BEHAVIOR UNDERSPECIFICATION IN LARGE HUMAN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large human language models (LHLMs), language models designed to simulate human responses, are increasingly used as scalable simulators of human behavior. Given a description of an individual, an LHLM generates responses from that person's perspective, enabling applications in behavioral science, marketing, public policy, and beyond. But what kind of human behavior do they actually model? Behavioral science has long documented a consequential gap between what people say they do and what they actually do. We hypothesize that LHLMs may not reliably distinguish between these constructs because their training data disproportionately capture what people say, while their training objectives do not explicitly differentiate among them. We refer to this ambiguity as behavior underspecification. To study it, we evaluate ten contemporary LHLMs on a cross-domain dataset comprising 98 matched measures of what people said and eventually did, derived from more than 37K human responses. With experiments, we find that behavior underspecification is widespread, that LHLMs model what people say better than what people do, and predictions of what people do often reflect what people say. The effect of behavior underspecification persists even in otherwise accurate models. Fine-tuning on human-response data does not necessarily resolve this problem and can sometimes exacerbate it. Together, our findings show that evaluating LHLMs requires asking not only how accurately they simulate human responses, but precisely what aspects of human behavior they simulate.

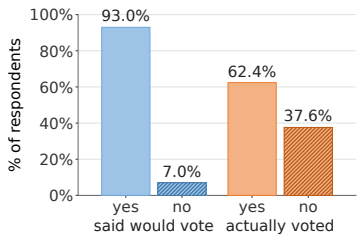
1 INTRODUCTION

LHLMs, language models specifically trained or conditioned to simulate human responses and conversations (Soni et al., 2024), are emerging as a powerful new class of human behavioral simulators. Their premise is deceptively simple: describe an individual through demographic, social, psychological, and contextual attributes, and then ask a language model to respond from that individual's perspective (Zhou et al., 2026). When evaluated at the level of population response distributions, LHLMs have demonstrated remarkable accuracy in some domains, a capability often attributed to their training on internet-scale traces of human expression and behavior, through which they learn the statistical regularities of human culture (Anthis et al., 2025).

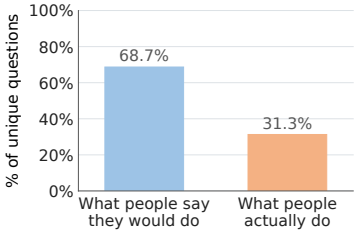
The deeper shift is potentially paradigm-changing for behavioral and social research, including market research, public policy, healthcare, and human-centered AI. Human behavior becomes something that can be computationally instantiated, experimentally controlled, and simulated at scale. This has fueled growing interest in synthetic participants in human studies (Ashokkumar et al., 2026; Kolluri et al., 2025) and human digital twins (Park et al., 2023) across industry and research communities. When such models are used as simulated users for training and evaluating large language models (LLMs) (Wang et al., 2025b; Abdulhai et al., 2025), or to build personalized and adaptive applications (Shaikh et al., 2025), they are often referred to as *user models* or *user simulators*. The appeal is clear. If LHLMs can approximate human responses with sufficient fidelity, they may enable rapid hypothesis testing, lower-cost experimentation, synthetic participant panels at unprecedented scale, and the simulation of behaviors difficult or impractical to study with real humans.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Yet, we observe that this promise rests on a foundational ambiguity: what is actually captured as "human behavior" in these models? The behavioral science and psychology literature have long recognized what people say they do and what they actually do as related but distinct constructs (Sheeran, 2002). Here, we use *what people say* (stated behavior) to refer to stated values, attitudes, and intentions, which capture how individuals describe themselves or anticipate acting. We use *what people do* (realized behavior) to refer to self-reports or observations of realized actions. The systematic discrepancy between the two is described as the say-do gap. For example, people may intend to exercise yet remain sedentary and report an intention to vote but fail to cast a ballot. Figure 1a illustrates a real-world, population-level say–do gap example where 93% of participants said they would vote, but U.S. voter records showed that only 62% of them actually did (Kuriwaki, 2026).



(a) Gap between reported voting intentions and actual voting behavior.



(b) Say–do imbalance in the SocSci210 (Kolluri et al., 2025).[1]

This say–do distinction can pose two fundamental challenges for developing and evaluating LHLMs. First, the composition of their training data is often opaque, making the prevalence of what people say vs do both unknown and potentially imbalanced. Because most internet-sourced data capture what people say, such as in forums and surveys, we hypothesize that what people say is substantially overrepresented relative to realized action. We tested this hypothesis using widely used SocSci210 dataset (Kolluri et al., 2025) of human responses (Figure 1b) and found that 69% of the data represented what people say, whereas only 31% captured actions. Second, even when both forms appear, training rarely distinguishes prospective statements such as "I plan to," "I would," or "I value" from realized actions such as "I did," "I chose," or "I purchased." Thus, the learned behavior is *underspecified* in both its evidentiary balance and its meaning: an LHLM presented as simulating human behavior may model what people say, what they do, or an indeterminate mixture of both. We call this phenomenon, *behavioral underspecification in LHLMs*.

This observation motivates the central question of our work: **What behavioral distribution do LHLMs actually learn when predicting human behavior? Do they model what people say, what they actually do, both equally well, one better than the other, or some implicit combination of the two?** The distinction matters because the appropriate behavioral target depends on the application. Public policy, marketing, and forecasting often seek to predict what people do, whereas recommendation systems, intelligent assistants, and psychological studies may instead target what people say they prefer or intend. Conflating the two can lead systems to optimize for the wrong target. For example, someone may say they prefer exercising in the morning but consistently go to the gym in the evening, a personalized LLM assistant modeling stated preferences may recommend a morning workout when an evening workout better suits the user.

To add this, we also observe that existing LHLM benchmarks largely focus on *simulation accuracy*, measuring how closely model-generated population-level response distributions match those of humans, without distinguishing between stated and realized behavior Hu et al. (2026). This creates a blind spot. A model can achieve high simulation accuracy while systematically reproducing the wrong behavioral construct. Determining whether and when LHLMs model what people say versus what they do is therefore essential for assessing their reliability as simulators of human behavior.

To address these concerns, we investigate whether contemporary LHLMs distinguish between these two constructs. We consider both general-purpose large language models (LLMs) repurposed as LHLMs and specially trained LHLMs, which we refer to as *human behavior models*. Specifically, we assess whether these models can accurately reconstruct the distinct population-level response distributions for what people say and what they do, comparing model-generated distributions with ground-truth data across 98 say–do pairs (what people said they do vs what they actually ended up doing) spanning multiple behavioral domains and collected from more than 37K human responses.

[1] As evaluated using GPT-4o as a judge.

Overall, we find that LHLMs predict what people say more accurately than what they do and often blur the distinction between the two, revealing widespread behavioral underspecification. This gap persists across general-purpose LLMs and specialized human behavior models and is not captured by simulation accuracy alone. Behavioral-science-inspired prompting improves predictions of realized behavior, while fine-tuning on human-response data does not necessarily resolve it but can sometimes exacerbate it. Together, these findings show that evaluating LHLMs requires asking not only *how accurately* they simulate human behavior, but *what behavior* they actually simulate.

Remark. We focus on behavioral underspecification at the population level, where aggregate distributions provide a well-defined and empirically tractable setting for isolating the say–do distinction. This choice allows us to characterize the phenomenon systematically before extending the analysis to the additional heterogeneity inherent in individual-level behavior.

2 Related Work

A growing body of work uses LLMs to simulate human behavior (Dillion et al., 2023; Grossmann et al., 2023; Bail, 2024; Gao et al., 2024), spanning public-opinion surveys (Argyle et al., 2023), economic decision-making (Horton et al., 2023; Xie et al., 2024), classic experiments in psychology and the social sciences (Aher et al., 2023; Binz & Schulz, 2023; Ashokkumar et al., 2026), and interacting populations of agents (Park et al., 2022; 2023; 2024). More recent work explicitly trains LLMs on large-scale human behavioral data for this purpose. SocSci210 pools 2.9 million individual responses from 210 social-science experiments, with fine-tuning improving predicted response distributions on unseen studies (Kolluri et al., 2025). Centaur is fine-tuned on trial-by-trial choices from 160 cognitive experiments and generalizes to held-out participants, task variants, and new domains (Binz et al., 2025). OSim extends this approach to 21.4 million interactions spanning conversational, social, and cognitive tasks (Zhou et al., 2026).

Yet these models remain limited in several ways. They can misrepresent demographic groups (Santurkar et al., 2023), flatten the diversity of human responses (Wang et al., 2025a), fail to reproduce human survey-response biases while exhibiting ordering and labeling biases of their own (Tjuatja et al., 2024; Dominguez-Olmedo et al., 2024), and diverge from humans on normative principles such as fairness and moral judgment (Hosseini & Khanna, 2025; Cheung et al., 2025; Dickerson et al., 2026). This work largely asks how accurately models reproduce human responses. We ask a different complementary question: *which* behavioral construct do they reproduce?

Our work also connects to personalization and user simulation, where simulated users are used to train and evaluate dialogue, recommendation, and conversational systems (Schatzmann et al., 2006; Balog & Zhai, 2024; Zhang et al., 2024; Naous et al., 2025). Recommender-systems research has long recognized that observed behavior, such as clicks and engagement, can diverge from users' stated preferences (Ekstrand & Willemsen, 2016; Kleinberg et al., 2023; 2024). Yet these user models are rarely evaluated on whether they reproduce stated preferences, realized behavior, or both. We aim to directly test this distinction. More details of related work is given in Appendix A.

3 Problem Formulation

Simulating human responses with LHLMs. Let M be an LHLM and $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{N}$ a sample of N participant profiles, often, characterized by demographic attributes. Given a question q, let y_i denote the response provided by participant i. We simulate a response for the same participant using M as $y_i' \sim M(\mathbf{p}_i, q)$ for $i = 1, \ldots, N$. For a response space $\mathcal{Y} = \{1, \ldots, K\}$, such as a Likert scale, the observed and model-implied population response distributions are,

$$P(k) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{y_i = k\}, \qquad P'(k) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{y_i' = k\}, \qquad k \in \mathcal{Y}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Given a distributional distance d, the simulation error is $d(P', P)$, where smaller values indicate greater fidelity to the human-response distribution.

[1]The individuals depicted are AI-generated, not actual study participants, protecting participant privacy and complying with licensing requirements. This study, however, uses profiles and responses from real participants.

