human duplicates satisfy it reasonably well, because the noise in a person's answer today is unrelated to the noise in the same answer two weeks later. LLM duplicates violate it in two ways. Their errors are shared, because models are trained on overlapping text, aligned with similar methods, and given the same prompt.[3] Their errors also load on the outcome, because a model reads the same cues that drive the outcome. We derive the bias that each violation causes and then prove a stronger result (Section 2). Once LLM errors may load on the outcome, no number of LLM measures identifies the effect (Proposition 5). More models, reruns, or prompts cannot fix this. Only a measure that does not see the outcome can.

[1] Examples include the sentiment of financial news headlines (Ludwig, Mullainathan, and Rambachan, 2025), the cognitive content of occupational tasks (Espinal Maya, 2026), and simulated survey respondents (Horton, 2023; Argyle et al., 2023). Dell and Rambachan (2026) review these uses.

[2] The instrument is "obviously related" to the regressor because both measure the same variable. Stacking uses both IV estimates at once and removes the arbitrary choice of which measure to instrument.

[3] The prompt is itself a source of common error. Two models from different developers that receive the same wording share any distortion the wording introduces.

We test the theory in three settings. The first is Twin-2K-500, which surveys 2,058 people, asks the same questions again two weeks later, and releases LLM twins that answer as each person. It therefore provides both a human retest and LLM duplicates of the same answers.[4] Twins from different developers share 77 to 88 percent of their errors on purchase decisions, as much as twins from the same developer. Their personas also contain the outcomes, so their errors load on the outcomes, and ORIV with two twins overstates the effects. The second setting is the comments data, where the LLM sees only the text. Its label still loads on the outcome, because it picks up what moderators reject beyond what raters call toxic. In the third setting we build new twins with Claude Haiku 5.5 from personas that withhold risk attitudes and altruism, so the model never sees the trait. These twins barely measure it, with validities of 0.03 and 0.09, yet in one arm they reproduce the human effect of altruism on empathy. Matching a human estimate is therefore no evidence that an LLM measures the trait.

The positive result is a design that keeps humans as the source of identification and uses LLMs to lower its cost (Section 5). It rests on three results. First, one human answer per person is not enough. Correcting a single answer with LLM measures, as in prediction-powered inference (Angelopoulos et al., 2023), makes the estimate more precise but leaves it attenuated. Its confidence interval therefore misses the truth more often as the sample grows (Proposition 7). Second, only part of the sample needs to answer twice. The first answer already predicts most of what the second adds, so the cheapest design asks a random sample once and a random third to half of it again (Proposition 8). Third, LLM measures should enter the estimator only as a correction with mean zero. In this role their bias cannot affect the estimate, and they reduce the number of humans needed by the share of the sampling variance they explain (Proposition 6). Two further results complete the design. We derive the lowest variance that any estimator can reach with these data (Proposition 9), and we show whom to ask again. At a given cost, retests should go to the people whose outcome and first answer disagree most (Proposition 10).

[4] Toubia et al. (2025) evaluate twins by item-level accuracy. They report an average of 71.7 percent for twins against 81.7 percent for the human retest. Section 4 and Appendix C use person-level scores as well, because a regression uses the score and not the item.

The savings are substantial and do not require good LLMs. In Twin-2K-500, estimating the effect of policy liberalism with a standard error of 0.03 takes 594 human answers when everyone is asked twice. Asking 340 people once and 113 of them again takes 453. Correcting with the persona data brings this to 394, a third fewer. The twins add nothing here, because a regression on the persona extracts more from the same data than the models do. In the comments data nothing but the text is observed for every comment, and the LLM label becomes the useful correction. It explains 18 percent of the sampling variance, against almost nothing for a text classifier trained on 2,000 rated comments, and raises the saving in human ratings from 16 to 24 percent for $54 of model calls. Targeting the retests cuts the variance of the moderation effect by a further 37 percent at the same budget. When the twins carry no information, as for risk and altruism, the design loses nothing.

The practical message is simple. Do not use an LLM measure as a duplicate, whether of another LLM measure or of a human answer. Ask a random human sample once and part of it twice. Use LLM measures, and anything else observed for everyone, to correct the human estimate, and measure in a pilot how much they help before paying for them.

The paper relates to three literatures. The first studies measurement error with repeated measures, from Spearman (1904) to the ORIV estimator and its applications to economic preferences and polygenic scores (Bokern et al., 2023; van Kippersluis et al., 2023).[5] The second studies inference with variables generated by machine learning and LLMs (Ludwig, Mullainathan, and Rambachan, 2025; Duan and Pelger, 2026; Egami and Shin, 2026; Canen and Enamorado, 2026; Dell and Rambachan, 2026). Most of these papers treat human labels as the truth, and Egami and Shin (2026) combine several imperfect LLM labels without any human labels. We instead treat human answers as measured with error and keep them as the source of identification. Human data must then correct both the LLM bias and the human error, and we apply the logic of prediction-powered inference (Angelopoulos et al., 2023) and two-phase sampling to an ORIV estimand.

[5] Spearman (1904) proposed dividing a correlation by the square root of the product of the two reliabilities. Fuller (1987) treats the general errors-in-variables problem. Chapman et al. (2023) apply corrections of this kind to a large set of economic preferences.

Closest to us, Espinal Maya (2026) applies ORIV to scores from two LLMs under the assumption that their errors are independent. We test that assumption and find that it fails. The third literature uses LLMs as simulated subjects (Horton, 2023; Argyle et al., 2023; Bisbee et al., 2024; Park et al., 2024), and we bring the measurement design of Gillen, Snowberg, and Yariv (2019) to that setting.[6]

Section 2 sets out the framework and reports simulations. Sections 3 and 4 present the Twin-2K-500 data and the evidence on LLM duplicates. Section 5 develops the design that combines human and LLM measures, Section 6 applies it to crowd ratings of online comments, and Section 7 concludes.

2 Framework

2.1 Measurement Error and ORIV

A researcher wants the coefficient β in a regression of an outcome y on a latent variable x^*,

$$y_i = \alpha + \beta x_i^* + \varepsilon_i. \tag{1}$$

The variable x^* is never observed. Instead the researcher has measures $x_{ik} = x_i^* + e_{ik}$ for $k = 1, \ldots, K$. These could be two elicitations from the same person, scores from two models, or one of each.

Assumption A *Each error has mean zero and is uncorrelated with x^*, with ε, and with every other error.*

Under Assumption A, OLS on a single measure converges to $\beta\lambda_a$, where $\lambda_a = \sigma^2_{x^*}/(\sigma^2_{x^*} + \sigma^2_{e_a})$

[6] Overgoor, Knight, and Bart (2025) show that correcting LLM labels toward a single human label corrects toward an attenuated target, and they use a second human rater as an instrument. Our retest plays the same role, and the two-phase design shows that the second rater is needed only for part of the sample.

is the reliability of measure a.[7] Instrumenting x_a with x_b gives

$$\operatorname{plim}\hat{\beta}_{IV} = \frac{\operatorname{Cov}(y, x_b)}{\operatorname{Cov}(x_a, x_b)} = \beta \, \frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \operatorname{Cov}(e_a, e_b)}, \tag{2}$$

which equals β when the two errors are uncorrelated. ORIV stacks the data so that each measure is instrumented by the other and estimates one coefficient with standard errors clustered by person (Gillen, Snowberg, and Yariv, 2019).[8] With $K > 2$ measures, each measure is instrumented by all the others, and the model is overidentified.[9]

Equation (2) carries the main point of this section. The correction depends on one number, the covariance between the two errors. It does not depend on how accurate either measure is. An accurate model whose errors are shared with another model is a worse instrument than a noisy model whose errors are its own.

2.2 Errors Shared across Models

We decompose the error of model m on replicate r into a part common to all models, a part specific to the model, and sampling noise, $e_{imr} = c_i + d_{im} + u_{imr}$.

Proposition 1 *Let x_a and x_b come from different models. Then ORIV converges to $\beta\sigma^2_{x^*}/(\sigma^2_{x^*} + \sigma^2_c)$. If x_a and x_b are two replicates of the same model, σ^2_c is replaced by $\sigma^2_c + \sigma^2_d$.*

ORIV therefore always lies between OLS and β. It removes idiosyncratic noise and treats the shared error as signal, so it estimates the effect of the models' consensus $x^* + c$ and not the effect of x^*. Reruns of one model are the weakest duplicates because they also share the model-specific error. The shared variance σ^2_c cannot be recovered from LLM measures alone. It requires one measure from outside the family of models, such as a human retest.

[7] The reliability ratio is the share of the variance of a measure that is signal. A measure with reliability 0.6 attenuates every coefficient on it by 40 percent.

[8] Each person appears twice in the stacked data, once with x_a instrumented by x_b and once with x_b instrumented by x_a. Each copy has its own intercept. Stata and R code is available at github.com/geighei/ORIV.

[9] The overidentification test checks whether all measures share one latent factor with mutually independent errors. With LLM measures, a rejection is direct evidence of shared error.

2.3 Compressed Scales

LLMs avoid extreme answers and differ in how much they do so. We model this as $x_{im} = (1 - \kappa_m)x_i^* + u_{im}$, where $\kappa_m > 0$ means model m compresses the scale. The error is now correlated with x^*, and IV converges to $\beta/(1-\kappa_a)$, the coefficient per unit of the model's own scale.

Proposition 2 *The standardized effect $\beta\sigma_{x^*}$ is identified from three covariances, whatever the values of κ_a and κ_b,*

$$\beta\,\sigma_{x^*} = \sqrt{\frac{\mathrm{Cov}(y,x_a)\,\mathrm{Cov}(y,x_b)}{\mathrm{Cov}(x_a,x_b)}}. \tag{3}$$

The unknown scales enter the numerator and the denominator in the same way and cancel. This is the identification of a one-factor model with congeneric measures (Jöreskog, 1971).[10] We use equation (3) throughout as the standardized ORIV estimator. The one-direction version $\mathrm{Cov}(y,x_b)/\sqrt{\mathrm{Cov}(x_a,x_b)}$ is correct only when the two models compress the scale equally.

2.4 Binary Labels

Let $x^* \in \{0,1\}$ with $P(x^*=1) = \pi$, and let model m label a true 0 as 1 with probability p_m and a true 1 as 0 with probability q_m. A true 1 can only be mislabeled downward, so the error is negatively correlated with x^* and never classical.[11]

Proposition 3 *If mislabeling is independent across models conditional on x^*, then*

$$\beta_{OLS} = \beta\,\frac{(1-p_a-q_a)\,\pi(1-\pi)}{\mathrm{Var}(x_a)} \quad < \quad \beta \quad < \quad \beta_{IV} = \frac{\beta}{1-p_a-q_a}. \tag{4}$$

ORIV overshoots, and OLS and ORIV bound β from below and above. With three or more labels β is point identified.[12]

[10] Measures are congeneric when they load on the same factor with possibly different loadings and error variances. The loadings appear once in the numerator and once in the denominator of equation (3).

[11] Aigner (1973) first showed that misclassification of a binary regressor is never classical. Kane, Rouse, and Staiger (1999) and Black, Berger, and Scott (2000) show that IV with a second misclassified measure overcorrects.

[12] Mahajan (2006) and Hu (2008) give identification results with additional measures. Egami and Shin (2026) apply this logic to labels from several LLMs without gold-standard labels.

2.5 Errors That Load on the Outcome

Suppose the model reads cues that also drive the outcome, so that $e_k = \delta\varepsilon + u_k$ for every model k.

Proposition 4 *When the u_k are independent, ORIV converges to*

$$\frac{\beta\sigma^2_{x^*} + \delta\sigma^2_\varepsilon}{\sigma^2_{x^*} + \delta^2\sigma^2_\varepsilon}. \tag{5}$$

For $\beta = \sigma^2_{x^*} = \sigma^2_\varepsilon = 1$ this is $(1+\delta)/(1+\delta^2)$, which exceeds 1 for $0 < \delta < 1$ and falls below 1 for $\delta < 0$ or $\delta > 1$.

The bias does not shrink as models are added, because every model carries the same δ. It also survives pairing a model with a human measure whenever the estimator uses $\mathrm{Cov}(y, x_k)$. An instrument removes measurement error in a regressor. It cannot remove a correlation between the regressor's error and the outcome that the regressor's own covariance with y carries.

2.6 What LLM Measures Alone Cannot Identify

Propositions 1 to 4 describe biases for given error structures. A stronger question is whether any amount of LLM data, however many models and runs, can recover ρ when the error structure is unknown. Let each LLM measure be $t_k = \kappa_k x^* + \delta_k\varepsilon + c + u_k$ for $k = 1, \ldots, K$, with a common error c, model-specific errors u_k, and loadings δ_k on the outcome shock, all unrestricted.

Proposition 5 *(i) For any K, the means and covariances of $(y, t_1, \ldots, t_K)$ are the same for every value of ρ in $(-1, 1)$, and so is their joint distribution when (x^*, ε) are jointly normal. The identified set for ρ is the whole interval. (ii) If instead $\delta_k = 0$ for every k, $K \geq 3$, and the loadings κ_k are nonzero with ratios κ_k/κ_l that are not all equal, $|\rho|$ is identified even when the error c is shared, and so is ρ once one measure is known to rise with the trait. (iii) Two human measures whose errors are independent of each other, of ε, and of the LLM errors identify ρ whatever the t_k.*

The proof is in Appendix A. Part (i) holds because the outcome and the LLM measures span the same two latent directions, and any direction other than y itself can be relabeled as x^* with the

remainder of y relabeled as ε. The LLM data cannot tell the trait from the outcome shock once the models are allowed to read both. Part (ii) shows that shared error alone is a smaller problem than outcome loading, because a factor model with three or more heterogeneous models separates the trait from the common error. Part (i) concerns second moments, which is all that covariance-based methods such as ORIV use. With non-normal latent variables, higher moments can in principle separate the trait from the outcome shock (Reiersøl, 1950), but only under restrictions on the shapes of distributions that nothing about LLM errors makes plausible. Part (iii) is why the rest of the paper keeps humans in the design. Outcome loading is a property of what the model sees, so it cannot be ruled out from LLM data, and only measures that do not see the outcome restore identification.

2.7 Monte Carlo Evidence

We simulate each case with $\beta = 1$, $\mathrm{Var}(x^*) = 1$, $y = x^* + \varepsilon$ with ε standard normal, $n = 2{,}000$, and 500 replications per design point.[13] Every simulated estimate is within 0.015 of the value the corresponding proposition implies.

Figure 1 fixes the reliability of each LLM measure at 0.6 and varies the share of its error variance that is common to all models. Half of the remaining error is specific to the model and half is sampling noise. A rerun shares the model-specific part. A second model does not. With no shared error, ORIV across two models recovers β. As the shared share rises, ORIV falls toward OLS and reaches it when all error is shared. Reruns of one model never get above 0.75. A single human measure used as the instrument recovers β at every level of sharing.

With a binary truth and a 10 percent misclassification rate in each direction, OLS converges to 0.80 and ORIV to 1.25. At 25 percent, the two bounds are 0.50 and 2.01. A two-class latent class model fitted to three labels recovers β within 0.01 at every rate up to 25 percent.[14] When

[13] The simulation code fixes the random seed and is part of the replication package.

[14] We fit the latent class model by EM (Dempster, Laird, and Rubin, 1977) on the three labels alone and regress y on the posterior probability that $x^* = 1$. The difference between x^* and this probability is uncorrelated with the probability, so it does not attenuate the slope.