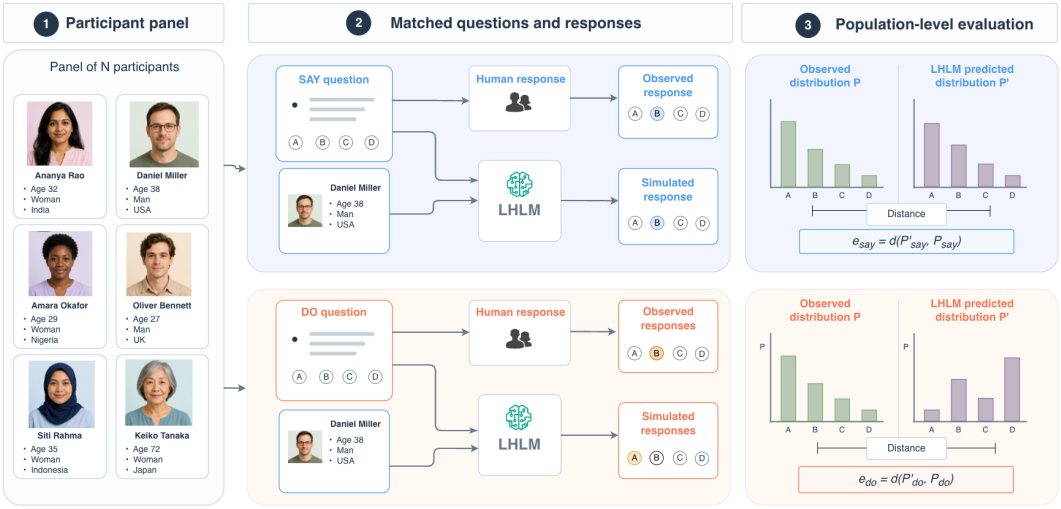


Figure 2: Overview of LHLM simulation of what people say versus do. (1) Human participants are characterized by demographic attributes. (2) For each paired 'say" and 'do" question, the LHLM role-plays each participant using their demographic profile and generates a response. (3) Simulated responses are aggregated into population-level distributions and compared with human distributions.

Simulating what people say versus what they do. Let $\mathcal{Q}$ be a set of matched say–do question pairs $(q_{\text{say}}, q_{\text{do}})$, where q_{say} elicits a stated intention, value, or attitude and q_{do} captures the corresponding realized action. Applying the simulation procedure above to each question yields question-specific observed human distributions $P_{\text{say}}, P_{\text{do}}$ and model-implied distributions $P'_{\text{say}}, P'_{\text{do}}$.

Definition 3.1 (**Fidelity Asymmetry**). For an LHLM M over a set of matched say–do pairs $\mathcal{Q}$,

$$\Delta_{\text{FA}}(M) = \mathbb{E}_{\mathcal{Q}}\left[d(P'_{\text{say}}, P_{\text{say}}) - d(P'_{\text{do}}, P_{\text{do}})\right]$$

We say that M exhibits *fidelity asymmetry* when $\Delta_{\text{FA}}(M) \neq 0$, meaning the model represents what people say and do with different accuracy. A negative value indicates fidelity to stated behavior, while a positive value indicates fidelity to realized behavior. For example, fidelity asymmetry occurs when a model reproduces what people say more accurately than what they actually do.

Definition 3.2 (**Target Ambiguity**). For $t \in \{\text{say}, \text{do}\}$, let $\bar{t}$ denote the alternative. Define,

$$\Delta^{t}_{\text{TA}}(M) = \mathbb{E}_{\mathcal{Q}}[d(P'_t, P_t) - d(P'_t, P_{\bar{t}})]$$

A target is distinguishable when $\Delta^{t}_{\text{TA}}(M) < 0$, meaning that the model-implied distribution is closer to its intended human target than to the alternative. We say that M exhibits *one-sided target ambiguity* when exactly one behavioral target is distinguishable, $\left(\Delta^{\text{say}}_{\text{TA}}(M) \geq 0\right) \oplus \left(\Delta^{\text{do}}_{\text{TA}}(M) \geq 0\right)$, and *two-sided target ambiguity* when neither target is distinguishable, $\left(\Delta^{\text{say}}_{\text{TA}}(M) \geq 0\right) \wedge \left(\Delta^{\text{do}}_{\text{TA}}(M) \geq 0\right)$. For example, one-sided ambiguity occurs when the model distinguishes what people say from what they do when predicting stated behavior but not when predicting realized behavior; two-sided ambiguity occurs when this distinction is absent in both directions.

Definition 3.3 (**Behavioral Underspecification**). We say that behavior is *underspecified* in an LHLM M when it exhibits fidelity asymmetry, target ambiguity, or both.

Remark: We note that, behavioral underspecification is intentionally distinct from the overall simulation accuracy of LHLMs. Instead, behavioral underspecification focuses on an LHLM's ability to faithfully distinguish between what people say and what they do. Consequently, an LHLM may achieve high overall simulation accuracy while still exhibiting behavioral underspecification and, conversely, may successfully distinguish between the two behavioral targets despite relatively low overall simulation accuracy. A comprehensive evaluation of LHLMs should therefore consider both.

4 BENCHMARKING LHLMS ON WHAT PEOPLE SAY VS WHAT THEY DO

4.1 SAY VS DO DATASET

We compiled datasets on what people say and do from public data repositories. Because data on what people do are scarce, we combined external behavioral records, such as U.S. voter files, with self-reported actions. Our dataset contains 98 matched say–do pairs across multiple domains and over 37,000 human responses. For example, one voting pair matches whether participants *intend to vote* (say) with whether they *voted* according to voter records (do) (Kuriwaki, 2026). Because response scales vary within and across pairs, we convert responses to a common binary yes/no scale using pair-specific rules developed with behavioral scientists, excluding neutral or ambiguous responses. While binarization loses response intensity, it preserves the directional distinction central to our analysis. More details on dataset processing and quality checks are in Appendices B and B.5.

4.2 EXPERIMENTAL SETUP

We evaluate ten LHLMs: seven general-purpose LLMs often used as LHLMs through persona-conditioned prompting and three human behavior models fine-tuned on human responses. The former comprise three closed-source frontier models—GPT-5.6 (OpenAI, 2026), Gemini 3.6 Flash (Google DeepMind, 2026), and Opus 5 (Anthropic, 2026)—and four open-weight models: GPT-OSS 120B (OpenAI, 2025), Llama-3.1-70B-Instruct (Grattafiori et al., 2024), Qwen2.5-14B-Instruct (Yang et al., 2024), and Qwen3-8B (Yang et al., 2025). The fine-tuned models are Socrates (Kolluri et al., 2025), OSim-8B (Zhou et al., 2026), and Centaur (Binz et al., 2025). For the prompt templates used to query each LHLM is given in Appendix H.

As the distributional error metric d, we use *total variation distance* (TVD), which has a simpler interpretation for binary outcomes. Because every response is mapped to yes or no, each distribution (P'_{say}, P_{say}, P'_{do}, and P_{do})is characterized by its share of yes responses. For two binary distributions, TVD is the absolute difference between these shares. Denoting the yes rate by p,

$$d(P'_{\text{say}}, P_{\text{say}}) = \left|p'_{\text{say}} - p_{\text{say}}\right| \qquad d(P'_{\text{do}}, P_{\text{do}}) = \left|p'_{\text{do}} - p_{\text{do}}\right|$$

Thus, each error is the difference between predicted and observed yes rates; an error of 13.1 means a 13.1% difference. This scale makes errors comparable across say-do pairs.

4.3 ESTABLISHING BEHAVIORAL UNDERSPECIFICATION

The Definitions 3.1 and 3.2 characterize behavioral underspecification through contrasts averaged over the matched say–do pairs. Prior work has documented substantial variation in LHLM simulation accuracy across domains and settings Kolluri et al. (2025). We therefore use hypothesis tests to determine whether each contrast is systematically supported across say–do pairs and operationalize fidelity asymmetry (FA) and target ambiguity (TA) as binary model-level indicators.

Fidelity Asymmetry. To establish fidelity asymmetry, we test the null hypothesis $H_0^{\text{FA}} : \Delta_{\text{FA}}(M) = 0$ against the alternative hypothesis $H_1^{\text{FA}} : \Delta_{\text{FA}}(M) \neq 0$. We then set,

$$\text{FA}(M) = \begin{cases} \text{Detected}, & \text{if } H_0^{\text{FA}} \text{ is rejected,} \\ \text{Not detected}, & \text{otherwise.} \end{cases}$$

Target Ambiguity. For each behavioral target $t \in \{\text{say}, \text{do}\}$, we test whether its model-implied distribution is closer to its intended human target than to the alternative, $H_{0,t}^{\text{TA}} : \Delta_{\text{TA}}^{t}(M) \geq 0$ vs. $H_{1,t}^{\text{TA}} : \Delta_{\text{TA}}^{t}(M) < 0$. Rejecting $H_{0,t}^{\text{TA}}$ detects target distinction for t. We then set,

$$\text{TA}(M) = \begin{cases} \text{Not detected}, & \text{if target distinction is detected for both say and do,} \\ \text{Detected}, & \text{otherwise.} \end{cases}$$

When target ambiguity is detected, we classify it as *one-sided* if target distinction is detected for exactly one behavioral target and *two-sided* if target distinction is detected for neither.

Type	Model	Distance in points (mean ± SD across datasets) $d(P'_{\text{say}}, P_{\text{say}})\downarrow$	$d(P'_{\text{say}}, P_{\text{do}})\uparrow$	$d(P'_{\text{do}}, P_{\text{do}})\downarrow$	$d(P'_{\text{do}}, P_{\text{say}})\uparrow$	Behavioral underspecification contrasts Δ_{FA}	$\Delta^{\text{say}}_{\text{TA}}$	$\Delta^{\text{do}}_{\text{TA}}$
Frontier closed-source	**GPT-5.6**	13.1 ± 10.5	29.6 ± 20.6	19.9 ± 14.2	22.2 ± 18.6	-6.8^{***}	-16.5^{***}	$-2.3^{\text{n.s.}}$
	Gemini Flash	12.0 ± 10.3	27.9 ± 18.9	16.8 ± 13.3	22.6 ± 19.2	-4.8^{**}	-15.9^{***}	-5.8^{*}
	Opus 5	11.5 ± 11.7	27.7 ± 20.5	18.0 ± 13.1	23.8 ± 20.6	-6.5^{***}	-16.2^{***}	-5.8^{*}
Open-weight	**GPT-OSS 120B**	14.0 ± 12.4	26.0 ± 18.4	18.7 ± 11.8	19.5 ± 16.7	-4.7^{**}	-12.0^{***}	$-0.7^{\text{n.s.}}$
	Llama 70B	14.0 ± 13.7	30.6 ± 20.4	20.4 ± 15.3	19.5 ± 17.4	-6.4^{**}	-16.7^{***}	$+0.8^{\text{n.s.}}$
	Qwen2.5-14B	19.5 ± 17.6	31.6 ± 21.4	21.1 ± 17.1	26.9 ± 22.1	$-1.5^{\text{n.s.}}$	-12.0^{***}	-5.9^{**}
	Qwen3-8B	18.5 ± 15.6	24.3 ± 18.1	17.8 ± 14.3	23.1 ± 19.9	$+0.7^{\text{n.s.}}$	-5.8^{*}	-5.4^{*}
Human-behavior models	**Socrates**	15.2 ± 10.4	19.5 ± 17.8	20.0 ± 14.1	27.6 ± 21.0	-4.8^{**}	$-4.3^{\text{n.s.}}$	-7.6^{**}
	OSim-8B	17.2 ± 12.3	18.8 ± 16.1	16.7 ± 14.6	18.4 ± 14.2	$+0.5^{\text{n.s.}}$	$-1.6^{\text{n.s.}}$	$-1.7^{\text{n.s.}}$
	Centaur	14.0 ± 12.7	25.9 ± 19.9	19.5 ± 14.9	20.9 ± 14.5	-5.5^{**}	-11.9^{***}	$-1.4^{\text{n.s.}}$

Table 1: How close each model comes to simulating what people say and what they actually do. All values given are in percentage points. The final three columns report behavioral underspecification contrasts: Δ_{FA} measures fidelity asymmetry, while $\Delta^{\text{say}}_{\text{TA}}$ and $\Delta^{\text{do}}_{\text{TA}}$ measure target distinction. Negative values indicate greater fidelity to what people say they do for Δ_{FA}, and greater proximity to the intended target for Δ^{t}_{TA}. Superscripts indicate significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; $^{\text{n.s.}}$ indicates no significant effect. See Appendix G for further details.

5 Experimental Results

In this section, we take a deeper look at behavioral underspecification in LHLMs. In particular, we examine the extent to which widely used LHLMs exhibit behavioral underspecification and how its dynamics vary across relevant factors.

5.1 Behavior Underspecification

Table 1 reports the different distributional accuracies in simulating human responses and the contrasts used to characterize behavioral underspecification introduced in Section 3. Moreover, Figure 3 summarizes how models vary along the two dimensions: fidelity asymmetry and target ambiguity. We first examine each dimension in turn, then consider how they jointly characterize behavioral underspecification across models.

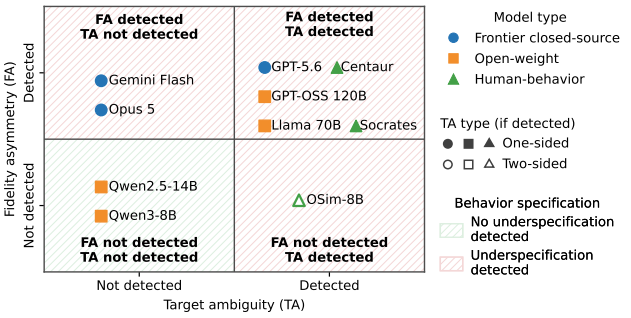


Figure 3: Behavioral underspecification across LHLMs. Models are categorized by detected fidelity asymmetry and target ambiguity.

Fidelity asymmetry is common and consistently favors what people say. Seven of the ten models exhibit significant fidelity asymmetry, with all significant Δ_{FA} values negative, meaning that models are relatively better at predicting what people say than what they actually do. However, contrary to expectations given the potential say-bias of open training data, neither Qwen2.5-14B nor Qwen3-8B exhibits significant fidelity asymmetry. Both, however, have the largest simulation errors, suggesting that poor simulation ability may obscure fidelity asymmetry before a model captures the underlying human-response distributions sufficiently well. This reinforces our remark in Section 3 that behavioral underspecification is distinct from overall simulation accuracy. Overall, we find that when fidelity differs across behavioral targets, it consistently favors what people say over what they do.

Target ambiguity is usually one-sided, and the ambiguous side is usually what people actually do. When predicting what people say, eight of ten models produce distributions detectably closer to the intended say target than to the do target ($\Delta^{\text{say}}_{\text{TA}} < 0, p < 0.05$). In contrast, only five models do so when predicting what people actually do ($\Delta^{\text{do}}_{\text{TA}} < 0, p < 0.05$). Consequently, five models exhibit one-sided target ambiguity. Four (GPT-5.6, GPT-OSS 120B, Llama 70B, and Centaur) distinguish the targets when predicting say but not do, while Socrates exhibits the reverse. Thus, target ambigu-

ity is predominantly one-sided: models generally preserve the say–do distinction for one target, but when that distinction breaks down, it is most often when predicting what people actually do.

Models without target ambiguity can still distinguish one target more strongly than the other. Gemini Flash, Opus 5, Qwen2.5-14B, and Qwen3-8B significantly distinguish what people say and what they do ($p < 0.05$), and therefore do not exhibit target ambiguity. However, this distinction is substantially stronger for what people say than what they do for Gemini Flash (-15.9% vs. -5.8%), Opus 5 (-16.2% vs. -5.8%), and Qwen2.5-14B (-12.0% vs. -5.9%). Only Qwen3-8B distinguishes the two targets with similar strength (-5.8% vs. -5.4%). Thus, even when both targets are distinguishable, models tend to separate what people say more clearly from what they do.

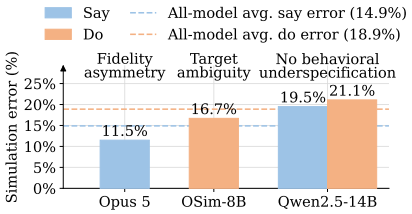


Figure 4: Simulation accuracy vs. behavioral underspecification.

Special cases highlight the value of evaluating behavioral underspecification alongside simulation accuracy. Qwen2.5-14B and Qwen3-8B are the models for which neither form of behavioral underspecification is detected, yet they have among the largest simulation errors (19.5%/21.1% and 18.5%/17.8% for say and do). The converse occurs: Opus 5 has the lowest simulation error for what people say (11.5%) but exhibits significant fidelity asymmetry, while OSim-8B has the lowest simulation error for what people do (16.7%) but exhibits two-sided target ambiguity. This is illustrated in Figure 4. Together, these cases illustrate the point raised in our remark in Section 3: low simulation error does not preclude behavioral underspecification, just as the absence of behavioral underspecification does not imply low simulation error. An evaluation of LHLMs should therefore consider both, which supports the thesis of this work.

Fine-tuned human-behavior models do not consistently reduce behavioral underspecification. Socrates, OSim-8B, and Centaur are fine-tuned human behavior models, whereas the frontier and open-weight models are general-purpose LLMs repurposed as LHLMs through prompting. Yet, all three human-behavior models exhibit target ambiguity: Socrates and Centaur show one-sided ambiguity alongside fidelity asymmetry, while OSim-8B shows two-sided ambiguity. Thus, we find that specialization for human-behavior simulation does not necessarily produce a clearer distinction between what people say and do. A possible explanation is that human-behavior fine-tuning optimizes response prediction without explicitly distinguishing these behavioral constructs; improvements in modeling human responses therefore need not translate into improved separation between say and do. More broadly, depending on the training data, sampling strategy, and fine-tuning objective, fine-tuning may preserve, amplify, or attenuate behavioral underspecification inherited from the pre-trained model, rather than systematically resolving it, as we discuss in Section 5.2.

5.2 Dynamics of Behavioral Underspecification

We next examine the dynamics underlying behavioral underspecification and how fine-tuning changes them. We characterize how LHLMs reproduce or distort the human gap between what people say and do, compare fine-tuned models with their base models, and use the Theory of Planned Behavior Ajzen (1991) to further understand these patterns and potential ways to mitigate them.

5.2.1 How LHLMs Transform the Human Say–Do Gap

Figure 5 provides a distributional view of how LHLMs transform the human say–do gap. Humans exhibit a clear gap between the two behavioral targets: the mean percentage of "yes" responses is 73.2% for say questions and 57.5% for do questions, corresponding to a 15.7% say–do gap. We use this human gap as a reference for examining how LHLMs position and separate the two.

Several distinct dynamics emerge. Many general-purpose LHLMs shift both distributions toward higher "yes" response rates while preserving, and in some cases amplifying, the say–do gap. Other models compress the gap. For example, Qwen3-8B retains a separation between say and do, but with a smaller gap than observed in humans. OSim-8B compresses the gap almost entirely, with its predicted say and do means nearly coinciding, while Socrates primarily shifts the predicted say distribution downward toward the do distribution. Thus, behavioral underspecification can arise through qualitatively different transformations of the human say–do gap—including shifts in its overall level and changes in its magnitude and direction.

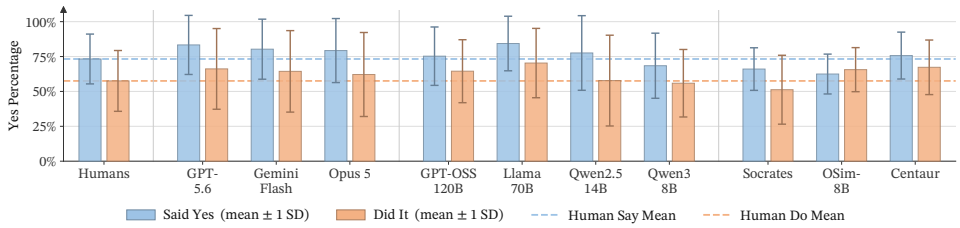


Figure 5: Mean response patterns across say–do pairs. For humans and each LHLM, we show the mean percentage of *yes* responses for say and do questions, with error bars indicating the standard deviation across matched say–do pairs (note that our distributions are binary hence *yes* response percentage is simplification of the actual distribution). The human distributions provide the reference gap between what people say and what they actually do. LHLM distributions show how this gap is preserved, shifted, compressed, or reversed across LHLMs.

5.2.2 The Impact of Fine-Tuning on Human Response Data

Llama 70B is the pre-trained model for Centaur, Qwen2.5-14B for Socrates, and Qwen3-8B for OSim-8B. Across these pairs, fine-tuning generally reduces simulation error ($d(P'_{\text{say}}, P_{\text{say}})$ and $d(P'_{\text{do}}, P_{\text{do}})$ in Table 1), indicating that specialization on human-response data can move predictions closer to human distributions. However, improved simulation accuracy does not necessarily reduce behavioral underspecification. Neither behavioral underspecification form is detected for the two Qwen pre-trained models, yet after fine-tuning, Socrates exhibits one-sided target ambiguity and OSim-8B exhibits two-sided target ambiguity. As shown in Figure 5, Socrates compresses the say–do gap by shifting the say distribution toward the human do distribution, while OSim-8B compresses the gap almost entirely. Fine-tuning also appear to make predicted yes rates more uniform across say–do pairs: their standard deviation decreases in all six comparisons suggesting reduced sensitivity to pair-level behavioral variation (refer Appendix C for more details).