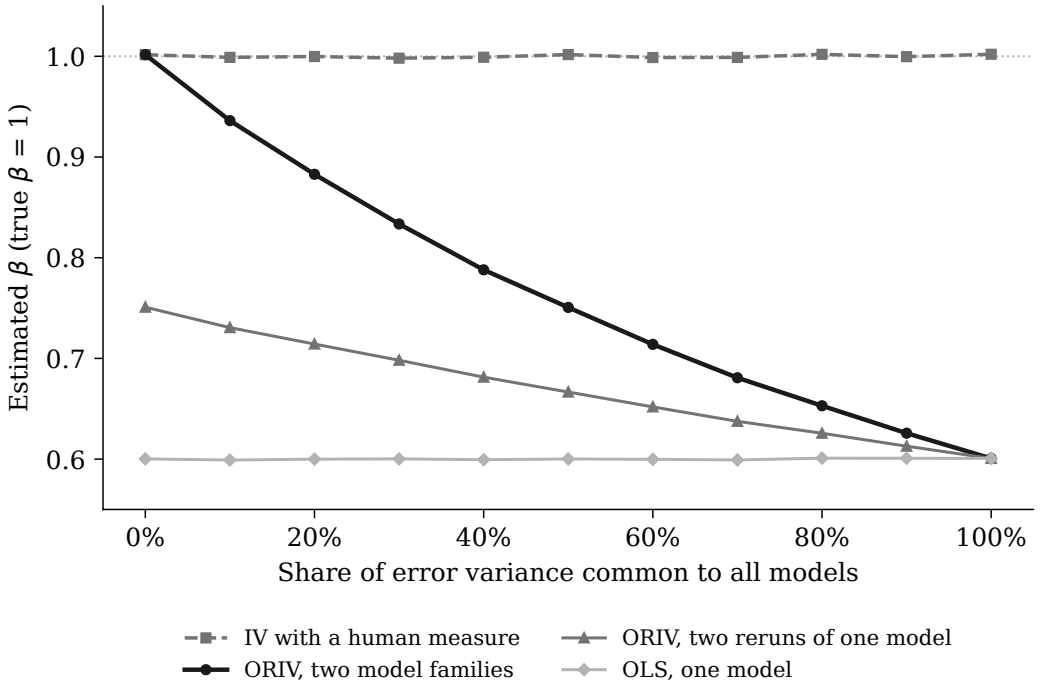


Figure 1. ORIV and Errors Shared across Models

Notes: Monte Carlo with $\beta = 1$, $\mathrm{Var}(x^*) = 1$, reliability of each LLM measure 0.6, $n = 2{,}000$, and 500 replications per point. The dotted line is the true β.

one model compresses the scale by $\kappa = 0.5$ and the other does not, raw-unit ORIV converges to 1.50, the one-direction standardized estimator falls to 0.71, and equation (3) stays between 0.999 and 1.001. When both models load on the outcome shock with $\delta = 0.2$, ORIV converges to 1.15. With $\delta = -0.2$ it converges to 0.77. In both cases IV with a human measure returns 1.00.

Table 1 reports ORIV inference at $n = 500$, a typical size for the human subsample of Section 5. Coverage is close to nominal down to a reliability of 0.15, where the first-stage F statistic is about 12. Below that the instruments are weak.[15]

3 Data

Toubia et al. (2025) surveyed 2,058 US adults in four waves between January and February 2025.

[15] Staiger and Stock (1997) suggest a first-stage F statistic of 10 as a threshold. Our F statistic is cluster-robust and comes from the stacked first stage.

Table 1. ORIV Inference at $n = 500$

Reliability of each measure	Median first-stage F	Median ORIV estimate	Coverage of 95 percent interval
0.90	11,365	1.00	0.942
0.60	446	1.00	0.956
0.30	55	0.99	0.958
0.15	12	0.99	0.934
0.10	5	0.98	0.886
0.05	1	0.74	0.826

Notes: Monte Carlo with $\beta = 1$ and 500 replications. Standard errors are clustered by person in the stacked regression.

Waves 1 to 3 cover demographics, 19 personality scales, cognitive tests, economic preferences, and a set of heuristics-and-biases experiments. Wave 4, two weeks after wave 3, repeats the heuristics-and-biases items and a pricing study under the same conditions. The data are public under a CC BY 4.0 license.

Each twin receives a persona built from one person's waves 1 to 3 answers and predicts that person's wave 4 answers.[16] We use four twins: GPT-4.1-mini with a text persona, GPT-4.1 with the same persona in JSON format, Gemini 2.5 Flash with a text persona, and GPT-4.1-mini with a persona that contains only demographics. For each person we therefore have six measures of every repeated item: the original human answer (h_1), the retest two weeks later (h_2), and four twin predictions of h_2.

We build three person-level scores, the same way from each of the six sources. Purchase propensity is the share of 40 grocery products the person would buy at the stated price.[17] Policy liberalism is mean support on a 1 to 5 scale for ten policies, such as a carbon tax and paid family leave, with the two conservative policies reverse coded. Perceived public support is the mean share of the public that the person believes supports the same ten policies. Two outcomes come only from the human waves 1 to 3 data. The spendthrift index averages the standardized answers to four

[16]The four twins differ in model, in the format of the persona (text or JSON), and in its content (all waves 1 to 3 answers or demographics only). We use the twin predictions released with the data and do not rerun any model.

[17]Each product appears once per wave with a stated price. The score therefore runs from 0 to 1 in steps of 0.025.

Table 2. Means and Standard Deviations by Source

Source	Purchase propensity	Policy liberalism
Human, wave 4	0.44 (0.17)	3.62 (0.85)
GPT-4.1-mini	0.30 (0.10)	3.73 (1.10)
GPT-4.1	0.41 (0.13)	3.85 (0.98)
Gemini 2.5 Flash	0.60 (0.12)	3.60 (1.51)
GPT-4.1-mini, demographics only	0.71 (0.17)	4.51 (0.51)

Notes: Means with standard deviations in parentheses. $N = 2{,}058$.

items of the tightwad and spendthrift scale of Rick, Cryder, and Loewenstein (2008).[18] Political views is self-placement on a five-point scale from very conservative to very liberal.

Table 2 shows that the twins do not answer on the human scale. GPT-4.1-mini compresses purchase propensity to about half the human standard deviation. Gemini stretches policy liberalism to almost twice the human standard deviation. The demographics-only twin places the average person close to "liberal" on every policy. These differences are the κ_m of Section 2, so we report every effect in standardized units using equation (3).

4 Results

All intervals in this section are 95 percent percentile intervals from 1,000 bootstrap draws over people.[19]

4.1 Humans Are Reliable and Twins Are Weakly Valid

The first column of Table 3 reports the correlation between the original answer and the retest. The remaining columns report the validity of each twin, defined as its correlation with the stable human trait,

$$\text{validity}(t) = \frac{\text{Cov}(t, h_2)}{\sqrt{\text{Cov}(h_1, h_2)\,\text{Var}(t)}}. \tag{6}$$

[18]The four items are self-placement from tightwad to spendthrift on an 11-point scale, trouble limiting spending, trouble spending (reversed), and similarity to a shopper who cannot resist a sale (reversed).

[19]Each draw resamples people with replacement and recomputes every statistic, including the human benchmark. The intervals therefore account for the correlation between the estimators being compared.

Table 3. Human Reliability and Twin Validity

	Human	Twin validity			
Score	retest	GPT-4.1-mini	GPT-4.1	Gemini	Demographics
Purchase propensity	0.86	0.22	0.18	0.24	0.10
Policy liberalism	0.92	0.75	0.72	0.73	0.63
Perceived public support	0.63	0.38	0.35	0.27	0.33

Notes: Human retest is the correlation between waves 1 to 3 and wave 4. Twin validity is the correlation of the twin score with the stable human trait. Every 95 percent bootstrap interval is narrower than ± 0.06. $N = 2{,}058$.

The numerator uses the retest, so it does not depend on whether the twin saw the original answer. Twins capture most of a person's politics and little of their shopping. Personas built from earlier answers beat demographics alone for both scores.

4.2 Twins Do Not Copy Earlier Answers

A twin that reproduced the noise in a person's earlier answer would covary more with h_1 than with h_2. We measure this as $1 - \mathrm{Cov}(t, h_2)/\mathrm{Cov}(t, h_1)$. Across the twelve combinations of twin and score, the measure lies between -0.15 and 0.02. For policy liberalism every interval lies within 0.035 of zero. The twins track the stable part of the human answer and not its noise, which is the pattern we would see if the repeated items were withheld from the persona.

4.3 Twins Share Most of Their Errors

Under Assumption A, the covariance between two twins equals

$$\frac{\mathrm{Cov}(t_a, h_2)\,\mathrm{Cov}(t_b, h_2)}{\mathrm{Cov}(h_1, h_2)}. \tag{7}$$

The excess over this benchmark is shared twin error, the σ_c^2 of Proposition 1. Table 4 reports it for all six pairs of twins. The share is the same for twins from one developer (GPT-4.1-mini and GPT-4.1) and from two developers (either one and Gemini). It is also the same when one of the twins saw only demographics. The shared error therefore comes from beliefs the models

Table 4. Shared Error between Twins

Score	Correlation between twins	Share of twin covariance not explained by the human trait
Purchase propensity	0.09 to 0.32	0.77 to 0.88
Policy liberalism	0.77 to 0.94	0.40 to 0.43
Perceived public support	0.65 to 0.88	0.84 to 0.87

Notes: Ranges across the six pairs of twins. The share is $1 - \mathrm{Cov}(t_a, h_2)\,\mathrm{Cov}(t_b, h_2)/[\mathrm{Cov}(h_1, h_2)\,\mathrm{Cov}(t_a, t_b)]$. $N = 2{,}058$.

hold in common about people like the persona, and not from the persona text or from a particular developer.

4.4 Twin Duplicates Overstate Effects

We now use each score as a regressor and ask whether twin duplicates recover the effect that the human retest recovers. Table 5 and Figure 2 report standardized effects. The human benchmark is ORIV on h_1 and h_2. Because humans are reliable, it exceeds OLS on h_2 by only 5 to 10 percent. Single twins miss in both directions, from 0.06 to 0.37 for purchase propensity. ORIV on two twins lands above the benchmark for every pair, at 0.27 to 0.53 against 0.20 for purchase propensity and at 0.90 to 0.92 against 0.73 for policy liberalism.

The mechanism is Proposition 4. Each persona contains the person's answers to the spendthrift items and their political views. The twin therefore infers purchase propensity and policy positions partly from the outcome, and its error loads on the outcome. The demographics-only twin reaches the same place by another route, since demographics predict both its answer and the outcome. Pairing a twin with the human retest leaves the bias in place, because the estimator still uses the twin's covariance with y.

4.5 Reruns, Fine-Tuning, and Published Benchmarks

Two further sets of results are in the appendix. Appendix B uses three more twin versions released with Twin-2K-500. Two runs of the same model share 95 percent of their purchase error, more

Table 5. Standardized Effects with Human and Twin Duplicates

Estimator	(1) Purchase propensity on spendthrift index	(2) Policy liberalism on political views
Panel A. Human measures		
OLS, retest	0.182 [0.143, 0.221]	0.693 [0.669, 0.716]
ORIV, original and retest	0.201 [0.158, 0.240]	0.725 [0.702, 0.747]
Panel B. OLS on one twin		
GPT-4.1-mini	0.168 [0.127, 0.212]	0.890 [0.881, 0.899]
GPT-4.1	0.136 [0.090, 0.181]	0.868 [0.857, 0.879]
Gemini 2.5 Flash	0.368 [0.326, 0.408]	0.887 [0.879, 0.894]
Demographics only	0.063 [0.020, 0.105]	0.737 [0.718, 0.754]
Panel C. ORIV on two twins		
GPT-4.1-mini and GPT-4.1	0.269 [0.204, 0.334]	0.911 [0.903, 0.920]
GPT-4.1 and Gemini	0.434 [0.356, 0.502]	0.915 [0.908, 0.922]
GPT-4.1-mini and Gemini	0.525 [0.449, 0.597]	0.914 [0.908, 0.921]
GPT-4.1-mini and demographics only	0.309 [0.168, 0.433]	0.908 [0.898, 0.919]
GPT-4.1 and demographics only	0.304 [0.159, 0.444]	0.912 [0.902, 0.923]
Gemini and demographics only	0.439 [0.264, 0.557]	0.904 [0.895, 0.913]
Panel D. ORIV on one twin and the human retest		
GPT-4.1-mini	0.393 [0.321, 0.479]	0.924 [0.912, 0.935]
GPT-4.1	0.391 [0.302, 0.481]	0.932 [0.920, 0.944]
Gemini 2.5 Flash	0.556 [0.485, 0.632]	0.937 [0.926, 0.947]
Demographics only	0.353 [0.195, 0.500]	0.918 [0.898, 0.935]

Notes: Standardized effects (standard deviations of the outcome per standard deviation of the trait). OLS rows report the correlation. ORIV rows use equation (3) divided by the standard deviation of the outcome. Brackets contain 95 percent bootstrap intervals from 1,000 draws over people. $N = 2{,}058$.

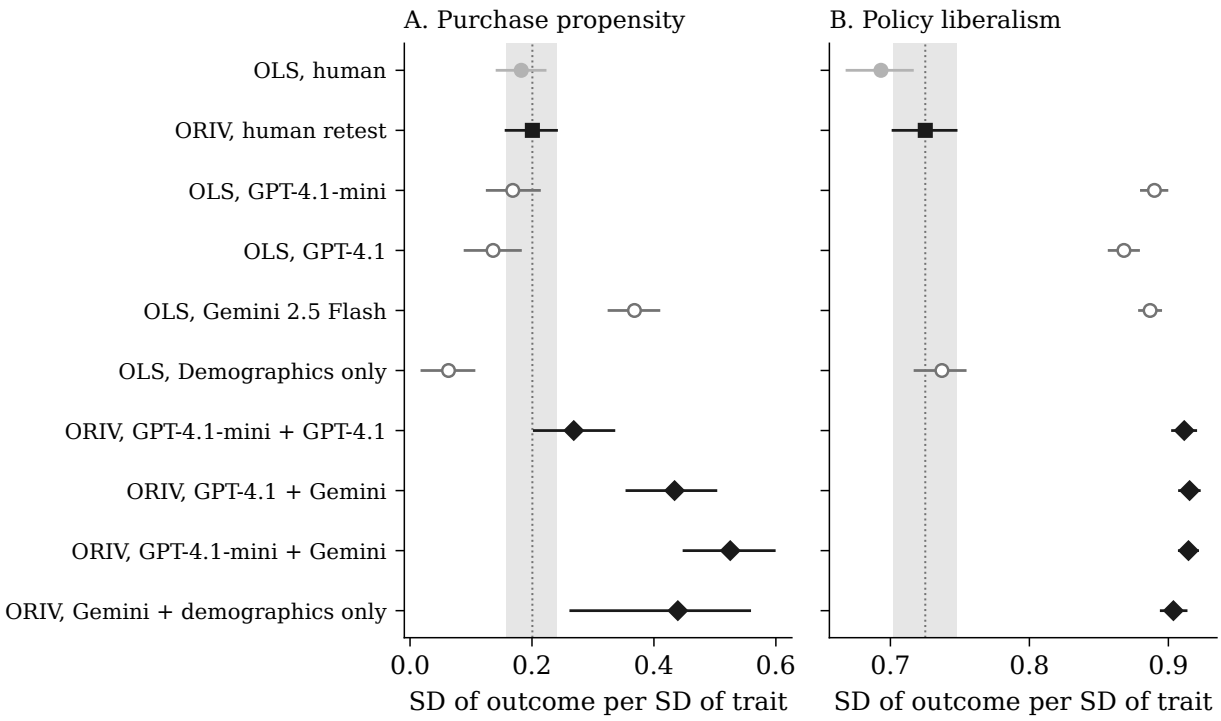


Figure 2. Effects Estimated with Human and Twin Duplicates

Notes: Panel A regresses the spendthrift index on purchase propensity. Panel B regresses political views on policy liberalism. Squares mark the human benchmark (ORIV on the original answer and the retest), whose interval is the shaded band. Hollow circles are OLS on one twin and diamonds are ORIV on two twins. Lines are 95 percent bootstrap intervals.

than any pair of developers, and ORIV on them barely moves OLS, as Proposition 1 predicts for replicates. A twin fine-tuned on 500 people is less over-coherent but also much less valid, so fine-tuning trades shared error for noise. Appendix C compares every statistic with published estimates. Measured by item accuracy, our twins reach 80 to 85 percent of the human retest on purchases, in line with the 86 to 88 percent reported by Toubia et al. (2025) and Park et al. (2024). Measured at the level of the person, which is what a regression uses, they recover only 11 to 25 percent of the retest correlation. Item accuracy rewards predicting the common answer, and a regression needs differences between people.