We further find that fine-tuning on human response data can improve simulation accuracy while worsening behavioral underspecification because the two objectives are not equivalent. Fine-tuning may move each target-specific prediction closer to its human distribution without preserving the relationship between the targets. For example, suppose humans respond "yes" 70% of the time for say questions and 50% for do questions, while a pre-trained model predicts 80% and 65%, respectively. Fine-tuning could shift these predictions to 72% and 60%, reducing simulation error from 10% to 2% for say and from 15% to 10% for do. Yet, the model's say–do gap would shrink from 15% to 12%, moving farther from the 20% human gap. Thus, improving each target in isolation can simultaneously make their relative relationship less faithful. Unless the training objective explicitly accounts for the say–do distinction, updates that improve simulation accuracy may still compress, shift, or even reverse the relationship between the two behavioral targets.

The Socrates comparison further suggests that these fine-tuning dynamics reflect a complex interplay of factors, including the relative proportion of say and do training examples. Although approximately 70% of its fine-tuning data are what people say and 30% are what people do (Figure 1b), fine-tuning shifts both its say and do predictions downward. Thus, the direction of the update does not simply follow the dominant behavioral target in the training data. One possibility is that the smaller do subset provides a stronger corrective signal relative to the pre-trained model, while the say data largely reinforces behavior already learned during pretraining. More broadly, differences in response distributions, domains, and behavioral contexts may cause say and do examples to exert different influence during fine-tuning despite their relative prevalence.

Together, these results separate two training goals: improving *human-simulation accuracy* and reducing *behavioral underspecification*. Human-response fine-tuning can improve the former without improving—and sometimes while exacerbating—the latter. Mitigating behavioral underspecification may therefore require training strategies that explicitly account for the distinction between what people say and what they actually do, rather than just increasing exposure to human-response data. Refer to Appendix C for more details.

5.2.3 THEORY OF PLANNED BEHAVIOR

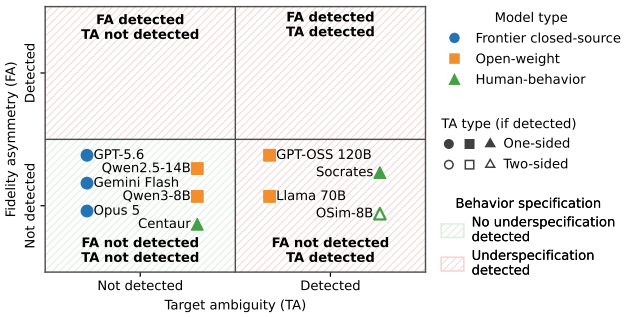


Figure 6: Behavioral underspecification after TPB.

As shown, our results suggest that what people actually do is particularly challenging for LHLMs to represent faithfully. We therefore turn to behavioral science for insights into factors that shape what people do. The Theory of Planned Behavior (TPB) Ajzen (1991) provides a well-studied framework for understanding how values, attitudes, and intentions relate to subsequent actions in humans. It posits that values, attitudes, subjective norms, and perceived behavioral control shape behavioral intentions, which in turn influence realized behavior, with perceived behavioral control directly affecting whether intentions translate into action. This motivates a simple question for LHLMs: rather than asking a model to predict what a person does independently, can explicitly conditioning that prediction on what the same person said improve its representation of realized behavior?

We test this idea using a TPB-inspired conditioning strategy. When simulating what people do, we provide the LHLM with the participant's corresponding say response. As shown in Figure 6, this intervention consistently shifts LHLMs toward less behavioral underspecification. Six of the ten LHLMs now exhibit neither form of behavioral underspecification. Three of the remaining four improve: each moves from exhibiting both fidelity asymmetry and target ambiguity to exhibiting target ambiguity alone, with the exception of OSim-8B, which continues to exhibit target ambiguity. Thus, explicitly linking what people say to what they do substantially reduces behavioral underspecification across models. We provide Detailed results in Appendix F.1. We also find that conditioning in the reverse direction, or asking for both answers in one call, does not yield the same improvement as discussed in Appendices F.2 and F.3.

These results suggest a promising direction for training human-behavior models. Rather than predicting what people do directly, first predict what a person would say they do, then condition their prediction of what the person would do on that predicted say response, while retaining direct prediction for say responses. Such a chain approach would explicitly model the transition TPB introduces.

Remark: This section is not intended to provide a comprehensive solution to behavioral underspecification, but rather to explore potential future directions for addressing it.

6 CONCLUSION

In this work, we identify behavioral underspecification as an overlooked but fundamental challenge for LHLMs. Behavioral underspecification arises when LHLMs, despite being presented as simulators of human behavior, fail to reliably distinguish between what people say and what they actually do. We find that behavioral underspecification is widespread. When fidelity differs between say and do, LHLMs consistently favor what people say. Moreover, when predicting what people actually do, LHLMs have greater difficulty identifying which aspect of human behavior they are intended to reproduce. Crucially, these failures are distinct from an LHLM's overall simulation accuracy: a model can closely reproduce human responses while still exhibiting substantial behavioral underspecification. Evaluations LHLMs based on simulation accuracy alone may therefore provide a misleadingly positive picture of an LHLMs' ability to faithfully simulate human behavior.

We further find that fine-tuning on human-response data does not necessarily resolve behavioral underspecification and can sometimes exacerbate it, suggesting that improving simulation accuracy alone is insufficient; training must also explicitly account for the two constructs. Encouragingly, explicitly linking what people say to what they do substantially reduces underspecification across models, pointing toward fine-tuning strategies that leverage established structures of human behavior. Ultimately, for LHLMs to become credible simulators of humans, we must move beyond asking how accurately they reproduce human responses and establish precisely what aspects of human behavior they actually simulate. This work takes a step toward making that distinction explicit.

AI USE STATEMENT

In this work, we used generative AI tools to identify candidate say–do pairs from open-data repositories and clean and reformat dataset. We have not used generative AI tools for propose or refine hypotheses, design or provide feedback on research methodology or experiments, help develop theoretical models or conceptual frameworks, and interpret results and generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs and support qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for create or edit software code (implementing some parts of the benchmarking), brainstorming, sourcing/searching for information and edit a research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by two authors. All data collections and cleaning done by coding agents were supervised and manually verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses only de-identified, publicly available survey and behavioral datasets, each accessed and used in accordance with its applicable license or data-use policy. Our analyses rely on aggregate rates of what people say and what they do rather than individual-level records, and no new human-subject data were collected for this work.

REPRODUCIBILITY STATEMENT

The data and evaluation procedures are described in Section 4 and the appendices. The say–do pair suite, including dataset-specific processing decisions, is documented in Appendix B together with the pair-quality checks of Appendix B.5. We report all models and prompting procedures used in our experiments, and use fixed random seeds for all resampling and permutation tests to ensure that the reported tables and figures can be reproduced.

REFERENCES

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. Consistently simulating human personas with multi-turn reinforcement learning. *Advances in Neural Information Processing Systems*, 38, 2025.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.

Icek Ajzen. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.

Icek Ajzen and Martin Fishbein. Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological bulletin*, 84(5):888–918, 1977.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*, 2025.

Anthropic. Claude Opus 5. Claude platform model documentation, 2026. URL `https://platform.claude.com/docs/en/models/opus-5/overview`. Model ID `claude-opus-5`, called here through the Amazon Bedrock inference profile `us.anthropic.claude-opus-5`. Accessed 2026-09-01.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Large language models can predict the results of social science experiments. *Nature*, 656(8126):115–122, 2026. doi: 10.1038/s41586-026-10742-x.

Christopher A. Bail. Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024. doi: 10.1073/pnas.2314021121.

Krisztian Balog and ChengXiang Zhai. User simulation for evaluating information access systems. *Foundations and Trends in Information Retrieval*, 18(1–2):1–261, 2024. doi: 10.1561/1500000098.

Anat Bardi and Shalom H. Schwartz. Values and behavior: Strength and structure of relations. *Personality and Social Psychology Bulletin*, 29(10):1207–1220, 2003. doi: 10.1177/0146167203254602.

Daryl J. Bem. Self-perception theory. volume 6 of *Advances in Experimental Social Psychology*, pp. 1–62. Academic Press, 1972. doi: https://doi.org/10.1016/S0065-2601(08)60024-6. URL https://www.sciencedirect.com/science/article/pii/S0065260108600246.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pp. 3141–3154, 2025. doi: 10.18653/v1/2025.naacl-long.162.

Vanessa Cheung, Maximilian Maier, and Falk Lieder. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122(25):e2412015122, 2025. doi: 10.1073/pnas.2412015122.

Robert B. Cialdini, Carl A. Kallgren, and Raymond R. Reno. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in Experimental Social Psychology*, volume 24, pp. 201–234. Academic Press, 1991. doi: 10.1016/S0065-2601(08)60330-5.

John P. Dickerson, Hadi Hosseini, Samarth Khanna, and Leona Pierce. Who gets the kidney? Human-AI alignment, indecision, and moral values. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26)*, pp. 8128–8147. ACM, 2026. doi: 10.1145/3805689.3806437.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023. doi: 10.1016/j.tics.2023.04.008.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. doi: 10.52202/079017-1458.

Michael D. Ekstrand and Martijn C. Willemsen. Behaviorism is not enough: Better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, pp. 221–224, 2016. doi: 10.1145/2959100.2959179.

Gerhard Fischer. User modeling in human–computer interaction. *User Modeling and User-Adapted Interaction*, 11(1–2):65–86, 2001. doi: 10.1023/A:1011145532042.

Martin Fishbein and Icek Ajzen. *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Addison-Wesley, Reading, MA, 1975.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1259, 2024. doi: 10.1057/s41599-024-03611-3.

Google DeepMind. Gemini 3.6 Flash. Gemini API model documentation, 2026. URL `https://ai.google.dev/gemini-api/docs/models`. Model ID `gemini-3.6-flash`. Accessed 2026-09-01.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Igor Grossmann, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. AI and the transformation of social science research. *Science*, 380 (6650):1108–1109, 2023. doi: 10.1126/science.adi1778.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic HCI research data: A case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2023. doi: 10.1145/3544548.3580688.

John J Horton, Apostolos Filippas, and Benjamin S Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Hadi Hosseini and Samarth Khanna. Distributive fairness in large language models: Evaluating alignment with human values. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Hadi Hosseini, Samarth Khanna, and Leona Pierce. The judgment-consequence gap: LLM moral reasoning in healthcare decisions. *arXiv preprint arXiv:2608.05583*, 2026.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. Simbench: Benchmarking the ability of large language models to simulate human behaviors. In *International Conference on Learning Representations*, volume 2026, pp. 28290–28341, 2026.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5906–5919, 2023. doi: 10.18653/v1/2023.findings-emnlp.393.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. MoralBench: Moral evaluation of LLMs. *ACM SIGKDD Explorations Newsletter*, 27(1):62–71, 2025. doi: 10.1145/3748239.3748246.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392, 2024. doi: 10.1038/s42256-024-00820-y.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *Management Science*, 2023. doi: 10.1287/mnsc.2022.03683. Articles in Advance.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Manish Raghavan. The inversion problem: Why algorithms should infer mental state and not just predict behavior. *Perspectives on Psychological Science*, 19(5):827–838, 2024. doi: 10.1177/17456916231212138.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S Bernstein. Finetuning llms for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30084–30099, 2025.

Austin C. Kozlowski and James Evans. Simulating subjects: The promise and peril of artificial intelligence stand-ins for social agents and interactions. *Sociological Methods & Research*, 54 (3):1017–1073, 2025. doi: 10.1177/00491241251337316.

Stephen J Kraus. Attitudes and the prediction of behavior: A meta-analysis of the empirical literature. *Personality and social psychology bulletin*, 21(1):58–75, 1995.

Shiro Kuriwaki. Cumulative CES Common Content, 2026. URL https://doi.org/10.7910/DVN/II2DB6.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pp. 24–49, 2025. doi: 10.18653/v1/2025.naacl-long.2.

Smitha Milli, Luca Belli, and Moritz Hardt. From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 714–722, 2021. doi: 10.1145/3442188.3445933.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. From individual to society: A survey on social simulation driven by large language model-based agents. *ACM Computing Surveys*, 58(11):1–41, 2026. doi: 10.1145/3800683.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. Flipping the dialogue: Training and evaluating user language models. *arXiv preprint arXiv:2510.06552*, 2025.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

OpenAI. GPT-5.6 Sol. OpenAI API model documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. Model ID gpt-5.6-sol. Accessed 2026-09-01.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1–18, 2022. doi: 10.1145/3526113.3545616.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024.

Elaine Rich. User modeling via stereotypes. *Cognitive Science*, 3(4):329–354, 1979. doi: 10.1207/s15516709cog0304_3.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7370–7392, 2024. doi: 10.18653/v1/2024.acl-long.399.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The Knowledge Engineering Review*, 21(2):97–126, 2006. doi: 10.1017/S0269888906000944.

Shalom H. Schwartz. Normative influences on altruism. In *Advances in Experimental Social Psychology*, volume 10, pp. 221–279. Academic Press, 1977. doi: 10.1016/S0065-2601(08)60358-5.

Shalom H. Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in Experimental Social Psychology*, volume 25, pp. 1–65. Academic Press, 1992. doi: 10.1016/S0065-2601(08)60281-6.

Omar Shaikh, Shardul Sapkota, Shan Rizvi, Eric Horvitz, Joon Sung Park, Diyi Yang, and Michael S Bernstein. Creating general user models from computer use. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–23, 2025.

Paschal Sheeran. Intention—behavior relations: a conceptual and empirical review. *European review of social psychology*, 12(1):1–36, 2002.

Paschal Sheeran and Thomas L. Webb. The intention–behavior gap. *Social and Personality Psychology Compass*, 10(9):503–518, 2016. doi: 10.1111/spc3.12265.

Nikita Soni, H Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. Large human language models: A need and the challenges. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8631–8646, 2024.

Jan E. Stets and Peter J. Burke. Identity theory and social identity theory. *Social Psychology Quarterly*, 63(3):224–237, 2000. doi: 10.2307/2695870.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 21147–21170, 2025. doi: 10.18653/v1/2025.acl-long.1028.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025a. doi: 10.1038/s42256-025-00986-z.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21082–21107, 2025b.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37, 2025c. doi: 10.1145/3708985.

Lorraine Whitmarsh and Saffron O'Neill. Green identity, green living? the role of pro-environmental self-identity in determining consistency across diverse pro-environmental behaviours. *Journal of Environmental Psychology*, 30(3):305–314, 2010. doi: 10.1016/j.jenvp.2010.01.003.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behavior? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. doi: 10.52202/079017-0501.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1807–1817, 2024. doi: 10.1145/3626772.3657844.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2204–2213, 2018. doi: 10.18653/v1/P18-1205.

Xuhui Zhou, Weiwei Sun, Weihua Du, Jiarui Liu, Haojia Sun, Qianou Ma, Tongshuang Wu, Yiming Yang, and Maarten Sap. OdysSim: Building foundation models for human behavior simulation. *arXiv preprint arXiv:2606.14199*, 2026.

Appendix

A Detailed Related Work **17**

B The Say–Do Dataset: Pair Annotations and Quality Checks **18**

B.1 Gap Categories 18

B.2 Temporal Design and Action Measurement 19

B.3 Domains 19

B.4 Joint Distribution 20

B.5 Pair Quality Checks 20

C Fine-Tuning on Human Behavior: Further Details **21**

D Behavioral Underspecification Is Not Localized **23**

D.1 Demographic groups 24

D.2 Domains 24

D.3 Properties of the behavior 24

E Demographic Differences Cancel Across Pairs **27**

F Detailed Results under Different Prompting Paradigms **31**

F.1 The Forward Paradigm, the Do Prediction Given the Say Answer 32

F.2 The Reverse Paradigm, the Say Prediction Given the Do Answer 32

F.3 The Joint Paradigm, Both Answers in One Call 33

G Details of Statistical Tests **34**

H Prompt Construction **36**

H.1 What varies across pairs, and what does not 37

H.2 Respondent, simulating the individual 38

H.3 Forward, statement given, action produced 40

H.4 Reverse, action given, statement produced 41

H.5 Joint, both sides in one generation 43

H.6 Response options, what the model is actually offered 45

H.7 Multi-item behavior batteries, asked act by act in one call 48

H.8 Measurement separation, ages, and per-occasion personas 50

H.9 Querying each model in its native format 51

A Detailed Related Work

LHLMs as Human Simulators. Recent work has explored LLMs as synthetic participants in social and behavioral research (Dillion et al., 2023; Grossmann et al., 2023; Bail, 2024; Gao et al., 2024; Mou et al., 2026; Kozlowski & Evans, 2025). Argyle et al. (2023) introduced *silicon sampling*, showing that demographic conditioning can reproduce some subgroup patterns in U.S. public-opinion surveys. Furthermore, related studies have used LLMs to simulate economic agents (Horton et al., 2023), replicate classic experiments in economics, psycholinguistics, social psychology (Aher et al., 2023), and cognitive psychology (Binz & Schulz, 2023), simulate trust behavior in economic games (Xie et al., 2024), generate synthetic HCI research data (Hämäläinen et al., 2023), and predict treatment effects in social-science experiments (Ashokkumar et al., 2026). LLM-based agent architectures further extend this approach to populated social systems (Park et al., 2022) and to temporally evolving and individualized simulations (Park et al., 2023; 2024).

Another line of work improves behavioral simulation by training directly on human-response data. Fine-tuning on large survey collections improves the prediction of opinion distributions across subpopulations (Suh et al., 2025) and countries (Cao et al., 2025), and Meister et al. (2025) benchmark how well such distributions are recovered. SocSci210 aggregates responses from 210 social-science experiments and fine-tunes an LLM to predict response distributions in unseen studies (Kolluri et al., 2025). Similarly, Centaur is fine-tuned on trial-level choices from cognitive experiments and generalizes across participants, task variations, and experimental domains (Binz et al., 2025). Most recently, OdysSim introduced a large-scale behavioral foundation model Osim and performed an investigation of behavioral foundation models (Zhou et al., 2026). Whereas these studies primarily establish proof of concept that LLMs and fine tuned models can approximate human responses, our work investigates this capability in depth, identifying when and why such simulations fail to distinguish what people say from what they do.

Limits of LHLMs. Recently, a growing line of work has examined the limits of behavioral fidelity in LHLMs as well. Prior studies show that these models may poorly represent certain demographic groups (Santurkar et al., 2023), underestimate the diversity of human responses (Wang et al., 2025a), and exhibit sensitivity to prompt wording and model updates (Bisbee et al., 2024). They may also fail to reproduce well-established human response biases in surveys (Tjuatja et al., 2024), and their survey answers can be driven by ordering and labeling artifacts rather than by the population being simulated (Dominguez-Olmedo et al., 2024). They are also often misaligned with humans on normative principles (Ji et al., 2025). Their allocations depart from human preferences over distributive fairness (Hosseini & Khanna, 2025), their moral decisions amplify cognitive biases such as omission bias (Cheung et al., 2025), and when allocating scarce medical resources they deviate from the attributes humans prioritize, rarely express indecision, and do not let judgments of patient responsibility shape the allocation as humans do (Dickerson et al., 2026; Hosseini et al., 2026). Our work identifies a complementary and previously underexplored failure mode, in which an LHLM may generate plausible human responses while modeling the wrong behavioral construct. Using matched say–do pairs, we evaluate whether LHLMs accurately recover the distributions of what people say vs what they do, and whether they preserve the empirical gap between them.

Personalization and User Modeling. Modeling individual users has long been central to adaptive systems (Rich, 1979; Fischer, 2001), and simulated users are a standard tool for training and evaluating dialogue and information-access systems (Schatzmann et al., 2006; Balog & Zhai, 2024). LLMs have renewed both threads, through persona-grounded dialogue (Zhang et al., 2018), benchmarks and methods for personalized generation and alignment (Salemi et al., 2024; Hwang et al., 2023; Kirk et al., 2024), and LLM-based user simulators for recommendation (Zhang et al., 2024; Wang et al., 2025c) and conversational assistants (Naous et al., 2025; Wang et al., 2025b; Abdulhai et al., 2025). The recommender-systems literature has also recognized that what users do can diverge from what they say they want. Behavioral signals such as clicks and engagement can misrepresent stated preferences and long-term value (Ekstrand & Willemsen, 2016; Milli et al., 2021; Kleinberg et al., 2023), motivating calls to infer mental states rather than just predict behavior (Kleinberg et al., 2024). Yet LLM-based user models are rarely evaluated on which of the two they reproduce. Our work brings this distinction to LHLMs and shows that it is often underspecified.

B THE SAY–DO DATASET: PAIR ANNOTATIONS AND QUALITY CHECKS

Our dataset consists of 98 say–do pairs. Each pair joins one survey item recording what respondents said to one recording what the same respondents did, and both answers are converted to yes or no. The details of each say-do pair and how the answers were binarized are provided as supplementary material. This appendix first describes the pairs along three dimensions, each coded independently of the others. *Gap category* records what the "say" expresses. *Temporal design* records when the action was observed relative to the statement. *Domain* records which area of life the action belongs to. Appendix D uses these dimensions to ask whether behavioral underspecification is confined to particular kinds of pairs. Table 2 gives the share of pairs at every level. The appendix then sets out the checks a candidate pair had to pass before it entered the dataset (Appendix B.5).