5 Combining Human and LLM Measures

LLM measures fail as duplicates of each other. This section asks what they can still do, and what a study that corrects measurement error with human answers should look like. The answer has three parts. A single human answer cannot do the job of two, even when it is combined with LLM measures for everyone. Two answers are needed, but only from a fraction of the people in the human sample. And information observed for everyone, including LLM twins, lowers the number of humans needed by the share of the sampling variance it explains, which in our data is small for the twins and larger for the persona data they were built from.

The results of this section are of two kinds. Propositions 1 to 5 and 7, and the consistency of every estimator below, are statements about probability limits that hold for any random sample. The variances in Propositions 6, 8, and 9, the optimal shares in Proposition 8, and the retest rule in Proposition 10 are first-order approximations that hold as the human sample grows. Assumption B in Appendix A.12 states the conditions for these approximations when the corrections are estimated: independent draws, retest probabilities known and set before the study sample is drawn, finite fourth moments, $\mathrm{Cov}(h_a, h_b) > 0$ and $\rho \neq 0$, and corrections that are least-squares projections or cross-fitted predictions. Theorem A1 gives the limit distribution under these conditions, and Section 5.5 measures how good the approximation is with 50 to 200 humans.

5.1 The Hybrid Estimator

Suppose the outcome y and the LLM twins $t_1, \ldots, t_K$ are available for all N people in a sample, as they are in a survey or an administrative file. Two human measures h_a and h_b of the regressor, such as an answer and a retest, are collected only for a random subsample of n people. Let $\hat{\rho}_n$ be the standardized ORIV estimate of equation (3) on the subsample, which is what a study with only these humans would report. The hybrid estimator corrects it with the information observed for everyone,

$$\hat{\rho}_H = \hat{\rho}_n - \hat{\Lambda}'(\bar{g}_n - \bar{g}_N), \tag{8}$$

where g_i collects five functions of person i's outcome and twins: $\bar{t}_i$, $\tilde{y}_i$, $\tilde{y}_i\bar{t}_i$, $\bar{t}_i^2$, and $\tilde{y}_i^2$, with $\tilde{y}_i$ the demeaned outcome and $\bar{t}_i$ the average of the standardized twins. The bars denote means over the subsample and over the full sample. The coefficient $\hat{\Lambda}$ comes from a regression, in the subsample, of the influence function of $\hat{\rho}_n$ on g. In practice we correct each of the covariances in equation (3) with g and then combine them, which is the same estimator to first order and remains stable when the effect is close to zero.

Proposition 6 *Let the subsample be a simple random sample and let h_a and h_b satisfy Assumption A. For any joint distribution of the twins with the latent variable, the outcome, and the human errors, $\hat{\rho}_H$ is consistent for the standardized effect ρ and asymptotically normal with variance*

$$V_\psi \left(\frac{1-R^2}{n} + \frac{R^2}{N} \right), \tag{9}$$

where V_ψ is the variance of the influence function of $\hat{\rho}_n$ and R^2 is the share of it explained by g. A study that measures only humans needs n_H people for the same precision, where $n = (1 - R^2)\, n_H / (1 - R^2 n_H / N)$.

The proof is in Appendix A. The twins enter equation (8) only through a term with mean zero, so their shared errors and their loading on the outcome cannot bias the estimate. Humans correct the measurement error. The twins only reduce the number of humans needed. When the full sample is much larger than the subsample, the share of human answers saved is R^2. This R^2 refers to the influence function, which depends on products of the outcome and the human answers, and not to the human answer alone. A twin can predict the human answer well and still explain little of V_ψ. A pilot that collects both human answers from a few hundred people measures R^2 directly.

We test the hybrid estimator on the data of Section 4. The outcome and the four twins are kept for all 2,058 people. For a random subsample of n people we keep both human answers and hide them for everyone else, and we draw 1,000 subsamples for each n. The benchmark is the human estimate on all 2,058 people. Both estimators are centered on the benchmark at every subsample size. The largest gap between the average hybrid estimate and the benchmark is 0.007,

against a bias of 0.07 to 0.32 when ORIV uses two twins as duplicates of each other. Controlling for demographics does not rescue the twin-only estimates. With 13 demographic variables as controls, ORIV on two twins puts the policy effect at 0.81 to 0.88, against a human benchmark of 0.47 with the same controls. Controls cannot remove outcome loading when the persona contains the outcome itself.

The precision gain is small. For policy liberalism, where the average twin correlates with the human answer at 0.74, the twins explain 11 percent of the sampling variance of the human estimate and save 9 to 10 percent of human answers. For purchase propensity, where that correlation is 0.28, they save 1 to 2 percent. At about £2 per human answer on Prolific and $0.04 per person for four model calls, twins for all 2,058 people roughly pay for themselves for policy liberalism, where they lower the cost of a study with a standard error of 0.03 by 4 percent, and cost more than they save for purchase propensity. Inference requires a minimum number of humans. With 50 humans the 95 percent interval of the hybrid covers the benchmark 83 percent of the time, with 100 humans 90 to 91 percent of the time, and from 200 humans on between 93 and 100 percent of the time. Section 5.5 gives intervals that keep their level with fewer humans.

5.2 One Human Answer Is Not Enough

A tempting shortcut is to collect one human answer from more people and correct it with LLM measures for everyone. This is prediction-powered inference with the human answer as the label (Angelopoulos et al., 2023). A second shortcut uses the human answer and an LLM measure as the two duplicates of ORIV.

Proposition 7 *Let h_a satisfy Assumption A with reliability λ_a. (i) The correlation of y with h_a, with or without a correction of the form in equation (8), converges to $\rho\sqrt{\lambda_a}$. (ii) Equation (3) with h_a and an LLM measure t as the two duplicates converges to ρ only if $\delta\sigma_\varepsilon^2 = \beta\,\mathrm{Cov}(u, e_a)$, where $\delta\varepsilon$ is the part of the error in t that loads on the outcome and u the rest. When the errors in t and h_a are uncorrelated, this requires $\delta = 0$.*

The correction removes noise and leaves the limit unchanged. Part (i) is the point that Overgoor, Knight, and Bart (2025) make for LLM labels of television advertisements, where human labels are themselves a noisy benchmark. More LLM measures and a larger full sample make the attenuated estimate more precise, so its confidence interval covers the truth less often as the study grows. Part (ii) follows from Proposition 4. A twin that loads on the outcome biases the mixed estimator even when its partner is a human.

Table 6 compares the estimators at a fixed budget of human answers. Panel A simulates a trait with human reliability 0.6 and a true standardized effect of 0.5, with three LLM measures that share an error and load on the outcome (Appendix Table A4). With 400 answers, one answer per person converges to 0.39, which is $0.5\sqrt{0.6}$. Prediction-powered inference makes the estimate more precise and lowers coverage from 24 to 16 percent. The mixed estimator converges to 0.64. Every design with two human answers is centered on 0.50 with coverage between 92 and 94 percent. Panel B repeats the exercise in Twin-2K-500 with 800 human answers. Human reliability there is 0.86 to 0.92, so one answer is attenuated by only 0.01 to 0.03. For purchase propensity this bias is smaller than the sampling error, and one answer has the lowest error of all designs. For policy liberalism the bias dominates. The interval of the corrected one-answer estimate covers the benchmark 77 percent of the time with 800 answers and 29 percent of the time with 2,000 (Figure 3). The mixed estimator puts the two effects at 0.46 and 0.93, against 0.20 and 0.73.

5.3 Who Needs to Answer Twice

The second human answer is needed to measure how noisy the first one is. That does not require a second answer from everyone. Let a random subsample S_1 of n_1 people give one answer and a random subsample S_2 of n_2 of them give a second. Moments observed on S_1 are corrected with information observed for everyone, as in equation (8). Moments observed only on S_2 are corrected with information observed on S_1, which includes the first answer. This is the regression estimator of two-phase sampling applied to the five moments of equation (3).

Proposition 8 *Let $S_2 \subseteq S_1$ be nested random samples and let h_a and h_b satisfy Assumption A. For*

Table 6. One and Two Human Answers at a Fixed Budget

	A. Simulation			B. Twin-2K-500					
	$\rho = 0.50$			Purchase (0.20)			Policy (0.73)		
Design	Mean	RMSE	Cover	Mean	RMSE	Cover	Mean	RMSE	Cover
One answer, humans only	0.39	0.120	0.24	0.19	0.027	0.97	0.70	0.028	0.82
One answer, corrected with LLM (and persona) data	0.39	0.118	0.16	0.19	0.025	0.98	0.70	0.027	0.77
One answer and one LLM measure as duplicates	0.64	0.148	0.20	0.46	0.267	0.00	0.93	0.206	0.00
Two LLM measures, no humans	0.60	0.097	0.00	0.38	0.179	..	0.91	0.186	..
Two answers, humans only	0.50	0.066	0.94	0.20	0.044	0.96	0.73	0.024	0.96
Two answers, corrected with LLM twins	0.50	0.060	0.93	0.20	0.043	0.95	0.73	0.022	0.95
Some answer twice, corrected with LLM (and persona) data	0.50	0.061	0.92	0.20	0.032	0.98	0.73	0.018	0.98

Notes: Mean, root mean squared error, and coverage of the 95 percent interval of the standardized effect. Panel A: 1,000 simulated populations of 2,000 people, human reliability 0.6, a budget of 400 human answers. Panel B: 300 random subsamples of the 2,058 Twin-2K-500 participants, a budget of 800 human answers, with the full-sample human estimate (in parentheses) as the benchmark. One answer uses 800 people (400 in Panel A), two answers use 400 (200), and the two-phase design of Proposition 8 uses 586 people of whom 214 answer twice for policy liberalism (619 and 181 for purchase propensity, 230 and 170 in Panel A). In Panel B the correction uses the twins and a cross-fitted prediction from persona data, in Panel A the three LLM measures. Two LLM measures average over the six pairs of twins. The twin-only estimate has no sampling variation from humans, so coverage is not reported.

any auxiliary information observed for everyone, the two-phase estimator is consistent for ρ *with variance*

$$V_\psi \left(\frac{1 - R_1^2}{n_2} + \frac{R_1^2 - R_0^2}{n_1} + \frac{R_0^2}{N} \right), \tag{10}$$

where R_0^2 *is the share of* V_ψ *explained by information observed for everyone and* R_1^2 *the share explained once the first answer is also known. If a first answer costs* c_1 *and a second costs* c_2*, the cheapest design for a given precision sets*

$$\frac{n_2}{n_1} = \sqrt{\frac{(1 - R_1^2)\, c_1}{(R_1^2 - R_0^2)\, c_2}} \tag{11}$$

whenever this ratio is below one, and has everyone answer twice otherwise.

The proof is in Appendix A. With $n_1 = n_2$ the variance reduces to that of Proposition 6. The first answer predicts most of what the second adds, because both measure the same trait with similar

Table 7. Human Answers Needed for a Standard Error of 0.03

	Purchase propensity				Policy liberalism			
	Once	Twice	Saved	Cost	Once	Twice	Saved	Cost
Humans only, everyone twice	0	1,040	..	5,551	0	297	..	1,588
Corrected with LLM twins, everyone twice	0	1,034	1%	5,605	0	278	7%	1,567
Corrected with twins and persona data, everyone twice	0	1,001	4%	5,427	0	242	19%	1,376
Some twice, humans only	878	351	24%	4,219	227	113	24%	1,210
Some twice, corrected with persona data	816	348	27%	4,038	182	106	34%	1,052
Some twice, corrected with twins and persona data	844	347	26%	4,191	186	106	33%	1,145

Notes: People who answer once and twice, the share of human answers saved relative to a human-only study, and the total cost in dollars, from Propositions 6 and 8 with V_ψ estimated on all 2,058 people. R_0^2 is measured out of sample as one minus the ratio of the sampling variances of the corrected and human-only estimators across random subsamples of 300 people, with any prediction trained on those 300 people only. R_1^2 is estimated on the full sample. Costs assume \$2.67 per human answer (about £2 on Prolific at 1.33 dollars per pound in October 2026) and \$0.04 per person for four model calls for all 2,058 people. Persona data are the 13 demographic variables and the first 30 principal components of the waves 1 to 3 answers, excluding the target items and the outcomes.

error. In Twin-2K-500, R_1^2 is 0.91 for policy liberalism and 0.93 for purchase propensity, and R_0^2 is between 0.01 and 0.22 depending on the information used. Equation (11) then asks a second answer from about one person in three. When the retest costs twice as much as the first answer, because it requires recontacting people, the share falls to one in four or five.

Table 7 reports the human answers and the cost needed for a standard error of 0.03. A human-only study of the policy effect needs 297 people measured twice, 594 answers at a cost of \$1,588. Asking 340 people once and 113 of them again reaches the same precision with 453 answers, 24 percent fewer, without any LLM. Adding a prediction from persona data raises the saving to 34 percent and lowers the cost to \$1,052. For purchase propensity the two-phase design saves 24 to 27 percent. Adding the LLM twins to the persona data does not raise these savings, and the twins cost \$82 for the full sample. Figure 3 confirms the ranking at every budget. At equal cost, the two-phase design has the lowest error among the designs that are centered on the benchmark, and its realized error is at or below the value implied by equation (10).

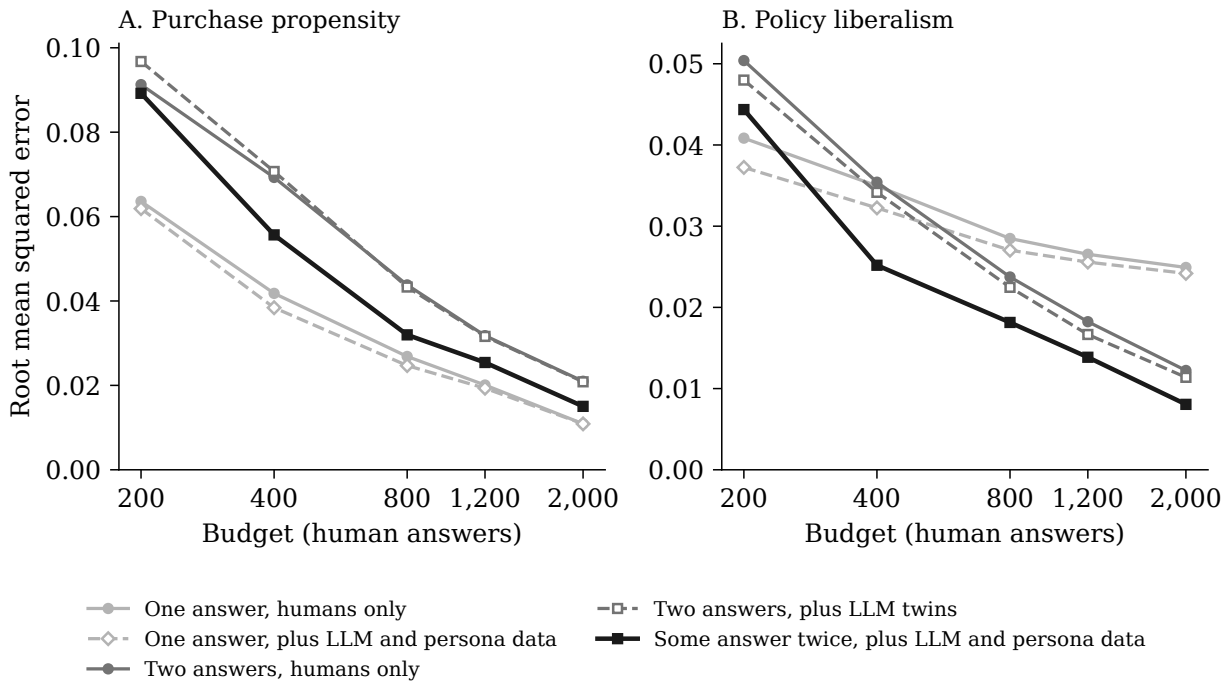


Figure 3. Error of the Effect Estimate by Budget

Notes: Root mean squared error of the standardized effect against the full-sample human benchmark across 300 random subsamples of Twin-2K-500 participants, for a budget measured in human answers. One-answer designs ask each person once. Two-answer designs ask each person twice. "Some answer twice" splits the budget as in equation (11). Panel A regresses the spendthrift index on purchase propensity. Panel B regresses political views on policy liberalism.