B.1 GAP CATEGORIES

Gap category specifies *what* is expressed on the "say" side. We distinguish six forms, ordered here from the most proximal to the action to the most distal.

Intention–Action. The "say" expresses a concrete behavioral intention, plan, willingness, or commitment toward a specific future action (e.g., "I intend to vote in the next election," "I plan to reduce food waste"). Under the Theory of Planned Behavior, intentions are considered the most proximal cognitive predictor of action, making intention–action gaps the canonical form studied in the behavioral literature (Ajzen, 1991; Sheeran, 2002).

Attitude–Action. The "say" expresses a general evaluation, value, concern, or interest rather than a commitment to a specific future action (e.g., "I am concerned about drought," "I am interested in politics"). Because attitudes are more distal determinants of actions than intentions, attitude–action relationships are typically weaker and more heterogeneous (Ajzen & Fishbein, 1977; Kraus, 1995).

Value–Action. The "say" rates how *important* something is to the respondent, either in their own life or as part of what a role or identity means to them (e.g., "How important is religion in your life?", "How important is attending religious services to what being a religious person means to you?"). Values are trans-situational priorities rather than evaluations of a single object, so they constrain behavior only weakly and only when the situation makes the value relevant (Schwartz, 1992; Bardi & Schwartz, 2003).

Norm–Action. The "say" is prescriptive and its target is other people rather than the respondent, in that it rates how people in some group ought to behave (e.g., "Thinking about gun owners who have children in their home, how important do you think it is for them to keep all of their guns in a locked place?", "Thinking about workers in general, how important do you think artificial-intelligence skills are for a worker to be successful?"). Because the statement is about a third party, the respondent can endorse the norm at essentially no personal cost, which is precisely the asymmetry that focus theory and the norm-activation literature predict (Cialdini et al., 1991; Schwartz, 1977). These are the only items in the dataset whose grammatical subject is not the respondent.

Belief–Action. The "say" is a factual or perceived-state judgement about the world, an outcome, or another person, with no evaluative or prescriptive content (e.g., "How much of a threat, if any, is the coronavirus for your personal health?", "How much do you think global climate change is currently affecting your local community?", "How important is getting enough physical exercise for preventing serious disease?"). In the expectancy-value tradition, beliefs sit behind attitudes rather than beside them, because they supply the premises from which an evaluation is formed (Fishbein & Ajzen, 1975). Subjective expectations about the respondent's own future, such as the percent chance of completing a degree, are coded here as well, since they report a probability judgement rather than a commitment.

Identity–Action. The "say" is a self-description, in which the respondent rates how well a label or category applies to them (e.g., "How well, if at all, does the phrase 'a supporter of the environment' describe you?", "Would you describe yourself as particularly concerned about helping the

environment as you go about your daily life?"). Identity is theorized to drive behavior through self-verification rather than through evaluation, and self-identity can predict behavior over and above attitudes and intentions (Stets & Burke, 2000; Whitmarsh & O'Neill, 2010).

Assignment rule. Categories are assigned by a single ordered test applied to the "say" item, so that every pair lands in exactly one category. Reading the item stem, we apply the following tests in order and stop at the first one that holds.

1. If the stem commits the respondent to a specific future action, the pair is **Intention–Action**.
2. Otherwise, if its subject is other people's conduct, the pair is **Norm–Action**.
3. Otherwise, if it asks the respondent to describe themselves with a label, the pair is **Identity–Action**.
4. Otherwise, if it asks for a judgement about a state of the world with no evaluative content, the pair is **Belief–Action**.
5. Otherwise, if it asks how important something is, the pair is **Value–Action**.
6. Otherwise the pair is **Attitude–Action**.

Attitude is therefore the residual category, and it remains the largest, at 36.7% of all say–do pairs. Value–Action, Identity–Action, and Belief–Action each cover between 13% and 16%, while Intention–Action covers 12.2% and Norm–Action 8.2%. That the canonical intention–action case is a minority of the dataset is deliberate, because restricting the benchmark to it would have excluded most of the available public data.

B.2 Temporal Design and Action Measurement

Temporal design specifies *when* the action was observed relative to the statement. We distinguish two designs.

Prospective. The "say" is collected before the target action occurs. Respondents express an intention or attitude at time T_1, and behavior is measured at a later time T_2. Because the reported statement cannot be influenced by knowledge of the eventual outcome, prospective designs are generally regarded as the methodological gold standard for studying say–do gaps (Sheeran & Webb, 2016).

Concurrent. The "say" and "do" are measured within the same survey wave. Although this design is considerably more common in social-science surveys, it is more susceptible to shared-method variance, including social-desirability bias and recall error that may jointly affect both measurements. It also does not fix the order of the two measurements. In a small number of same-session designs the action was recorded first and the statement elicited afterwards, so the statement can be a post-hoc account of a choice the respondent had already made. We do not give these cases a level of their own, because they raise the same shared-method concern as the rest of the concurrent pairs.

Action measurement. We separately record how the action itself was measured, distinguishing external records, respondents' own reports, and incentive-compatible tasks, but we do not treat it as a dimension of its own, because only 6.1% of all say–do pairs use an action measure that is not the respondent's own report. That figure is a property of what is publicly available rather than a design choice, since external measures require record linkage and incentivized tasks require a laboratory. Appendix D therefore splits the pairs by temporal design and action measurement together.

B.3 Domains

Domain records the area of life the behavior belongs to. We use seven groups, which partition the dataset, and they are collapsed from the finer subject labels carried by the source surveys so that every group is large enough to stratify on.

Table 2: Joint distribution of the taxonomy, as a percentage of all say–do pairs. Rows are the seven domains. The left block crosses domain with gap category and the right block with temporal design, and this is the only place the temporal breakdown within a domain is reported. One pair is 1.0%, and cells are rounded independently, so rows and columns can differ from their totals by 0.1.

	Gap category							Temporal design	
Domain	Att.	Val.	Iden.	Bel.	Int.	Norm	**All**	Prosp.	Conc.
Environment and Sustainability	18.4	1.0	7.1	2.0	2.0	–	30.6	4.1	26.5
Religion and Spirituality	–	13.3	4.1	2.0	–	–	19.4	6.1	13.3
Health and Wellness	4.1	–	2.0	5.1	7.1	–	18.4	4.1	14.3
Privacy and Tech Adoption	12.2	–	–	1.0	–	1.0	14.3	–	14.3
Social Behavior	–	–	1.0	1.0	–	7.1	9.2	–	9.2
Civic Participation	2.0	1.0	–	–	2.0	–	5.1	2.0	3.1
Education and Work	–	–	–	2.0	1.0	–	3.1	2.0	1.0
All	36.7	15.3	14.3	13.3	12.2	8.2	100	18.4	81.6

Environment and Sustainability. Climate, conservation, sustainable consumption, and agriculture, covering recycling and energy use, food waste, biodiversity and forest protection, organic certification, and talking to others about climate change. This is the largest group, at 30.6% of all say–do pairs.

Religion and Spirituality. Religious and supernatural belief and observance, covering attendance at services in person and online, private prayer, activity in a congregation, and belief in astrology or reincarnation.

Health and Wellness. Personal health behavior, covering COVID-19 vaccination, boosters, masking and home testing, physical exercise, smoking cessation, and organ-donor registration.

Privacy and Tech Adoption. Behavior toward technology and personal data, covering reading privacy policies, adjusting account settings, clearing cookies, using encrypted messaging, and adopting artificial-intelligence tools at work.

Social Behavior. Conduct toward other people where the action has a direct social or ethical cost, covering honesty in incentivized tasks, trust in strangers, firearm-safety practices, and participation in community organizations.

Civic Participation. Political and civic action, covering voting, political activity on social media, attending demonstrations, and displaying the flag.

Education and Work. Educational attainment and workplace behavior, covering completing a degree, taking up job training, and working from home. This is the smallest group, at 3.1% of all say–do pairs.

B.4 Joint Distribution

Table 2 crosses domain with each of the other two dimensions. Although coverage is uneven, with the environment, religion, and health groups together holding 68.4% of all say–do pairs, the dataset covers a large variety of domains and types of say-do pairs. The three dimensions also overlap. Most Value pairs concern religious observance, most Norm pairs concern social behavior, and no privacy or social-behavior pair is prospective.

B.5 Pair Quality Checks

A stated item and a behavioral item drawn from the same survey do not by themselves make a usable say–do pair. Because a pair's gap is a disagreement between the two binarized answers, and *both* directions of disagreement count, a pair is admissible only when each direction would be a

contradiction. Saying yes and not acting has to be an inconsistency, and so does saying no and acting. Where only one of the two is, the disagreement in the other direction is not a gap at all but a circumstance, and averaging the two would report a quantity that is part inconsistency and part noise. Every pair in the dataset is scored in both directions, and no pair uses a one-sided rule.

Screening for that property is stricter than asking whether the two items are about the same topic, and it removed many candidate pairs. Each candidate was reviewed one pair at a time, against the verbatim wording of both items and the cross-tabulation of the two binarized answers, and the decision and its reason were recorded per pair. In the largest source we drew from, these checks rejected about 41% of the pairs that had already passed a topical screen. A candidate had to satisfy the following, of which the first six are requirements and the last is a flag we recorded rather than a cut we applied.

1. The stated side asks the respondent for an attitude, belief, value, priority, normative judgement, or intention. An item that reports the respondent's own past or habitual behavior is a behavioral item however value-laden it sounds.
2. The behavioral side asks factually about the respondent's own conduct, meaning an act performed, a frequency of acts, a membership, or a possession that exists only because of an act. A status the respondent did not enact does not qualify.
3. The behavior is a plausible enactment of the stated position, about the same object, and at least partly under the respondent's control.
4. The statement names an act and a direction. Items that report only how much a respondent cares, an emotion, or a perception are excluded, because any behavior is compatible with them.
5. The act is not gated on something the respondent may not have, such as an intervening purchase, a capital outlay, a home, an employer programme, or a nearby opportunity. Otherwise not acting can be circumstance rather than inconsistency.
6. The statement is about the respondent, or else about a group the respondent belongs to and that the data lets us restrict the sample to. A judgement about what governments, companies, or people in general ought to do is excluded on its own, because the respondent's own inaction contradicts nothing, and this is why the norm items in the dataset are paired only where the relevant subgroup is recoverable.

The pairs that survive are ones where a reader can say what a discordant respondent is being inconsistent about. This screen applies to pairs. It is separate from the rule of Section 4 that applies to respondents, which drops a respondent whose answer on either side is neutral or ambiguous.