5.4 Efficient Estimation and Targeted Retests

The estimator of Proposition 8 corrects each moment with linear functions of the information available at each stage. Two questions remain. Is there a better estimator for the same design, and should everyone in the first sample have the same chance of a retest? The design makes both questions answerable, because the researcher chooses who is asked again and therefore knows each person's probability of being asked. Let W_0 denote the information observed for everyone, W_1 add the first human answer, and $\pi(W_1)$ be the probability that a person in the first sample is asked again.

Proposition 9 *Let ψ be the influence function of the standardized ORIV estimator with both answers for everyone, $h_0 = E[\psi \mid W_0]$, $h_1 = E[\psi \mid W_1]$, and $\sigma^2(W_1) = \mathrm{Var}(\psi \mid W_1)$. No regular*

estimator has asymptotic variance below

$$\frac{\mathrm{Var}(h_0)}{N} + \frac{E[(h_1 - h_0)^2]}{n_1} + \frac{1}{n_1} E\left[\frac{\sigma^2(W_1)}{\pi(W_1)}\right]. \tag{12}$$

The two-phase estimator with inverse-probability weights attains the bound when its corrections equal h_0 and h_1, and remains consistent and asymptotically normal for any other corrections.

The proof is in Appendix A, and Theorem A1 gives the asymptotic distribution when the corrections are estimated by cross-fitting. Prediction-powered inference is the special case with one phase, a single human answer, and the attenuated target of Proposition 7. Equation (10) is the variance of this estimator with linear corrections and a constant π, and it equals the bound when the conditional means are linear in the features. Better predictions of ψ, linear or not, move the variance toward the bound, and their errors can never bias the estimate.

The bound also shows who should be asked again. The last term weights each person's remaining uncertainty by the inverse of their retest probability, so retests are worth most where that uncertainty is largest.

Proposition 10 *Let a first answer cost c_1 and a second answer c_2. For a given expected cost, the variance in equation (12) is smallest when*

$$\pi^*(W_1) = \min\{1,\ \lambda\,\sigma(W_1)\}, \qquad \sigma(W_1) = \left|\frac{\rho}{2}\left(\frac{\tilde{y}}{\mathrm{Cov}(y, h_b)} - \frac{h_a - E[h_a]}{\mathrm{Cov}(h_a, h_b)}\right)\right| \mathrm{sd}(h_b \mid W_1), \tag{13}$$

where $\lambda = \sqrt{c_1/(c_2 A)}$ with $A = E[(h_1 - h_0)^2]$ when no probability reaches one. At a given expected number of retests, the retest term of the variance falls by the factor $(E[\sigma])^2/E[\sigma^2]$ relative to a constant probability.

The proof is in Appendix A. The rule is a version of Neyman allocation (Neyman, 1934). It asks again the people whose outcome and first answer, each scaled by its covariance with the retest, disagree most. These are the people whose second answer moves the estimate. A pilot gives every quantity in equation (13), and the estimator weights each retested person by $1/\pi$, so a poor

pilot costs precision but never introduces bias. The pilot must be a separate sample. Estimating the retest probabilities on the same people whose answers enter the estimate makes the design adaptive, which our results do not cover. We keep a floor of 0.10 on π to bound the weights.

Table 8 tests the rule. Each dataset is split at random into a pilot half, which sets the design, and a study half, which plays the population. At the same budget of human answers we compare the random design of Proposition 8 with the targeted design of Proposition 10, averaging over ten splits. The gain follows the dispersion of $\sigma(W_1)$, as Proposition 10 predicts. For moderation, where rejected comments are rare and carry most of the information, targeting asks again only 18 percent of the first sample against 53 percent under random retests, and cuts the variance of the estimate by 37 percent. For likes and in the simulation it cuts the variance by 13 percent. In Twin-2K-500 the slopes vary little and only about 100 people are asked again, so the noise from the weights offsets the predicted gain of 8 to 9 percent. Targeting pays when the outcome or the first answer is skewed and the retest sample has several hundred people.

5.5 Inference with Few Humans

The delta-method interval of Proposition 8 undercovers with few humans. With 50 people measured twice, the 95 percent interval covers the benchmark 83 to 91 percent of the time (Table 8, Panel B). The cause is not weak identification, since the human measures in our data have reliabilities of 0.59 to 0.92. It is that the correction regressions are fitted on few people and the plug-in variance is too small. Two changes fix this. The correction for each person is fitted without that person, which keeps it mean zero exactly, and the variance is estimated by the jackknife over the first sample. With these changes coverage is 91 to 97 percent with 50 people and 93 to 98 percent with 100 or more. Computing the interval for the Fisher transform of ρ adds one to two points when the effect is large. When the effect itself may be zero, Theorem A1 shows how to invert tests of a smooth moment condition instead of using a standard error.

Table 8. Targeted Retests and Inference with Few Humans

Panel A. Random and targeted retests at a fixed budget

		Share asked again		Variance, targeted over random	
Data	Budget	Random	Targeted	Realized	Predicted
Simulation	400	0.75	0.52	0.87	0.86
Purchase propensity	400	0.29	0.26	1.08	0.92
Policy liberalism	400	0.36	0.26	0.98	0.91
Comments, moderation	4,000	0.53	0.18	0.63	0.74
Comments, likes	4,000	0.57	0.44	0.87	0.89

Panel B. Coverage of 95 percent intervals, everyone asked twice

	People	Delta	Jackknife	Jackknife, z
Simulation	50	0.90	0.97	0.98
	100	0.91	0.96	0.94
Purchase propensity	50	0.91	0.96	0.96
	100	0.93	0.94	0.95
Policy liberalism	50	0.83	0.91	0.92
	100	0.92	0.93	0.94

Notes: Panel A: each dataset is split ten times into a pilot half and a study half, with 100 draws per split. The budget is in human answers. The random design asks again a constant share of the first sample, chosen by Proposition 8. The targeted design uses equation (13) with parameters from the pilot half and a floor of 0.10. Both use the inverse-probability weighted estimator with the same auxiliary information (the average twin in Twin-2K-500, the LLM label for comments). Realized is the ratio of mean squared errors around the study-half benchmark, predicted is the ratio of the variances in equation (12). Panel B: 400 draws per row. Delta uses the plug-in standard error. Jackknife uses leave-one-out corrections and a jackknife standard error, on the scale of ρ or of its Fisher transform z.

5.6 When the Retest Is Imperfect

Assumption A requires the errors in the answer and the retest to be independent. A person who remembers the first answer, or who reads the question the same way twice, violates it. Let the two errors have covariance $\mu > 0$, and let $\kappa = \mu / \mathrm{Cov}(h_a, h_b)$ be the share of the retest covariance that they account for. The covariances of the two answers with y do not change and $\mathrm{Cov}(h_a, h_b)$ rises by μ, so equation (3) converges to $\rho\sqrt{1-\kappa}$. Because μ cannot exceed the larger of the two error variances, this limit lies between ρ and the one-answer limit $\rho\sqrt{\lambda}$ of the less reliable answer.

Three consequences follow. First, the design does not change the target. Every estimator in this section is a function of the same five population moments, so the two-phase, targeted, and corrected estimators converge to the same $\rho\sqrt{1-\kappa}$ as a study that asks everyone twice. In the

simulation of Table 6, Panel A, with a correlation between the two human errors of 0 to 0.3, both designs are centered within 0.003 of $\rho\sqrt{1-\kappa}$. Second, memory makes the first answer a better predictor of the second. It raises R_1^2, and the optimal design asks fewer people again. In Twin-2K-500, replacing each retest by a mix that puts weight 0.3 on the first answer raises the saving in human answers from 24 to 32 percent and lowers the estimate by 0.004 to 0.008, because the human answers are already reliable. Third, the bias works against our conclusions about twins. Memory makes the human benchmark understate the effect, so the twin estimates could be right only if memory were large. The smallest overstatement in Table 5, 25 percent for policy liberalism, requires $\kappa = 0.36$. The answer would then have a reliability of 0.59 despite a retest correlation of 0.92, and a person's errors two weeks apart would have to be correlated at 0.81. For purchase propensity every pair of twins requires κ of 0.44 or more. A retest asked in different words or in a different format reduces memory without changing the estimator, because equation (3) allows the two measures to have different scales (Proposition 2).

5.7 What Predicts Human Answers

The value of any information observed for everyone is the share of V_ψ it explains. We compare four sources. The outcome alone explains 1 to 5 percent. The average of the four twins explains 5 percent for purchase propensity and 14 percent for policy liberalism when the coefficients are estimated on all 2,058 people. A ridge regression of the human answer on persona data, cross-fitted on the same people, explains 27 and 35 percent, and adding the twins to it changes neither number by more than one percentage point. The twins are built from the same persona data, and a regression extracts more of the relevant information from it than the models do. With the prediction trained on only 300 people, the persona regression still explains 12 and 22 percent, against 1 and 8 percent for the twins.

More models help little. The share explained rises from 3 percent with one twin to 5 percent with four for purchase propensity, and from 10 to 14 percent for policy liberalism. This is Proposition 1 at work. Models that share most of their errors add little independent information, whether

they correct a human estimate or duplicate each other.

The advantage of the twins is that they need no training data. They are the natural source of auxiliary information when the persona is text, when the human subsample is too small to train a regression, or when the target was never asked of anyone. In Twin-2K-500 none of these conditions holds, and a regression on the structured persona data dominates.

5.8 How Many Personas

The last term of equation (10) gives the number of people with personas, and hence with LLM measures, that a study needs. The variance is within 10 percent of its limit with infinitely many personas when

$$N \geq \frac{10R_0^2}{(1-R_1^2)/n_2 + (R_1^2 - R_0^2)/n_1}, \tag{14}$$

which for a design in which everyone answers twice is $N \geq 10R_0^2 n/(1-R_0^2)$. Personas are cheap only when the persona data already exist. If a persona costs c_N, the cost-minimizing number is $N = n_2\sqrt{R_0^2 c_2/((1-R_1^2)c_N)}$ (Appendix A).

For the two-phase designs of Table 7 with twins and persona data, the 10 percent rule requires 643 personas for policy liberalism, about twice the 292 people who give a first answer. For purchase propensity it requires 780, fewer than the 1,191 people who give a first answer, so personas for the human sample alone are enough. When the persona data exist and only model calls must be paid for, the cost-minimizing number is 1,307 for policy liberalism and 2,831 for purchase propensity. When each persona requires a new survey costing \$1, the optimum falls below the size of the human sample in both cases, and personas beyond the people already in it do not pay. With R_0^2 close to zero, as for the twins alone, personas add nothing and the right number is zero. The number of personas is therefore not a separate design choice. It follows from how much they explain, which a pilot measures.

5.9 Five Rules

The framework and the evidence give five rules for researchers who use LLM measures. First, keep outcome information out of the model's input, because the largest bias in Section 4 comes from models that see the outcome. Second, do not use an LLM measure as a duplicate, of another LLM measure or of a human answer, and do not replace the second human answer by an LLM correction of the first. Third, ask a random human subsample once and part of it twice, choose the share with equation (11) or target the retests with equation (13) when the outcome or the first answer is skewed, and correct the estimate with whatever predicts human answers for everyone, LLM twins included. With fewer than 200 humans, report jackknife intervals. Fourth, report standardized effects using equation (3). Fifth, do not use ORIV on binary labels. Report OLS and ORIV as bounds, or ask for three labels or a continuous score.

5.10 Held-Out Twins for Risk Attitudes and Altruism

The twins of Twin-2K-500 are weakly informative about purchases and saw the outcome for both scores. The natural next step is a trait the models have not seen, with twins built from personas that exclude the outcome. Twin-2K-500 already measures risk attitudes and altruism more than once in waves 1 to 3. Risk attitudes are elicited with three certainty-equivalent lists over 50-50 lotteries, each with 14 rows. Altruism is elicited with a dictator game and with five trustee decisions in a trust game, alongside an ultimatum offer and a trust transfer that mix altruism with strategic motives. Table 9 shows that the three risk lists correlate at 0.62 to 0.70, close to the duplicate risk elicitations of Gillen, Snowberg, and Yariv (2019), and that 95 percent of people switch once in each list. Dictator giving correlates at 0.51 with the share returned as a trustee.

These human duplicates make the study possible without new human data. We generate new twins with Claude Haiku 5.5 from personas that withhold the whole economic preferences block, and the model completes the three risk lists and the allocation games for all 2,058 people. A random half of the personas also contain the outcome items (the empathy scale and the spendthrift items), which identifies the loading of Proposition 4 directly. A second run covers a random 300

Table 9. Human Duplicates for Risk Attitudes and Altruism in Twin-2K-500

Pair of measures	Correlation
Panel A. Risk attitudes (sure choices in a certainty-equivalent list)	
Lottery ($6, $0) and lottery ($8, $2)	0.62
Lottery ($6, $0) and lottery ($10, $0)	0.70
Lottery ($8, $2) and lottery ($10, $0)	0.63
Panel B. Altruism	
Dictator giving and share returned as trustee	0.51
Dictator giving and trust transfer	0.40
Dictator giving and ultimatum offer	0.37

Notes: Waves 1 to 3 of Twin-2K-500, $N = 2{,}058$. Each lottery pays either amount with probability one half. The share of people with a single switch point is 0.95 in each list. Variable definitions are in Appendix E.

people. The human measures are the first risk list and the average of the other two, and dictator giving and the share returned as a trustee. The outcomes are the spendthrift index for risk and the empathy scale for altruism. The run cost about $9.

Table 10 reports the results, and they are stark. Held out from the persona, these preferences are almost invisible to the model. The validity of the twin is 0.03 for risk aversion and 0.09 for altruism, against human reliabilities of 0.73 and 0.51. The twin answers with perfect internal consistency, switching once in every risk list (95 percent of humans do), but it compresses the scale. Its risk scores have a standard deviation of 2.1 against 3.7 for humans, and three quarters of twins give exactly $3 in the dictator game. Two runs correlate at only 0.20, and all of their covariance is error unrelated to the trait. ORIV on two runs is therefore uninformative, with intervals that span most of the admissible range.

Two lessons follow. First, a twin that barely measures a trait can still reproduce its effect. In the arm without the outcome items, OLS on the twin's dictator giving puts the effect of altruism on empathy at 0.14, against a human benchmark of 0.13, although the twin's validity is 0.09. The model infers generosity from personality answers that also predict empathy, which is the loading of Proposition 4 operating through correlated content rather than through the outcome item itself. In the other arm the same estimate is 0.03, and the difference between arms is not significant.

Table 10. Held-Out Twins for Risk Attitudes and Altruism

	Risk aversion	Altruism
Panel A. Measures		
Correlation between the two human measures	0.73	0.51
Validity of the twin	0.03	0.09
	[−0.01, 0.08]	[0.03, 0.15]
Correlation between two runs (300 people)	0.20	0.21
Shared error between two runs	1.00	0.99
Panel B. Standardized effect on the outcome		
Human ORIV, personas without the outcome	−0.067	0.129
OLS on the twin, personas without the outcome	−0.017	0.138
Human ORIV, personas with the outcome	−0.088	0.132
OLS on the twin, personas with the outcome	−0.019	0.026
Outcome loading of the twin	0.02	−0.12
	[−0.11, 0.15]	[−0.26, 0.06]
Panel C. Correction with the twin		
Share of the human-only variance explained	0.00	0.00
Humans measured twice for a standard error of 0.03, without and with the twin	1,403, 1,417	2,029, 2,029

Notes: Twins generated by Claude Haiku 5.5 from waves 1 to 3 personas without the economic preferences block, $N = 2{,}058$, with a second run for 300 people. Human measures: sure choices in the first risk list and the average of the other two lists, dictator giving and the share returned as a trustee. Outcomes: the spendthrift index for risk and the empathy scale for altruism. Validity and shared error are defined as in Tables 3 and 4. Outcome loading is the change in the twin's effect between the two arms minus the change in the human effect. Brackets are 95 percent intervals from 1,000 bootstrap draws.

Agreement with a human estimate is no evidence that a twin measures the trait. Second, the design of this section loses nothing when the twins are useless. They explain none of the sampling variance, the hybrid estimate stays centered on the human benchmark at every subsample size, and the number of humans needed is the same as without them. A cheap pilot with a weak model costs a few dollars and reveals this before any human data are bought. One limitation remains. The human duplicates for these traits come from a single session, so session-specific error enters both measures and the human benchmark slightly understates noise.

6 A Second Application: The Toxicity of Online Comments

Twin-2K-500 is a demanding test for LLMs, because the persona data that built the twins are themselves available to the researcher. Most uses of LLMs in economics are different. A model labels text, and nothing else observed for every document measures the same thing. We therefore repeat the analysis on comments from the Civil Comments platform, released by Jigsaw for a public competition. Each comment was rated toxic or not by four to ten crowd raters, and the platform recorded two outcomes, whether the publication's moderators rejected the comment and how many likes it received.

We use a random sample of 28,682 comments with four to ten raters. The raters of each comment are split at random into two halves, and each human measure is the share of a half that rates the comment toxic. Different people produce the two halves, so their errors are independent given the comment's latent toxicity, which plays the role of the retest in Twin-2K-500. Both outcomes are residualized on publication and calendar-month fixed effects, and the likes outcome is the logarithm of one plus the number of likes, for approved comments only. Claude Sonnet 5.5 reads each comment, and nothing else, and reports the percentage of crowd raters it expects to call the comment toxic. A second run covers a random subset of 3,177 comments. Labels for 494 comments failed because of connection errors unrelated to the comments, which leaves 28,188. The labels cost $54, or about $0.002 per comment.

Table 11 reports the results. Human measures are noisy. The two rater halves correlate at 0.59, so a single half attenuates the effect of toxicity on moderation from 0.142 to 0.105, by 26 percent. The LLM label is less valid than a rater half. Its correlation with latent toxicity is 0.47, against 0.77 for a half. Two runs of the model correlate at 0.99, and 77 to 79 percent of their covariance is error they share, as Proposition 1 predicts for reruns.

The LLM error loads on the outcome. Regressed on the LLM label alone, moderation responds with a standardized effect of 0.288, twice the human benchmark of 0.142, even though the label is less valid than a human half. ORIV with two runs gives 0.298, against 0.189 for the human estimate on the same comments, and the mixed estimator with a rater half and the LLM label gives

Table 11. Toxicity of Online Comments: Human Raters and an LLM

	Moderation	Likes
Panel A. Measures		
Correlation between the two rater halves	0.59	0.58
Validity of the LLM label	0.47	0.44
Shared error between two LLM runs	0.77	0.79
Panel B. Standardized effect of toxicity		
One rater half	0.105	0.053
ORIV, two rater halves (benchmark)	0.142	0.072
LLM label alone	0.288	0.063
One rater half and the LLM label as duplicates	0.291	0.099
ORIV, two LLM runs (subset)	0.298	0.063
ORIV, two rater halves (same subset)	0.189	0.033
Panel C. Share of the human-only variance explained		
Text regression trained on 2,000 comments	0.01	0.00
LLM label	0.18	0.07
Panel D. Human answers for a standard error of 0.03 (once, twice, saved)		
Humans only, everyone twice	0, 2,612	0, 1,574
Corrected with the LLM, everyone twice	0, 2,176, 17%	0, 1,419, 10%
Some twice, humans only	1,501, 1,450, 16%	943, 938, 10%
Some twice, corrected with the LLM	1,251, 1,353, 24%	833, 868, 18%

Notes: 28,188 comments (26,580 approved comments for likes). A human answer is the share of one random half of a comment's raters who rate it toxic. Outcomes are residualized on publication and month fixed effects. Validity is the correlation of the LLM label with latent toxicity, $\mathrm{Cov}(t,h_b)/\sqrt{\mathrm{Cov}(h_a,h_b)\,\mathrm{Var}(t)}$. The LLM is Claude Sonnet 5.5. The two runs overlap on 3,127 comments (2,932 approved). Panel C reports one minus the ratio of sampling variances of the corrected and human-only estimators across 200 random subsamples of 2,000 comments rated twice. Panel D uses Propositions 6 and 8. For likes the correction combines the LLM label with the text regression. Saved is the share of human answers saved relative to the first row.

0.291. The label captures what moderators reject beyond what crowd raters call toxic, such as off-topic or abusive content that raters did not flag, which is the error of Proposition 4. For likes the loading is smaller and the mixed estimator gives 0.099 against 0.072.

Panel C turns to the correction. Here the LLM is worth having. A ridge regression on the words of the comment, trained on 2,000 comments with human ratings, explains essentially none of the sampling variance of the human estimate. The LLM label explains 18 percent for moderation and 7 percent for likes, without any training data. The bias of the label does not matter for this use, because it enters the estimator only through a correction with mean zero, and the corrected estimates are centered on the benchmark at every budget.

Panel D gives the human answers needed. Rater reliability is lower than in Twin-2K-500, so the first answer predicts less of the second (R_1^2 is 0.81 to 0.83), and the cheapest design asks about half of the comments twice rather than a third. With the LLM correction it saves 24 percent of human answers for moderation and 18 percent for likes, against 16 and 10 percent for the same design without the LLM. At equal budgets the ranking is the same as in Figure 3. With 8,000 human answers the root mean squared error of the moderation effect is 0.039 with one answer per comment, 0.022 with two, and 0.018 with the two-phase design corrected with the LLM. With 16,000 answers the interval of the one-answer estimate covers the benchmark in 0 to 2 percent of samples and the two-answer intervals in 98 to 100 percent.

The two applications agree on the role of LLMs. As duplicates they fail in both, because their errors are shared across runs and load on the outcome. As a correction they are harmless in both, and they pay off when nothing else observed for everyone measures the trait. For text, $54 of model calls for 28,000 comments save another 8 percentage points of human answers on top of what the two-phase design saves by itself.

7 Conclusion

LLMs make duplicate measures cheap, and ORIV turns duplicate measures into unbiased estimates. The link between the two is one assumption, that the errors in the duplicates are independent of each other and of the outcome. We have shown why LLM duplicates violate this assumption, what the violation does to ORIV, and how large it is in data where a human retest is available.

Three results stand out. Errors are shared across models from different developers, so more models do not mean more independent information. Errors load on the outcome whenever the model can see information related to it, and this bias survives every combination of models. Item accuracy, the standard measure of LLM fidelity, hides both problems because it rewards predicting the common answer. And when the trait itself is held out from the persona, as with risk attitudes and altruism, the twins carry almost no information about it, even when their regression estimates

happen to match the human ones.

The practical message follows directly. LLM duplicates are not a substitute for the second elicitation that Gillen, Snowberg, and Yariv (2019) built into their study, and neither is a single human answer corrected with LLM measures. The second elicitation is needed, but only for part of the sample. A researcher should ask a random human sample once, ask a random third of it again, and correct the estimate with whatever predicts human answers for everyone. This cuts the human answers needed by a quarter to a third, and targeting the retests at the people whose outcome and first answer disagree most cuts further when either is skewed. LLM measures can be part of that correction, and they cannot bias it. In survey data they add little to the persona data from which they were built. In text, where no such data exist, they add a meaningful share, as the comments application shows. In both cases a pilot measures their value before a researcher pays for them. The design inherits the quality of the human retest and adds no bias of its own. If people remember their first answer, every design converges to the same attenuated value, so a retest worded differently or asked later protects the estimate more than any choice about who is asked again.

References

Aigner, Dennis J. 1973. "Regression with a Binary Independent Variable Subject to Errors of Observation." *Journal of Econometrics* 1 (1): 49–59.

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. "Prediction-Powered Inference." *Science* 382 (6671): 669–74.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31 (3): 337–51.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis* 32 (4): 401–16.

Black, Dan A., Mark C. Berger, and Frank A. Scott. 2000. "Bounding Parameter Estimates with Nonclassical Measurement Error." *Journal of the American Statistical Association* 95 (451): 739–48.

Bokern, Paul, Jona Linde, Arno Riedl, Hans Schmeets, and Peter Werner. 2023. "The Convergent and External Validity of Risk Preference Elicitation Methods: Controlling for Measurement Error in a Large Population Sample." Netspar Discussion Paper 08/2023-038.

Canen, Nathan, and Ted Enamorado. 2026. "When Predictions Become Regressors: A Split-Sample Correction for Biases in Downstream Inference." arXiv 2608.02909.

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer. 2023. "Econographics." *Journal of Political Economy Microeconomics* 1 (1): 115–61.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21 (1): C1–C68.

Dell, Melissa, and Ashesh Rambachan. 2026. "The Measurement Revolution? Credible Measurement and Inference in the Age of AI." NBER Working Paper 35744 and arXiv 2608.23524.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B* 39 (1): 1–38.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3): 522–50.

Duan, Junting, and Markus Pelger. 2026. "Inference with AI-Generated Covariates." NBER Working Paper 35481.

Egami, Naoki, and Sooahn Shin. 2026. "Debiased Inference for AI-Generated Data without Gold-Standard Labels: Identification via Multiple Imperfect Measurements." arXiv 2608.18294.

Espinal Maya, Cristian. 2026. "Measuring What Cannot Be Surveyed: LLMs as Instruments for Latent Cognitive Variables in Labor Economics." arXiv 2604.02403.

Fuller, Wayne A. 1987. *Measurement Error Models*. New York: Wiley.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy* 127 (4): 1826–63.

Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92 (5): 1644–55.

Horton, John J. 2023. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" NBER Working Paper 31122.

Hu, Yingyao. 2008. "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution." *Journal of Econometrics* 144 (1): 27–61.

Jöreskog, Karl G. 1971. "Statistical Analysis of Sets of Congeneric Tests." *Psychometrika* 36 (2): 109–33.

Kane, Thomas J., Cecilia Elena Rouse, and Douglas Staiger. 1999. "Estimating Returns to Schooling When Schooling Is Misreported." NBER Working Paper 7235.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2025. "Large Language Models: An Applied Econometric Framework." NBER Working Paper 33344.

Mahajan, Aprajit. 2006. "Identification and Estimation of Regression Models with Misclassification." *Econometrica* 74 (3): 631–65.

Murphy, Ryan O., Kurt A. Ackermann, and Michel J. J. Handgraaf. 2011. "Measuring Social Value Orientation." *Judgment and Decision Making* 6 (8): 771–81.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–625.

Overgoor, Gijs, Samsun Knight, and Yakov Bart. 2025. "Correcting Toward the Wrong Target: LLM-Annotated Regressors Under Noisy Human Benchmarks." Working paper, Southern Methodist University.

Park, Joon Sung, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. 2024. "LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals." arXiv 2411.10109.

Reiersøl, Olav. 1950. "Identifiability of a Linear Relation between Variables Which Are Subject to Error." *Econometrica* 18 (4): 375–89.

Rick, Scott I., Cynthia E. Cryder, and George Loewenstein. 2008. "Tightwads and Spendthrifts." *Journal of Consumer Research* 34 (6): 767–82.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–66.

Shao, Jun, and Dongsheng Tu. 1995. *The Jackknife and Bootstrap*. Springer Series in Statistics. New York: Springer.

Spearman, Charles. 1904. "The Proof and Measurement of Association between Two Things." *American Journal of Psychology* 15 (1): 72–101.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557–86.

Toubia, Olivier, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. 2025. "Twin-2K-500: A Dataset for Building Digital Twins of over 2,000 People Based on Their Answers to over 500 Questions." arXiv 2505.17479.

van Kippersluis, Hans, Pietro Biroli, Titus Galama, and others. 2023. "Overcoming Attenuation Bias in Regressions Using Polygenic Indices." *Nature Communications* 14.

A Proofs

Throughout, $\sigma^2 \equiv \mathrm{Var}(x^*)$ and $\sigma_\varepsilon^2 \equiv \mathrm{Var}(\varepsilon)$. All components of measurement error are independent of x^* unless stated otherwise.

A.1 The Probability Limit of ORIV

With two measures, the stacked data contain each person twice. In the first copy the regressor is x_a and the instrument is x_b. In the second copy the regressor is x_b and the instrument is x_a. Each copy has its own intercept, so the estimator demeans within copy, and

$$\hat{\beta}_{ORIV} = \frac{\widehat{\mathrm{Cov}}(y,x_b) + \widehat{\mathrm{Cov}}(y,x_a)}{2\,\widehat{\mathrm{Cov}}(x_a,x_b)} \quad \xrightarrow{p} \quad \frac{\mathrm{Cov}(y,x_a) + \mathrm{Cov}(y,x_b)}{2\,\mathrm{Cov}(x_a,x_b)}. \tag{A1}$$

Every proof below evaluates the covariances in equation (A1).

A.2 Proof of Proposition 1

Let $x_a = x^* + c + d_a + u_a$ and $x_b = x^* + c + d_b + u_b$ for two different models, with all components mutually independent and independent of ε. Then $\mathrm{Cov}(y,x_k) = \beta\sigma^2$ for $k = a,b$, and $\mathrm{Cov}(x_a,x_b) = \sigma^2 + \sigma_c^2$. Substituting into equation (A1) gives $\beta\sigma^2/(\sigma^2+\sigma_c^2)$.

For two replicates of model a, $x_{a1} = x^* + c + d_a + u_{a1}$ and $x_{a2} = x^* + c + d_a + u_{a2}$. Now $\mathrm{Cov}(x_{a1},x_{a2}) = \sigma^2 + \sigma_c^2 + \sigma_d^2$, and the limit is $\beta\sigma^2/(\sigma^2+\sigma_c^2+\sigma_d^2)$.