C Fine-Tuning on Human Behavior: Further Details

Three models in our study are fine-tuned on the responses of individual people, and each can be compared with the model it was trained from. Centaur is compared with Llama 70B,[2] Socrates with Qwen2.5-14B, and OSim-8B with Qwen3-8B. Section 5.2 reports that fine-tuning does not remove behavioral underspecification and that it compresses the gap between the predicted say and do rates. This appendix gives the details. We first compare the verdicts and the simulation errors of each base with those of its fine-tune, and then show what the fine-tuning does to the predictions that produces both results. Every comparison uses the 98 say–do pairs and the independent design of Table 1.

Neither form of behavioral underspecification is removed by fine-tuning. No fine-tune moves from exhibiting a form to not exhibiting it. Centaur keeps the fidelity asymmetry of its base (Δ_{FA} goes from -6.4 to -5.5 points) and the same one-sided target ambiguity, with the do target the ambiguous one in both. Socrates acquires a fidelity asymmetry its base did not show (-1.5, not significant, to -4.8, $q < 0.01$), and the target distinction its base showed on both sides becomes one-sided target ambiguity, with the say target the ambiguous one. OSim-8B, like its base, shows no fidelity asymmetry ($+0.7$ and $+0.5$, both not significant), but the target distinction its base showed on both sides becomes two-sided target ambiguity.

[2] Centaur is an adapter on the Llama-3.1-70B base checkpoint, which we did not run as a standalone model, so the base it is compared with here is the instruction-tuned release of the same model.

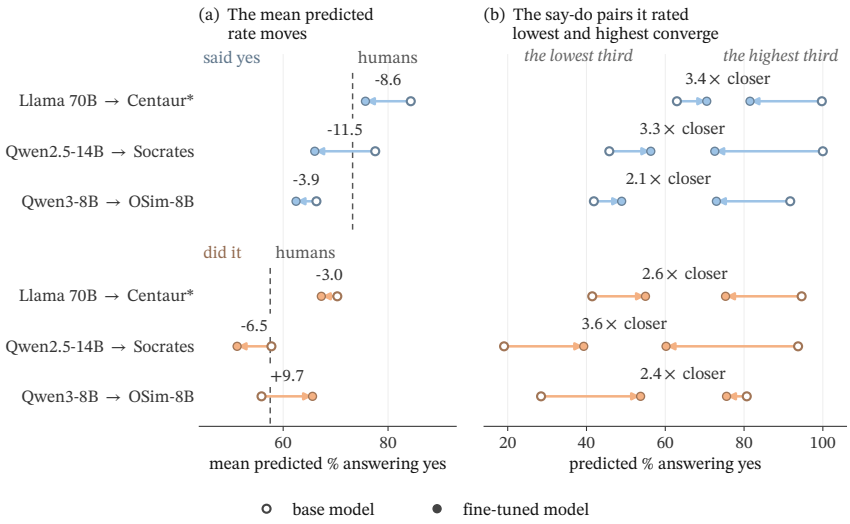


Figure 7: **What fine-tuning on human behavior does to a base model.** Each row is one base model paired with the model fine-tuned from it, split by question side, over the same say-do pairs as Table 1. Open circles are the base model and filled circles the fine-tune, so every arrow runs from base to fine-tune. (a) The mean predicted yes rate, against the human rate for that side (dashed). (b) The same rate computed separately over the third of say-do pairs on which the base model gave its lowest yes rates and the third on which it gave its highest, with the middle third left out. The label gives how many times closer together the two groups end up, greyed and marked n.s. where a paired bootstrap over say-do pairs does not establish the change. Centaur is asterisked because its adapter targets a checkpoint not collected here, so the base shown for it is the instruction-tuned release of the same model.

Simulation error changes little. The simulation error $d(P'_{\text{say}}, P_{\text{say}})$ or $d(P'_{\text{do}}, P_{\text{do}})$ falls in five of the six comparisons (three fine-tunes, two sides each), by 0.9 to 4.4 points, and is unchanged for Centaur on the say side. None of the six changes is significant (all $q > 0.2$). The small and uncertain gain in accuracy therefore comes with no gain in the separation of the two targets.

Fine-tuning levels the predictions across say–do pairs. Within a pair, the fine-tunes vary their answers across respondents more than any other model. Their predicted yes rate is near-unanimous, above 95% or below 5%, on 2% to 8% of pairs, against 14% to 35% for the other models. Across pairs, however, their predictions become more alike. Order the pairs by the yes rate the base model predicts for each. The fine-tune raises the lowest third of that ordering and lowers the highest third, leaving the two 2.1 to 3.6 times closer together (Figure 7b), and a bootstrap over pairs establishes the change in all six comparisons. Ordering the pairs by the human yes rate instead, which does not depend on any model, gives the same narrowing, by 1.9 to 2.7 times, although the bootstrap does not establish it for the say side of Socrates or of OSim-8B. Five of the six mean rates fall, by 3.0 to 11.5 points, because the highest third falls further than the lowest third rises. The do side of OSim-8B rises by 9.7 points, because there the larger move is at the other end (Figure 7a).

The levelled predictions follow the human rate less closely. A prediction about people should rise and fall with the human yes rate from one pair to the next. We measure this by the slope of the predicted rate against the human rate across the pairs, which is 1 when the prediction moves point for point with the human rate and 0 when it ignores which pair is asked about. Every fine-tune is flatter than its base, falling from 0.50–0.83 to 0.30–0.55, against 0.83–0.96 for the frontier models. The fall is established in four of the six comparisons, all but the say side of Socrates and of OSim-8B (Figure 8).

The say and do rates move closer together. The same levelling applies between the two sides. Each fine-tune's mean predicted say and do rates sit closer together than its base's. The say–do gap narrows from 14.0 to 8.4 points for Centaur and from 19.8 to 14.8 for Socrates, and for OSim-8B

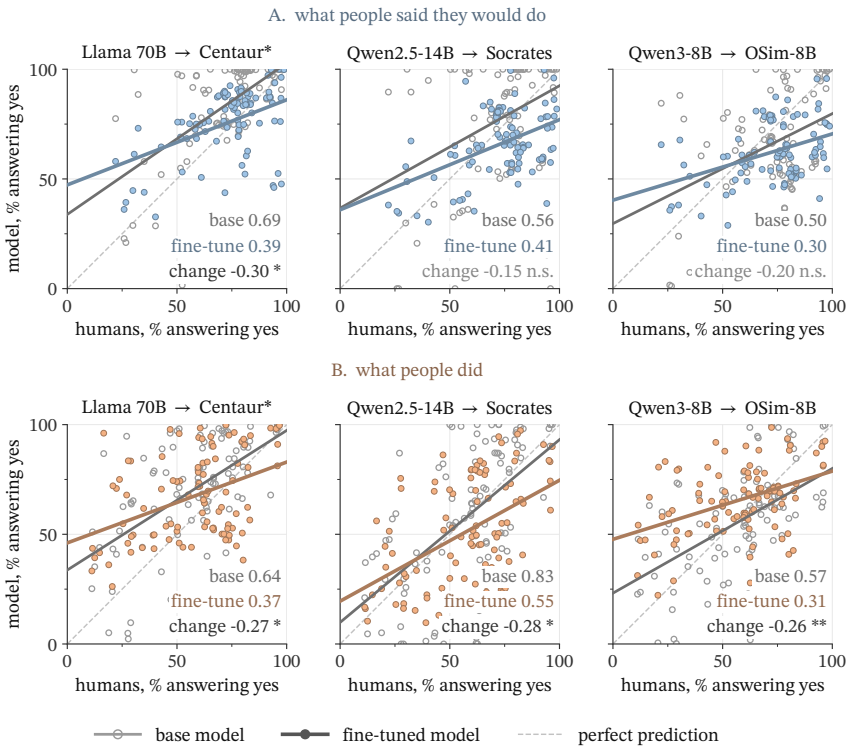


Figure 8: **How closely each model follows the people it is asked about.** One panel per base and fine-tune pair per question, with the share of people who answered yes on the horizontal axis and the share the model predicts on the vertical, one point per say-do pair. Grey open circles are the base model and filled circles the model fine-tuned from it, coloured by question. The lines are least squares fits, and the slope printed for each is the rate at which it follows the human rate, where 1 is one-for-one tracking and 0 is a prediction that ignores which say-do pair is being asked about. The third number is the change, marked n.s. and set in grey where a bootstrap over say-do pairs does not establish it. The dashed diagonal is perfect prediction. Centaur is asterisked because its adapter targets a Llama-3.1-70B checkpoint not collected here, so the base shown for it is the instruction-tuned release of the same model.

it goes from 10.5 to -3.1, a do rate slightly above the say rate. The closer the two predicted rates, the more similar the distance from each prediction to the two human targets, which is how target ambiguity arises.

D Behavioral Underspecification Is Not Localized

Section 5.1 detects fidelity asymmetry and target ambiguity over all 98 say–do pairs. This section asks whether either form is carried by one part of the data, a group of people, a topic, or a kind of behavior, rather than being spread across it. For each subset we re-run the tests of Section 4.3 on that subset alone, for all ten models, with the correction across the ten models applied within the subset. Demographic groups and domains are drawn as the quadrant charts of Figure 3, one chart per subset. The behavioral properties have too many categories for that, so their figure gives the quadrant of each model in each category as one cell of a grid. If the underspecification were localized, it would be present in some subsets and absent from the rest. Instead, every subset we can test shows it for at least six of the ten models.

One caution applies to every figure here. Target ambiguity is detected when target distinction is not, so a subset with fewer pairs, and so less power to establish distinction, drifts toward target ambiguity and away from detected fidelity asymmetry. Across the domains and behavioral categories, fidelity asymmetry is detected in 39% of model–subset cells with 30 or more pairs and in 8% of those with

fewer, and two-sided target ambiguity rises from 15% to 38%. We therefore read a detected form in a small subset as evidence and an undetected one as uninformative. A subset of five or fewer pairs, on which the tests cannot reach $q < 0.05$ at all, is greyed out and not classified.

D.1 Demographic groups

Figures 9 and 10 split the respondents of every pair into two groups on one attribute at a time and compare each group's predicted yes rates with its own observed ones. For age, education and income the groups are the bottom and top thirds within each source, and for gender they are men and women. Politics is split into the left and right thirds of an ideology scale, or into Democrats and Republicans where a source records only party, and race into White and Black respondents, the two groups recorded in the most sources. An attribute uses only the pairs whose source records both of its groups, 81 to 96 of them, so the two panels of a row cover the same pairs.