OLS on x_a converges to $\beta\sigma^2/(\sigma^2+\sigma_c^2+\sigma_d^2+\sigma_u^2)$. Because variances are nonnegative, for $\beta > 0$,

$$\beta_{OLS} \le \beta_{ORIV}^{\text{reruns}} \le \beta_{ORIV}^{\text{models}} \le \beta, \tag{A2}$$

with the inequalities reversed for $\beta < 0$. Finally, let $h = x^* + e_h$ be a human measure with e_h independent of every model error and of ε. IV of x_a with instrument h converges to $\mathrm{Cov}(y,h)/\mathrm{Cov}(x_a,h) = \beta\sigma^2/\sigma^2 = \beta$. □

A.3 Proof of Proposition 2

Let $x_m = (1-\kappa_m)x^* + u_m$ with $\kappa_m < 1$ and u_a, u_b, ε mutually independent. Then $\mathrm{Cov}(y, x_m) = \beta(1-\kappa_m)\sigma^2$ and $\mathrm{Cov}(x_a, x_b) = (1-\kappa_a)(1-\kappa_b)\sigma^2$. Therefore

$$\frac{\mathrm{Cov}(y,x_a)\,\mathrm{Cov}(y,x_b)}{\mathrm{Cov}(x_a,x_b)} = \frac{\beta^2(1-\kappa_a)(1-\kappa_b)\sigma^4}{(1-\kappa_a)(1-\kappa_b)\sigma^2} = \beta^2\sigma^2, \tag{A3}$$

and the square root equals $|\beta|\sigma$. The sign of β equals the sign of $\mathrm{Cov}(y, x_a)$ because $1-\kappa_a > 0$. Dividing by σ_y gives the correlation between y and x^*, which is the standardized effect we report.

IV of x_a with instrument x_b converges to $\beta(1-\kappa_b)\sigma^2/[(1-\kappa_a)(1-\kappa_b)\sigma^2] = \beta/(1-\kappa_a)$. The one-direction standardized estimator converges to

$$\frac{\mathrm{Cov}(y,x_b)}{\sqrt{\mathrm{Cov}(x_a,x_b)}} = \beta\sigma\sqrt{\frac{1-\kappa_b}{1-\kappa_a}}, \tag{A4}$$

which equals $\beta\sigma$ if and only if $\kappa_a = \kappa_b$. □

A.4 Proof of Proposition 3

Write $x_k = p_k + (1-p_k-q_k)x^* + v_k$, where $E[v_k \mid x^*] = 0$. This holds because $P(x_k = 1 \mid x^*) = p_k + (1-p_k-q_k)x^*$. Nondifferential misclassification means that (x_a, x_b) is independent of ε given x^*. Conditional independence across models means that v_a and v_b are uncorrelated. Then

$$\mathrm{Cov}(y,x_k) = \beta(1-p_k-q_k)\pi(1-\pi), \tag{A5}$$

$$\mathrm{Cov}(x_a,x_b) = (1-p_a-q_a)(1-p_b-q_b)\pi(1-\pi). \tag{A6}$$

IV of x_a with instrument x_b therefore converges to $\beta/(1-p_a-q_a)$, and equation (A1) gives

$$\beta_{ORIV} = \frac{\beta}{2}\left(\frac{1}{1-p_a-q_a} + \frac{1}{1-p_b-q_b}\right), \tag{A7}$$

which exceeds β in absolute value whenever $0 < p_k + q_k < 1$ for some k.

OLS on x_a converges to $\beta(1-p_a-q_a)\pi(1-\pi)/\operatorname{Var}(x_a)$. To show that this is smaller than β in absolute value, decompose

$$\operatorname{Var}(x_a) = (1-p-q)^2\pi(1-\pi) + \pi q(1-q) + (1-\pi)p(1-p), \tag{A8}$$

dropping the subscript a. The claim is $D \equiv \operatorname{Var}(x_a) - (1-p-q)\pi(1-\pi) > 0$. Collecting terms,

$$D = \pi q(1-q) + (1-\pi)p(1-p) - (1-p-q)(p+q)\pi(1-\pi). \tag{A9}$$

For $0 < \pi < 1$ we have $\pi(1-\pi) < \pi$ and $\pi(1-\pi) < 1-\pi$. Also $1-p-q \le 1-q$ and $1-p-q \le 1-p$. Hence $(1-p-q)q\,\pi(1-\pi) < q(1-q)\pi$ and $(1-p-q)p\,\pi(1-\pi) < p(1-p)(1-\pi)$ whenever $p+q > 0$, so $D > 0$.

With three labels that are independent given x^*, the two-class latent class model is identified up to the labeling of the classes, and the restriction $p+q<1$ fixes the labeling (Hu, 2008). Let $P = E[x^* \mid x_a, x_b, x_c]$. Then $y = \alpha + \beta P + \beta(x^* - P) + \varepsilon$. The term $x^* - P$ is uncorrelated with P by construction, and ε is uncorrelated with P because misclassification is nondifferential. OLS of y on P is therefore consistent for β. □

A.5 Proof of Proposition 4

Let $x_k = x^* + \delta\varepsilon + u_k$ with u_a, u_b independent of each other, of x^*, and of ε. Then $\operatorname{Cov}(y, x_k) = \beta\sigma^2 + \delta\sigma_\varepsilon^2$ and $\operatorname{Cov}(x_a, x_b) = \sigma^2 + \delta^2\sigma_\varepsilon^2$, and equation (A1) gives the expression in the proposition. With $\beta = \sigma^2 = \sigma_\varepsilon^2 = 1$ it equals $(1+\delta)/(1+\delta^2)$, which is greater than 1 if and only if $\delta > \delta^2$, that is, if and only if $0 < \delta < 1$, and smaller than 1 for $\delta < 0$ or $\delta > 1$.

Now pair x_k with a human measure $h = x^* + e_h$ whose error is independent of ε and u_k. IV of x_k with instrument h uses only the human covariance with y and converges to $\operatorname{Cov}(y,h)/\operatorname{Cov}(x_k,h) =$

β. The symmetric estimator of Proposition 2 applied to (x_k, h) uses $\mathrm{Cov}(y, x_k)$ and converges to

$$\sqrt{\frac{(\beta\sigma^2 + \delta\sigma_\varepsilon^2)\,\beta\sigma^2}{\sigma^2}} = \sigma\sqrt{\beta^2 + \beta\delta\sigma_\varepsilon^2/\sigma^2}, \tag{A10}$$

which differs from $\beta\sigma$ whenever $\delta \neq 0$. This is the sense in which the bias survives pairing a model with a human measure. □

A.6 Proof of Proposition 5

(i) Normalize $\mathrm{Var}(x^*) = 1$ and let (x^*, ε) be independent normal variables, independent of c and the u_k. Every observed variable is a linear function of $(x^*, \varepsilon, c, u_1, \ldots, u_K)$, and y and the t_k depend on (x^*, ε) only through the plane they span. Take any $r \in (-1, 1)$ and any variable z in that plane with $\mathrm{Var}(z) = 1$ and $\mathrm{Corr}(y, z) = r$, which exists because the plane is two-dimensional and $|r| < 1$. Define $x^{*\prime} = z$, $\beta' = \mathrm{Cov}(y, z)$, and $\varepsilon' = y - \beta' z$. Then ε' is uncorrelated with z, so the two are independent normal variables, and (z, ε') spans the same plane as (x^*, ε) because z is not proportional to y. Each t_k can therefore be written as $\kappa_k' x^{*\prime} + \delta_k' \varepsilon' + c + u_k$ for unique (κ_k', δ_k'). The primed model satisfies every assumption of the original model, generates exactly the same random variables $(y, t_1, \ldots, t_K)$, and has standardized effect $\mathrm{Corr}(y, x^{*\prime}) = r$. Since r was arbitrary, every value in $(-1, 1)$ is observationally equivalent.

(ii) With $\delta_k = 0$, $\mathrm{Cov}(y, t_k) = \beta\kappa_k$ and, for $j \neq k$, $\mathrm{Cov}(t_j, t_k) = \kappa_j\kappa_k + \sigma_c^2$. If $\beta = 0$ all covariances with y are zero and $\rho = 0$ is identified. Otherwise the ratios $r_k = \kappa_k/\kappa_1 = \mathrm{Cov}(y, t_k)/\mathrm{Cov}(y, t_1)$ are identified and $\kappa_k = s\, r_k$ for an unknown scale s. Because the r_k are not all equal and $K \geq 3$, there are distinct j, k, l with $r_k \neq r_l$, and $r_j \neq 0$ because every κ_j is nonzero. Then $s^2 = (\mathrm{Cov}(t_j, t_k) - \mathrm{Cov}(t_j, t_l))/(r_j(r_k - r_l))$. This gives $|\kappa_1|$, then $|\beta| = |\mathrm{Cov}(y, t_1)/\kappa_1|$ and $|\rho| = |\beta|/\sigma_y$. Changing the sign of every κ_k and of β leaves all second moments unchanged, so the sign of ρ is identified only with the sign of one loading, which holds when one measure is known to rise with the trait.

(iii) This is Proposition 2 applied to the two human measures, whose covariances with y and

with each other do not involve the t_k. □

A.7 Proof of Proposition 6

Let S be the subsample of n people, drawn at random from the N people of the full sample, and treat the N people as independent draws. Write $f_{ya,i} = (y_i - \mu_y)(h_{a,i} - \mu_a)$, define $f_{yb,i}$ and $f_{ab,i}$ in the same way, and let C_{ya}, C_{yb}, and C_{ab} be their means. The influence function of the symmetric estimator is

$$\psi_i = \frac{\rho}{2}\left(\frac{f_{ya,i} - C_{ya}}{C_{ya}} + \frac{f_{yb,i} - C_{yb}}{C_{yb}} - \frac{f_{ab,i} - C_{ab}}{C_{ab}}\right). \tag{A11}$$

Under Assumption A the symmetric estimator is consistent for ρ on any random sample (Proposition 2). The human-only estimator divides by the standard deviation of y in the subsample, which adds $-(\rho/2)\left((y_i - \mu_y)^2/\sigma_y^2 - 1\right)$ to ψ_i. We include this term in ψ from here on. By the delta method $\hat{\rho}_n - \rho = \bar{\psi}_S + o_p(n^{-1/2})$. The added term is a function of $\tilde{y}^2$, which is an element of g, so it changes V_ψ and R^2 but not the variance of the residual of ψ on g.

Because S is random, $\bar{g}_S$ and $\bar{g}_N$ have the same expectation for any g with finite variance. The estimated coefficient $\hat{\Lambda}$ converges to $\Lambda = V_g^{-1} C_{g\psi}$, where V_g is the variance of g and $C_{g\psi}$ is its covariance with ψ. Hence

$$\hat{\rho}_H - \rho = \bar{\psi}_S - \Lambda'(\bar{g}_S - \bar{g}_N) + o_p(n^{-1/2}), \tag{A12}$$

which has mean zero. No assumption on the twins enters except the finite variance of g. The twins' errors may be shared across models, compressed, or correlated with ε, and the estimator remains consistent.

For the variance, $\mathrm{Var}(\bar{\psi}_S) = V_\psi/n$ and $\mathrm{Var}(\bar{g}_S - \bar{g}_N) = V_g(1/n - 1/N)$. Because S is contained in the full sample, $\mathrm{Cov}(\bar{\psi}_S, \bar{g}_N) = C_{g\psi}/N$ and $\mathrm{Cov}(\bar{\psi}_S, \bar{g}_S - \bar{g}_N) = C_{g\psi}(1/n - 1/N)$. Therefore

$$\mathrm{Var}(\hat{\rho}_H) = \frac{V_\psi}{n} - \left(\frac{1}{n} - \frac{1}{N}\right) C_{g\psi}' V_g^{-1} C_{g\psi} = V_\psi\left(\frac{1 - R^2}{n} + \frac{R^2}{N}\right), \tag{A13}$$

with $R^2 = C_{g\psi}' V_g^{-1} C_{g\psi} / V_\psi$. Asymptotic normality follows from the central limit theorem applied to the linear terms. A study that uses only humans, with n_H people, has variance V_ψ / n_H. Setting the two variances equal gives $n = (1 - R^2)\, n_H / (1 - R^2 n_H / N)$.

When the N people are instead treated as a fixed population, as in the subsampling exercise of Section 5, both variances lose the term V_ψ / N. The human-only variance becomes $V_\psi(1/n - 1/N)$ and the hybrid variance becomes $V_\psi(1 - R^2)(1/n - 1/N)$. Their ratio is exactly $1 - R^2$, which is how Section 5 measures R^2 out of sample. □

A.8 Proof of Proposition 7

Let $h_a = x^* + e_a$ with e_a independent of x^* and ε, and let $\lambda_a = \sigma_{x^*}^2 / (\sigma_{x^*}^2 + \sigma_{e_a}^2)$. Then $\mathrm{Cov}(y, h_a) = \beta \sigma_{x^*}^2$ and

$$\mathrm{Corr}(y, h_a) = \frac{\beta \sigma_{x^*}^2}{\sigma_y \sqrt{\sigma_{x^*}^2 + \sigma_{e_a}^2}} = \frac{\beta \sigma_{x^*}}{\sigma_y} \sqrt{\lambda_a} = \rho \sqrt{\lambda_a}. \tag{A14}$$

The raw estimator is a smooth function of the sample moments of y and h_a and converges to this value. The corrected estimator adds a term $-\hat{\Lambda}'(\bar{g}_S - \bar{g}_N)$, which converges to zero for the reason given in the proof of Proposition 6. It therefore has the same probability limit. More LLM measures or a larger N change its variance and leave its limit at $\rho \sqrt{\lambda_a}$.

For the second part, let $t = \kappa x^* + u + \delta \varepsilon$ with u independent of x^* and ε, and allow $\mathrm{Cov}(u, e_a) = \sigma_{ua}$. Then $\mathrm{Cov}(y, h_a) = \beta \sigma_{x^*}^2$, $\mathrm{Cov}(y, t) = \beta \kappa \sigma_{x^*}^2 + \delta \sigma_\varepsilon^2$, and $\mathrm{Cov}(h_a, t) = \kappa \sigma_{x^*}^2 + \sigma_{ua}$. Equation (3) gives

$$\frac{1}{\sigma_y} \sqrt{\frac{\beta \sigma_{x^*}^2 \left(\beta \kappa \sigma_{x^*}^2 + \delta \sigma_\varepsilon^2\right)}{\kappa \sigma_{x^*}^2 + \sigma_{ua}}}, \tag{A15}$$

which equals $\rho = \beta \sigma_{x^*} / \sigma_y$ if and only if $\delta \sigma_\varepsilon^2 = \beta \sigma_{ua}$. With $\sigma_{ua} = 0$ and $\beta \neq 0$ this requires $\delta = 0$. A loading on the outcome and a correlation with the human error can offset each other only on this knife edge. □

A.9 Proof of Proposition 8

Write the standardized effect as a function of five moments,

$$\rho = \text{sign}(m_1)\sqrt{\frac{m_1 m_2}{m_3 - m_4 m_5}}\Big/\sigma_y, \tag{A16}$$

with $m_1 = E[\tilde{y}h_a]$, $m_2 = E[\tilde{y}h_b]$, $m_3 = E[h_a h_b]$, $m_4 = E[h_a]$, $m_5 = E[h_b]$, and $\tilde{y}$ the outcome minus its mean. Let ψ be the influence function of ρ in these moments. Let g_0 collect functions of information observed for everyone and g_1 add functions that also involve h_a. The functions $\tilde{y}h_a$ and h_a are elements of g_1.

Moments observed on S_1 are estimated by $\bar{f}_{S_1} - \hat{L}_0'(\bar{g}_{0,S_1} - \bar{g}_{0,N})$. Moments observed only on S_2 are estimated by $\bar{f}_{S_2} - \hat{L}_1'(\bar{g}_{1,S_2} - \hat{\mu}_{g_1})$, where $\hat{\mu}_{g_1} = \bar{g}_{1,S_1} - \hat{L}_0'(\bar{g}_{0,S_1} - \bar{g}_{0,N})$ is the corrected mean of g_1. Because the moments observed on S_1 are elements of g_1, their estimator is also of the second form with a coefficient that selects them, so every moment is estimated by the same formula. Let Λ_1 be the coefficient of the population projection of ψ on g_1 and Λ_0 that of ψ on g_0. Because g_0 is contained in g_1, the projection of $\Lambda_1' g_1$ on g_0 is $\Lambda_0' g_0$. By the delta method,

$$\hat{\rho} - \rho = \bar{e}_{2,S_2} + \bar{e}_{1,S_1} + \bar{e}_{0,N} + o_p(n_2^{-1/2}), \tag{A17}$$

with $e_2 = \psi - \Lambda_1' g_1$, $e_1 = \Lambda_1' g_1 - \Lambda_0' g_0$, and $e_0 = \Lambda_0' g_0$, each centered at its mean. Each term has mean zero whatever the relation between the auxiliary information and x^*, ε, or the human errors, so $\hat{\rho}$ is consistent for ρ.

The samples are nested and random. The covariance between $\bar{e}_{2,S_2}$ and $\bar{e}_{1,S_1}$ is $\text{Cov}(e_2, e_1)/n_1$, which is zero because e_2 is uncorrelated with every function in g_1. The same argument gives zero for the other two pairs. Hence

$$\text{Var}(\hat{\rho}) = \frac{\text{Var}(e_2)}{n_2} + \frac{\text{Var}(e_1)}{n_1} + \frac{\text{Var}(e_0)}{N} = V_\psi \left(\frac{1 - R_1^2}{n_2} + \frac{R_1^2 - R_0^2}{n_1} + \frac{R_0^2}{N} \right). \tag{A18}$$

For the design, minimize $c_1 n_1 + c_2 n_2$ subject to $A_2/n_2 + A_1/n_1 = v$, with $A_2 = 1 - R_1^2$, $A_1 =$

$R_1^2 - R_0^2$, and v the target variance divided by V_ψ net of the term in N. The first-order conditions give $c_j = \lambda A_j / n_j^2$, so $n_j \propto \sqrt{A_j / c_j}$ and

$$\frac{n_2}{n_1} = \sqrt{\frac{(1-R_1^2)\,c_1}{(R_1^2 - R_0^2)\,c_2}}. \tag{A19}$$

When this ratio exceeds one, the constraint $n_2 \leq n_1$ binds and everyone in the subsample answers twice. Adding $c_N N$ to the cost and A_0/N with $A_0 = R_0^2$ to the constraint gives in the same way $N/n_2 = \sqrt{R_0^2 c_2 / ((1-R_1^2) c_N)}$.

When the auxiliary information includes a prediction trained on S_1, each person in S_1 receives the prediction of a model trained without their fold, and $\bar{g}_{0,N}$ is the average over folds of the full-sample means under each fold's model. Conditional on the models, the people of each fold are a random sample, so every correction keeps mean zero. □

A.10 Proof of Proposition 9

Treat the N people as independent draws. Each enters S_1 independently with probability $p_1 = n_1/N$ and, given W_1, is asked again with probability $\pi(W_1) \geq \underline{\pi} > 0$, both chosen by the researcher. Let R_1 and R_2 indicate the two events. With both answers for everyone, ρ is a smooth function of five means, and in a nonparametric model the plug-in estimator is efficient with influence function ψ. The observed data are a monotone coarsening of the full data with known coarsening probabilities that depend only on what is observed at each stage. By the representation of Robins, Rotnitzky, and Zhao (1994) for such data, every regular asymptotically linear estimator has an influence function of the form

$$\phi = h_0(W_0) + \frac{R_1}{p_1}\big(h_1(W_1) - h_0(W_0)\big) + \frac{R_1 R_2}{p_1 \pi(W_1)}\big(\psi - h_1(W_1)\big) \tag{A20}$$

for some functions h_0 and h_1, plus terms with mean zero that only add variance. Let $h_0^* = E[\psi \mid W_0]$ and $h_1^* = E[\psi \mid W_1]$, and let ϕ^* be ϕ evaluated at them. At these functions the three terms of ϕ^* are

uncorrelated, because each has mean zero given the information of the stage before, so

$$\mathrm{Var}(\phi^*) = \mathrm{Var}(h_0^*) + \frac{E[(h_1^* - h_0^*)^2]}{p_1} + \frac{1}{p_1} E\left[\frac{\sigma^2(W_1)}{\pi(W_1)}\right]. \tag{A21}$$

For any other h_0 and h_1,

$$\phi - \phi^* = \Big(1 - \frac{R_1}{p_1}\Big)(h_0 - h_0^*) + \frac{R_1}{p_1}\Big(1 - \frac{R_2}{\pi(W_1)}\Big)(h_1 - h_1^*), \tag{A22}$$

and each term is uncorrelated with ϕ^* and with the other, because $E[R_1/p_1] = 1$, $E[R_2/\pi \mid W_1] = 1$, $E[h_1^* - h_0^* \mid W_0] = 0$, and $E[\psi - h_1^* \mid W_1] = 0$. Hence

$$\mathrm{Var}(\phi) = \mathrm{Var}(\phi^*) + \frac{1-p_1}{p_1} E[(h_0 - h_0^*)^2] + \frac{1}{p_1} E\left[\frac{1-\pi(W_1)}{\pi(W_1)}(h_1 - h_1^*)^2\right], \tag{A23}$$

which is minimized by $h_0 = h_0^*$ and $h_1 = h_1^*$. Dividing by N gives the bound in the text. The estimator of Section 5 has influence function ϕ with h_0 and h_1 equal to the limits of its corrections, so it attains the bound when those limits are the conditional means. When they are not, ϕ still has mean zero, so the estimator remains consistent and asymptotically normal, with the larger variance $\mathrm{Var}(\phi)/N$. With π constant and least-squares projections on nested features, the three terms of ϕ are again uncorrelated, $\mathrm{Var}(h_0) = V_\psi R_0^2$, $E[(h_1 - h_0)^2] = V_\psi(R_1^2 - R_0^2)$, and $E[(\psi - h_1)^2] = V_\psi(1 - R_1^2)$, which gives equation (10). □

A.11 Proof of Proposition 10

With expected cost $n_1(c_1 + c_2 E[\pi])$ fixed at B, the variance in Proposition 9, net of the term in N, is proportional to

$$\big(c_1 + c_2 E[\pi]\big)\big(A + E[\sigma^2/\pi]\big), \qquad A = E[(h_1 - h_0)^2],\ \sigma^2(W_1) = \mathrm{Var}(\psi \mid W_1). \tag{A24}$$

The derivative with respect to $\pi(w)$ at a point where $0 < \pi(w) < 1$ is proportional to $c_2(A + E[\sigma^2/\pi]) - (c_1 + c_2 E[\pi])\sigma^2(w)/\pi(w)^2$. Setting it to zero gives $\pi(w) = \lambda\sigma(w)$ with λ common to all w. Let $\mathscr{C}$ be the set where the constraint $\pi \leq 1$ binds. Substituting $\pi = \lambda\sigma$ off $\mathscr{C}$ and $\pi = 1$ on it, the first-order condition becomes $\lambda^2 = (c_1 + c_2 P(\mathscr{C}))/(c_2(A + E[\sigma^2 \mathbf{1}_{\mathscr{C}}]))$, which reduces to $\lambda^2 = c_1/(c_2 A)$ when no constraint binds. For a given expected share $E[\pi]$, minimizing $E[\sigma^2/\pi]$ subject to $0 < \pi \leq 1$ is a convex problem, and its solution has the form $\pi = \min\{1, \lambda\sigma\}$ by the first-order condition above. The optimal design is therefore in this family, and the condition on λ selects the best member.

For the standardized effect, write $\tilde{h}_b = h_b - E[h_b]$. Given W_1, the influence function ψ of equation (3) is linear in h_b, with slope

$$\beta(W_1) = \frac{\rho}{2}\left(\frac{\tilde{y}}{\mathrm{Cov}(y, h_b)} - \frac{h_a - E[h_a]}{\mathrm{Cov}(h_a, h_b)}\right), \tag{A25}$$

so $\sigma(W_1) = |\beta(W_1)|\,\mathrm{sd}(h_b \mid W_1)$. When σ is constant, π^* is constant and the design is that of Proposition 8. Otherwise the retest term of the variance falls, at a given expected number of retests, by the factor $(E[\sigma])^2/E[\sigma^2]$. □

A.12 Assumption B and Theorem A1: Asymptotic Theory with Estimated Corrections

The corrections used in Section 5 are estimated, either as linear projections on a fixed set of features or as cross-fitted predictions. The following conditions cover both.

Assumption B *(B1) The N people are independent draws from a population, and* $n_1/N \to p_1 \in (0, 1]$. *(B2) Each person enters the first sample independently with probability* p_1 *and, given* W_1*, is asked again with a probability* $\pi(W_1)$ *that is known to the researcher, satisfies* $\pi(W_1) \geq \underline{\pi} > 0$*, and does not depend on the sample itself. (B3)* y*,* h_a*,* h_b *and every feature have finite fourth moments, and the second-moment matrices of the features are nonsingular. (B4)* $\mathrm{Cov}(h_a, h_b) > 0$ *and* $\rho \neq 0$*. (B5) Each correction is either a linear projection on a fixed, finite set of features, estimated by*

least squares, or a prediction estimated by K-fold cross-fitting that converges in mean square to a fixed function. Neither limit needs to equal the conditional mean.

Condition (B2) describes independent selection. The code draws samples of fixed size without replacement, which has the same first-order behavior when n_1/N and n_2/n_1 converge, and gives smaller variances by the finite-population factors $1-n_1/N$ and $1-n_2/n_1$. This is why coverage in our subsampling exercises exceeds 95 percent when the human sample is a large share of the population. Condition (B2) also requires that the retest probabilities be set before the study sample is drawn, for instance from a separate pilot as in Section 5.

Theorem A1 *Under Assumption B, let $\bar{h}_0$ and $\bar{h}_1$ be the limits of the estimated corrections. Then*

$$\sqrt{n_1}\,(\hat{\rho}-\rho) \to \mathcal{N}\big(0,\ p_1 \operatorname{Var}(\bar{\phi})\big), \tag{A26}$$

where $\bar{\phi}$ is the term ϕ of the proof of Proposition 9 evaluated at $\bar{h}_0$ and $\bar{h}_1$. When $\bar{h}_0$ and $\bar{h}_1$ are the conditional means, or least-squares projections on nested features with π constant, the variance equals $p_1 \operatorname{Var}(\bar{h}_0) + E[(\bar{h}_1-\bar{h}_0)^2] + E[(\psi-\bar{h}_1)^2/\pi(W_1)]$. The variance estimator that computes each person's contribution to the moments and estimates their variance by regression on the features of W_0 is consistent.

Proof. Write the estimator of each moment as the sample mean of the augmented term ϕ of the proof of Proposition 9 with $\hat{h}$ in place of h. The difference between this mean and the same mean with $\bar{h}$ is a sum of terms of the form $(R_1/p_1)(1-R_2/\pi)(\hat{h}_1-\bar{h}_1)$ and $(1-R_1/p_1)(\hat{h}_0-\bar{h}_0)$. Under cross-fitting each $\hat{h}$ is estimated without the person it is evaluated at, so given the other folds each term has mean zero because π and p_1 are known. Its variance is bounded by $\underline{\pi}^{-1}E[(\hat{h}-\bar{h})^2]$, which converges to zero. The difference is therefore $o_p(n_1^{-1/2})$, and no rate of convergence is needed beyond consistency, because the selection probabilities are known by design (Chernozhukov et al., 2018). For linear corrections estimated by least squares on the same sample, the estimated coefficients converge at rate $n_1^{-1/2}$ to the projection coefficients under (B3), and they

multiply terms such as $\bar{g}_{S_1} - \bar{g}_N$ that are themselves $O_p(n_1^{-1/2})$, so their estimation error is again $o_p(n_1^{-1/2})$. This is the standard argument for the regression estimator in two-phase sampling. The mean with $\bar{h}$ is a mean of independent terms and obeys the central limit theorem under the moment conditions. The delta method applied to $\rho = \text{sign}(m_1)\sqrt{m_1 m_2/(m_3 - m_4 m_5)}/\sigma_y$, which is differentiable when $\text{Cov}(h_a, h_b) > 0$ and $\rho \neq 0$, gives the limit. Consistency of the variance estimator follows from the same argument applied to second moments. The jackknife variance of Section 5.5 is consistent because $\hat{\rho}$ is a smooth function of sample means (Shao and Tu, 1995). Fitting each person's correction without that person changes each mean by $O_p(n_1^{-1})$ and leaves this argument intact. When $\rho = 0$ the square root is not differentiable, and confidence sets should be obtained by inverting tests of $m_1 m_2 = \rho_0^2 \sigma_y^2 (m_3 - m_4 m_5)$, a smooth function of the moments. □

A.13 The Measures Used in Section 4

Let $h_j = x^* + e_j$ for waves $j = 1, 2$, with e_1 and e_2 independent of each other and of x^*. Let a twin be $t = \lambda_t x^* + \gamma_t e_1 + \eta_t$, where γ_t measures copying of the wave 1 to 3 error and η_t is independent of e_2. We assume that x^* is stable over the two weeks between waves. Then

$$\text{Cov}(h_1, h_2) = \sigma^2, \qquad \text{Cov}(t, h_2) = \lambda_t \sigma^2, \qquad \text{Cov}(t, h_1) = \lambda_t \sigma^2 + \gamma_t \sigma_{e_1}^2. \tag{A27}$$

Validity. $\text{Cov}(t, h_2)/\sqrt{\text{Cov}(h_1, h_2)\,\text{Var}(t)} = \lambda_t \sigma/\sigma_t = \text{Corr}(t, x^*)$.

Copying. $1 - \text{Cov}(t, h_2)/\text{Cov}(t, h_1) = \gamma_t \sigma_{e_1}^2/(\lambda_t \sigma^2 + \gamma_t \sigma_{e_1}^2)$, which is zero if and only if $\gamma_t = 0$.

Shared error. For two twins,

$$\text{Cov}(t_a, t_b) = \lambda_a \lambda_b \sigma^2 + \gamma_a \gamma_b \sigma_{e_1}^2 + \text{Cov}(\eta_a, \eta_b), \qquad \frac{\text{Cov}(t_a, h_2)\,\text{Cov}(t_b, h_2)}{\text{Cov}(h_1, h_2)} = \lambda_a \lambda_b \sigma^2. \tag{A28}$$

The difference is the covariance of the twins' errors, including any copying they share. Section 4 finds γ_t close to zero, so the difference measures $\text{Cov}(\eta_a, \eta_b)$. If the trait changes between waves in a way unrelated to its level, $x_2^* = x^* + \nu$, none of these expressions changes. If it partly reverts

with a stable variance, $x_2^* = \rho x^* + \nu$ with $\rho < 1$ and $\mathrm{Var}(x_2^*) = \sigma^2$, then $\mathrm{Cov}(h_1, h_2)$ and $\mathrm{Cov}(t, h_2)$ both fall by the factor ρ. Validity is then understated by the factor $\sqrt{\rho}$, and a reported shared share s corresponds to a true share $1 - (1-s)/\rho$. The retest correlations bound ρ from below by 0.86 for purchases and 0.92 for policy, so validity is understated by at most 7 percent, and the shared share for purchases is at least 0.73 to 0.86 instead of 0.77 to 0.88. The human ORIV benchmark is unaffected, because $\mathrm{Cov}(y, h_2)$ and $\mathrm{Cov}(h_1, h_2)$ fall by the same factor.

B Reruns, Repeated Questions, and Fine-Tuning

Three further twin versions released with Twin-2K-500 test the framework without new model calls. Table A1 reports them.

Reruns. A second run of GPT-4.1-mini at its default temperature agrees with the main run on 87 percent of items. For purchase propensity the two runs correlate at 0.71, and 95 percent of their covariance is shared error, more than for any pair of developers (77 to 88 percent). ORIV on the two runs moves the effect only from 0.17 (OLS) to 0.18, as Proposition 1 predicts for replicates. For policy liberalism the shared share is 0.42 for the two runs and for every pair of developers alike, so on this score almost all twin error is common to all models.

Repeated questions. The version labeled "repeating questions" in the release matches the original answers on 67 percent of purchase items and the retest on 67 percent, the same as the main twin. Its copying measure and its validity are unchanged. In every released version the copying channel is closed, so the test of Section 4 serves as a check and not as a finding.

Fine-tuning. A GPT-4.1-mini model fine-tuned on 500 people is evaluated on the 1,558 people outside its training set. It is far less over-coherent. Its OLS effect of policy liberalism on political views is 0.41, against 0.89 for the base twin and 0.74 for humans. It also shares less error with other twins (0.32 to 0.37 of the covariance for policy, against 0.40 to 0.43). The gain comes at a cost in validity, which falls from 0.76 to 0.41 for policy and from 0.22 to 0.06 for purchases, and its purchase scores are 2.7 times as dispersed as human ones. ORIV that pairs the fine-tuned twin

with another twin still overstates the policy effect (0.86 to 0.89 against 0.74), by less than pairs of base twins. Fine-tuning trades shared error for noise. It does not make twins valid duplicates.

Table A1. Reruns, Repeated Questions, and Fine-Tuning

	Purchase propensity		Policy liberalism	
	(1)	(2)	(3)	(4)
Panel A. One twin: validity and OLS effect				
GPT-4.1-mini, main setting	0.22	0.168	0.75	0.890
GPT-4.1-mini, default temperature	0.18	0.130	0.75	0.886
GPT-4.1-mini, repeating questions	0.23	0.167	0.75	0.890
GPT-4.1-mini, fine-tuned[a]	0.06	0.074	0.41	0.409
Panel B. Two twins: shared error share and ORIV effect				
GPT-4.1-mini, two runs	0.95	0.175	0.42	0.900
GPT-4.1-mini and GPT-4.1	0.88	0.269	0.42	0.911
GPT-4.1-mini and Gemini	0.77	0.525	0.42	0.914
Fine-tuned and GPT-4.1-mini[a]	0.94	0.234	0.37	0.858
Fine-tuned and Gemini[a]	0.86	0.510	0.32	0.894
Panel C. Human benchmark: ORIV on original and retest				
All people		0.201		0.725
People outside the fine-tuning sample[a]		0.200		0.737

Notes: Columns (1) and (3) report twin validity in Panel A and the share of twin covariance not explained by the human trait in Panel B. Columns (2) and (4) report standardized effects on the spendthrift index and on political views. [a]Evaluated on the 1,558 people outside the fine-tuning training set. Bootstrap intervals from 1,000 draws are in the replication package. $N = 2{,}058$ unless marked.

C Benchmarks against Published Estimates

Table A2 places our estimates next to published ones. We compute each of our statistics the way the cited paper computes it.[20]

Four comparisons stand out. First, our human measures are more reliable than the preference elicitations in Gillen, Snowberg, and Yariv (2019), where two elicitations of risk aversion correlate at about 0.6. Purchase propensity averages 40 decisions and policy attitudes are stable, so attenuation is small. ORIV raises the human coefficient by 5 to 10 percent, less than the 12 to 22 percent

[20]Item accuracy is the share of items on which two answers are identical, for the 40 binary purchase items and the 10 five-point policy items. Normalized accuracy divides twin accuracy by human retest accuracy, as in Toubia et al. (2025) and Park et al. (2024).

Table A2. Benchmarks against Published Estimates

Statistic	This paper	Published	Source
Correlation between two human elicitations	0.86 (purchase), 0.92 (policy), two weeks apart	About 0.6 (risk), 0.58 to 0.67 across waves	Gillen, Snowberg, and Yariv (2019)
Twin item accuracy relative to human retest	0.80 to 0.85 (purchase), 0.60 to 0.62 (policy)	0.88 (twins, all items)	Toubia et al. (2025)
		0.86 (surveys and interviews)	Park et al. (2024)
Same, demographics-only agent	0.70 (purchase), 0.57 (policy)	0.74	Park et al. (2024)
Person-level validity relative to human retest	0.11 to 0.25 (purchase), 0.66 to 0.78 (policy)	Not reported	
Correlation between two LLM measures	0.09 to 0.32 (purchase), 0.77 to 0.94 (policy)	0.76	Espinal Maya (2026)
Standard deviation, LLM relative to human	0.56 to 1.00 (purchase), 0.60 to 1.78 (policy)	About 0.51	Bisbee et al. (2024)
ORIV relative to OLS, human duplicates	1.10 (purchase), 1.05 (policy)	1.12 (height), 1.22 (education)	van Kippersluis et al. (2023)
		1.25 (LLM duplicates)	Espinal Maya (2026)
Twin ORIV relative to human ORIV	1.34 to 2.62 (purchase), 1.25 to 1.26 (policy)	Not reported	

Notes: Ranges across the four twins or the six pairs of twins. The published standard deviation ratio is computed from Table 1 of Bisbee et al. (2024) for affective polarization among partisans. The 0.88 for Toubia et al. (2025) is 71.7 divided by 81.7.

that van Kippersluis et al. (2023) find for polygenic indices. Corrected correlations between preference measures in Gillen, Snowberg, and Yariv (2019) and Bokern et al. (2023) rise far more, often from below 0.45 to above 0.6. Section 5 of the paper turns to those preferences.

Second, item accuracy and person-level validity tell different stories. On purchase items, twins reach 80 to 85 percent of human retest accuracy, which matches the 86 to 88 percent reported by Toubia et al. (2025) and Park et al. (2024). The demographics-only twin reaches 70 percent, close to the 74 percent Park et al. (2024) report for demographic agents. At the level of the person, the same twins recover only 11 to 25 percent of the retest correlation. Figure A1 shows the gap. Purchase decisions have a common answer for most products, and a twin that predicts it scores well on accuracy while missing who buys more than whom. A regression uses only the second kind of information.

Third, agreement between models is not evidence of independent errors. Espinal Maya (2026) reports a correlation of 0.76 between two LLMs and treats it as support for ORIV. Our policy twins agree at 0.77 to 0.94, yet 40 to 43 percent of their covariance is shared error. Agreement measures the sum of shared signal and shared error, and ORIV needs the second to be zero.

Fourth, LLMs do not simply compress. Bisbee et al. (2024) find synthetic responses with about half the human standard deviation. Our twins range from 0.56 to 1.78 times the human standard deviation depending on model and score, which is why we use the scale-robust estimator of equation (3). Bisbee et al. (2024) also find that partisanship predicts attitudes more strongly in synthetic data than in survey data. Our twins show the same over-coherence between ideology and policy positions, and Proposition 4 explains why it biases regression estimates.

D LLM Prompts and Protocol

D.1 Persona

For a new sample, each person's demographics fill the following template. The persona contains no answer to any task or outcome question, except in the randomized arm described below.

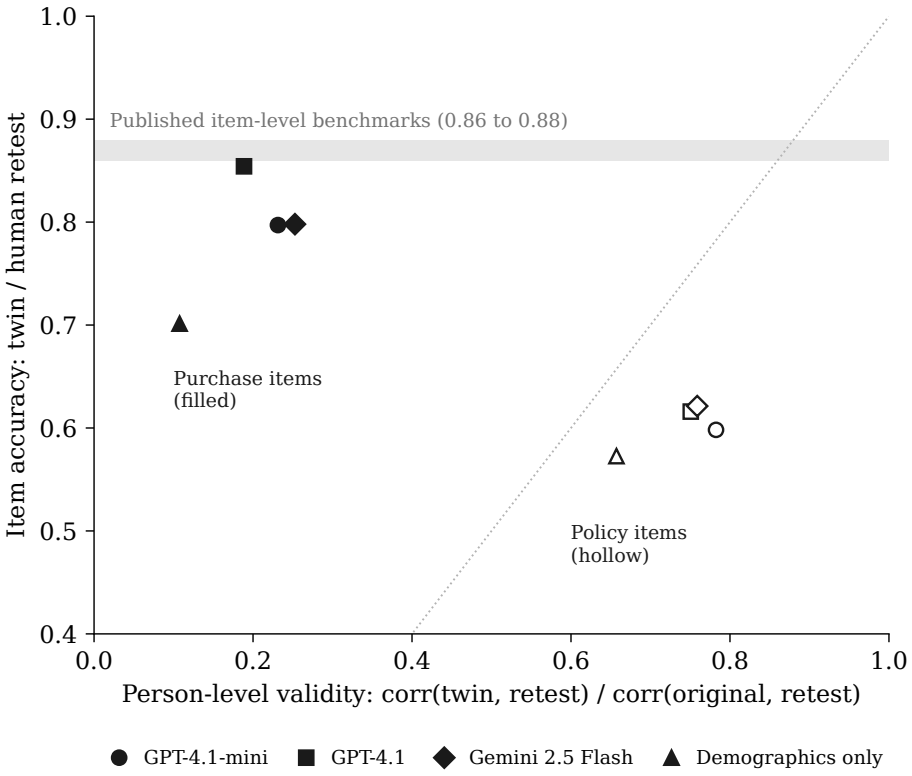


Figure A1. Item Accuracy and Person-Level Validity

Notes: Vertical axis: share of items on which the twin matches the wave 4 answer, divided by the same share for the waves 1 to 3 answer. Horizontal axis: correlation of the twin score with the wave 4 score, divided by the correlation of the waves 1 to 3 score with the wave 4 score. The shaded band marks published item-level estimates. The dotted line is the 45-degree line.

```
You are taking part in a study as the following person. Age: {age}.
Gender: {gender}. Highest education: {education}. Household income:
{income band}. Region: {region}. Employment: {employment status}.
Answer every question as this person would. Answer only with the
requested format.
```

In the randomized arm, for two models and for a random half of personas, the template adds the participant's answers to the three outcome questions. Comparing the two halves identifies the loading δ of Proposition 4.

D.2 Tasks

The model receives the same instructions as human participants, one part per message, in the participant's randomized order, within a single conversation. Comprehension checks are omitted. After each part the model returns a JSON object. For the multiple price list it returns a JSON object with one entry per row, such as `"row_1": "A"`. For the certainty equivalents it returns the first row at which it prefers the sure amount for each lottery. For the dictator game it returns the amount sent, and for the slider it returns the selected position on each item.

D.3 Models and Settings

Models from at least three developers answer every persona, including at least one open model run from Hugging Face. We record the exact model version string and date of every call. Calls use temperature 1 where the model accepts a temperature setting. For one closed and one open model, each persona is answered three times to estimate sampling noise separately from model-specific error. For open models we also record the log-probability of each option, which gives a continuous measure of each choice. For one open model family we run both the base and the instruction-tuned version.

D.4 Held-Out Twins within Twin-2K-500

The same protocol applies to the existing Twin-2K-500 participants. Each persona is rebuilt from the person's waves 1 to 3 answers, in the question-and-answer format of Toubia et al. (2025), with the economic preferences block removed. The tasks are the three certainty-equivalent lists, the dictator game, the trust game, and the ultimatum game, worded as in the original survey. The risk lists state that option 1 keeps the lottery and option 2 takes the sure amount, because the original screen conveyed this by position. The outcome items, the empathy scale (QID232) and the four spendthrift items (QID31 to QID34), are removed from every persona and then added back for a random half of people. The outcomes are the spendthrift index for risk attitudes and the empathy

scale for altruism. Each call is about 19,000 input tokens (about 32,000 in Claude's tokenizer). The run in Section 5.10 uses Claude Haiku 5.5 (model `claude-haiku-5-5`, effort set to low, structured JSON output) for all 2,058 people, on 18 October 2026, with a second run for a fixed random subset of 300 people. Every one of the 2,362 answers parsed, and the run cost about $9. The analysis estimates the hybrid of Section 5 for human subsamples from 50 people to the full sample, and ORIV on pairs of twins with and without controls for the demographics in the persona. The code is in the replication package (folder `code/new_twins`), with a synthetic model that tests every step without API calls.

D.5 Toxicity Labels for Online Comments

For Section 6, Claude Sonnet 5.5 (model `claude-sonnet-5-5`, effort set to low, structured JSON output) receives each comment with the following instructions. System: "You rate online comments the way a crowd of raters would. A toxic comment is a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion." User: the comment text, followed by "What percentage of crowd raters (0 to 100) would rate this comment as toxic?" and a request to answer as an integer. The definition of toxicity is the one shown to the crowd raters. The model sees neither the ratings, nor the publication, nor the moderation decision, nor the likes. The data are a random sample of 30,000 comments with at least four raters from the Kaggle competition "Jigsaw Unintended Bias in Toxicity Classification," of which 28,682 have at most ten raters. The code is in the replication package (folder `code/jigsaw`).

E Variables and Simulation Parameters

Table A3. Variable Definitions in Twin-2K-500

Variable	Items	Construction
Purchase propensity	`1_Q295` to `40_Q295`	Share of products the person would buy (1 = buy)
Policy liberalism	QID287, rows 1 to 10	Mean support on a 1 to 5 scale, rows 9 and 10 reversed as 6 minus the answer
Perceived public support	QID290, rows 1 to 10	Mean percent of the public believed to support each policy, rows 9 and 10 as 100 minus the answer
Spendthrift index	QID31 to QID34	Mean of z-scores of QID31, QID32, 6 minus QID33, and 6 minus QID34
Political views	QID22	1 = very conservative to 5 = very liberal
Risk aversion, list k	QID250, QID251, QID252	Number of the 14 rows in which the person takes the sure amount
Dictator giving	QID231	Dollars sent out of \$5
Ultimatum offer	QID224	Dollars offered out of \$5
Trust transfer	QID117	Dollars sent out of \$5
Trustee return share	QID118 to QID122	Mean share returned across the five tripled amounts

Notes: Item codes refer to the question catalog released with Toubia et al. (2025). In the formatted twin files, the first data row holds question text and is dropped. People are matched across files on `TWIN_ID`.

Table A4. Monte Carlo Designs

Design	Parameters
Shared errors	Total error variance 2/3 per measure (reliability 0.6). Share s of it common to all models, $(1-s)/2$ model-specific, $(1-s)/2$ sampling noise. Human error variance 2/3. s from 0 to 1 in steps of 0.1
Binary labels	$\pi = 0.5$, $p = q$ from 0 to 0.25 in steps of 0.05, independent across labels
Equal compression	$\kappa_a = \kappa_b$ from 0 to 0.5, noise variance 0.3
Unequal compression	$\kappa_a = 0$, κ_b from 0 to 0.5, noise variance 0.3
Outcome loading	$e_k = \delta\varepsilon + u_k$, δ from -0.4 to 0.4, $\mathrm{Var}(u_k) = 0.5$, human error variance 0.5
Three labels	As in binary labels, latent class model fitted by EM with 200 iterations
Designs of Section 5	$y = x^* + \varepsilon$, $\mathrm{Var}(x^*) = 1$, $\mathrm{Var}(\varepsilon) = 3$, so $\rho = 0.5$. Two human answers $x^* + e$ with reliability 0.4, 0.6, or 0.8. Three LLM measures $0.8x^* + c + u_k + 0.3\varepsilon/\sqrt{3}$ with $\mathrm{Var}(c) = 0.5$ common to all and $\mathrm{Var}(u_k) = 0.3$. $N = 2{,}000$, a budget of 400 human answers, 1,000 replications
Coverage	$n = 500$, reliability λ from 0.05 to 0.9, error variance $(1-\lambda)/\lambda$, cluster-robust standard errors

Notes: In every design $\beta = 1$, $\mathrm{Var}(x^*) = 1$, $\varepsilon \sim N(0,1)$, 500 replications, and $n = 2{,}000$ unless stated otherwise.