Every group shows behavioral underspecification for at least eight of the ten models. Fidelity asymmetry is detected for the same seven models as over all pairs in eight of the twelve groups. Among younger, more educated and left-leaning respondents the set gains or swaps one or two models, and among right-leaning respondents it is detected for four (GPT-5.6, Gemini Flash, Opus 5 and Centaur). Target ambiguity differs more between the two groups of an attribute, mainly for Gemini Flash and Opus 5, which distinguish both targets over all pairs. Both show target ambiguity among younger, more educated, male, higher-income, left-leaning and Black respondents and distinguish both targets in the other group of each attribute, except that Gemini Flash also shows target ambiguity among right-leaning respondents. Neither form is therefore confined to a group of people. Groups can still differ inside a single pair. Appendix E shows that they do, but that the differences point opposite ways across pairs and cancel on average.

D.2 Domains

Figure 11 splits the pairs by topic. Civic behavior (5 pairs) and education and work (3) are too small to test and are not drawn. Target ambiguity is detected for every model in all five remaining domains, and in health it is two-sided for every model. Fidelity asymmetry is detected in environment (30 pairs) for five models and in privacy and technology (14) for four, and for no model in religion, health or social behavior (9 to 19 pairs). Given the caution above, we do not read the latter as an absence of fidelity asymmetry. The topic of a pair therefore does not decide whether its predictions are underspecified.

D.3 Properties of the behavior

We coded each say-do pair on nine properties of the behavior and its measurement, listed with their categories in Table 3. Seven of them describe the behavior itself, and the other two, the gap category and the relative timing, are the taxonomy axes defined in Appendix B.1 and Appendix B.2. We chose these because the behavioral-science account of the say-do gap makes each one a plausible reason for doing to fall short of saying, so if the underspecification were localized we would expect it to track at least one of them. It largely does not.

Figure 12 re-runs both tests in every category of the nine properties, 33 of which hold enough pairs to test. Target ambiguity is detected in 314 of the 330 model–category cells, and every category shows behavioral underspecification for at least six of the ten models. Fidelity asymmetry is again detected mainly in the larger categories. Some model shows it in nine of the eleven categories that hold 30 or more pairs and in eight of the twenty-two smaller ones. The two large exceptions are household-visible behaviors (32 pairs) and behaviors under high personal agency (39), in which no model shows it. Where fidelity asymmetry does fall along a property, as from free behaviors (six models) to trivial (four) to modest cost (none), the number of pairs falls with it (53, 25 and 17), so these data cannot separate the property from the loss of power. The cells in which a model shows neither form are few (13) and scattered across properties. The largest group of them is behaviors that need an institution (19 pairs), where the three frontier models and Qwen2.5-14B show neither. Behavioral underspecification therefore does not attach to a recognizable kind of behavior. The rows are not independent evidence, since one pair belongs to a category of every property, and the same-survey, self-reported pairs make up the largest category of several.

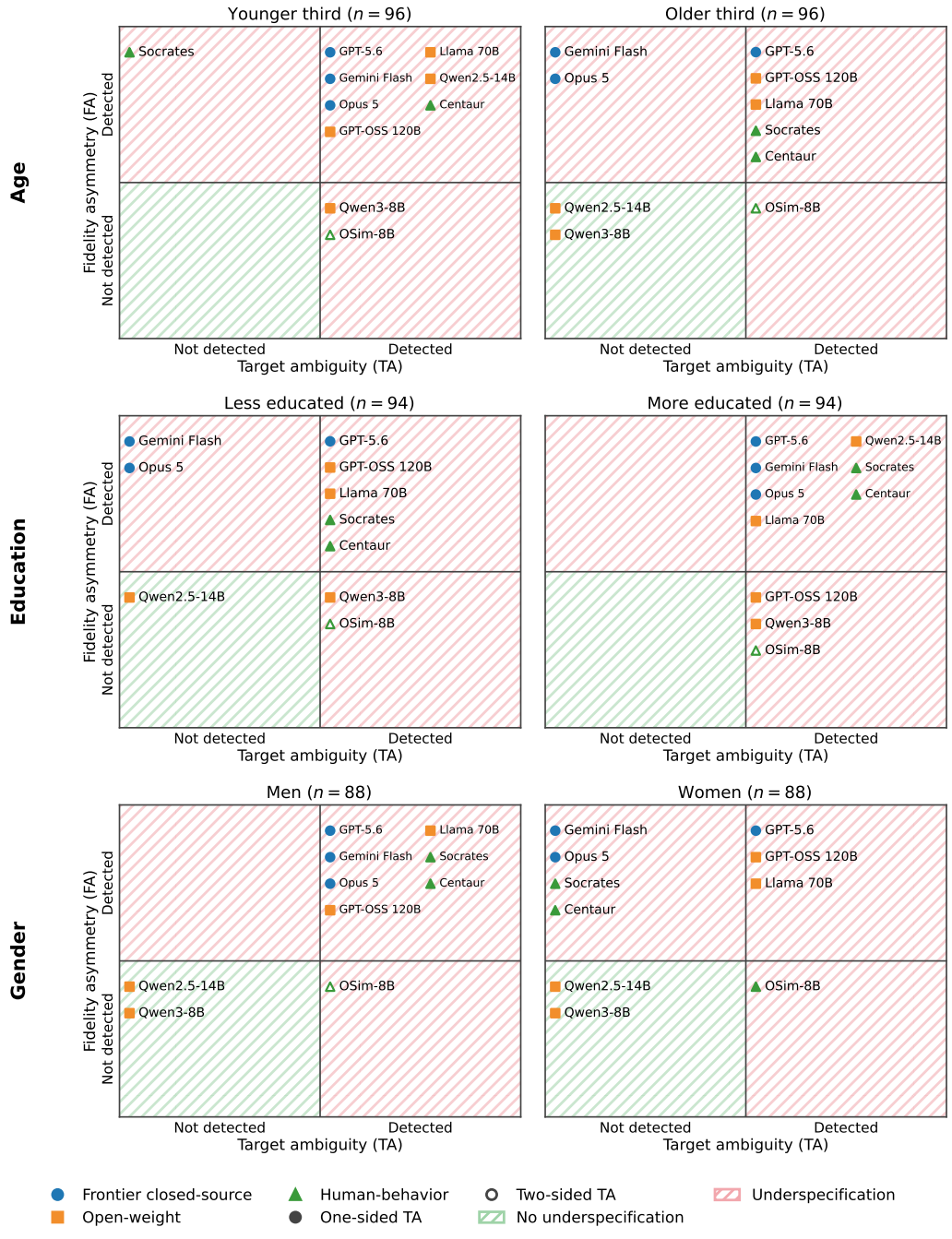


Figure 9: **Behavioral underspecification by demographic group: age, education and gender.** Each panel is the quadrant chart of Figure 3 recomputed from one group's own predicted and observed yes rates, for all ten models. The two panels of a row are the two groups of one attribute over the same say–do pairs, and n is the number of pairs. Hollow markers are two-sided target ambiguity. Figure 10 gives income, politics and race.

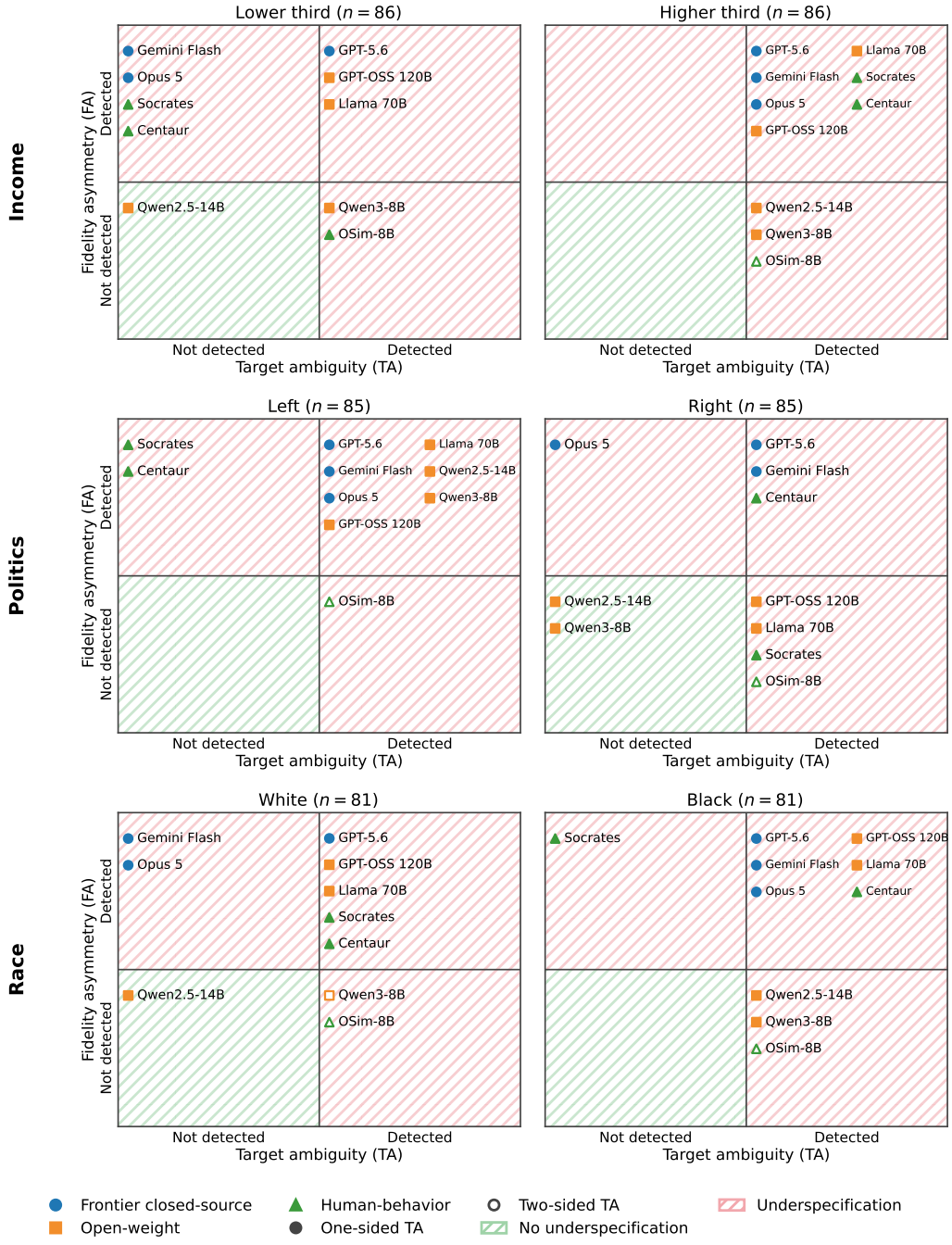


Figure 10: **Behavioral underspecification by demographic group: income, politics and race.** As Figure 9. Left and right are the thirds of an ideology scale, or Democrats and Republicans where a source records only party.

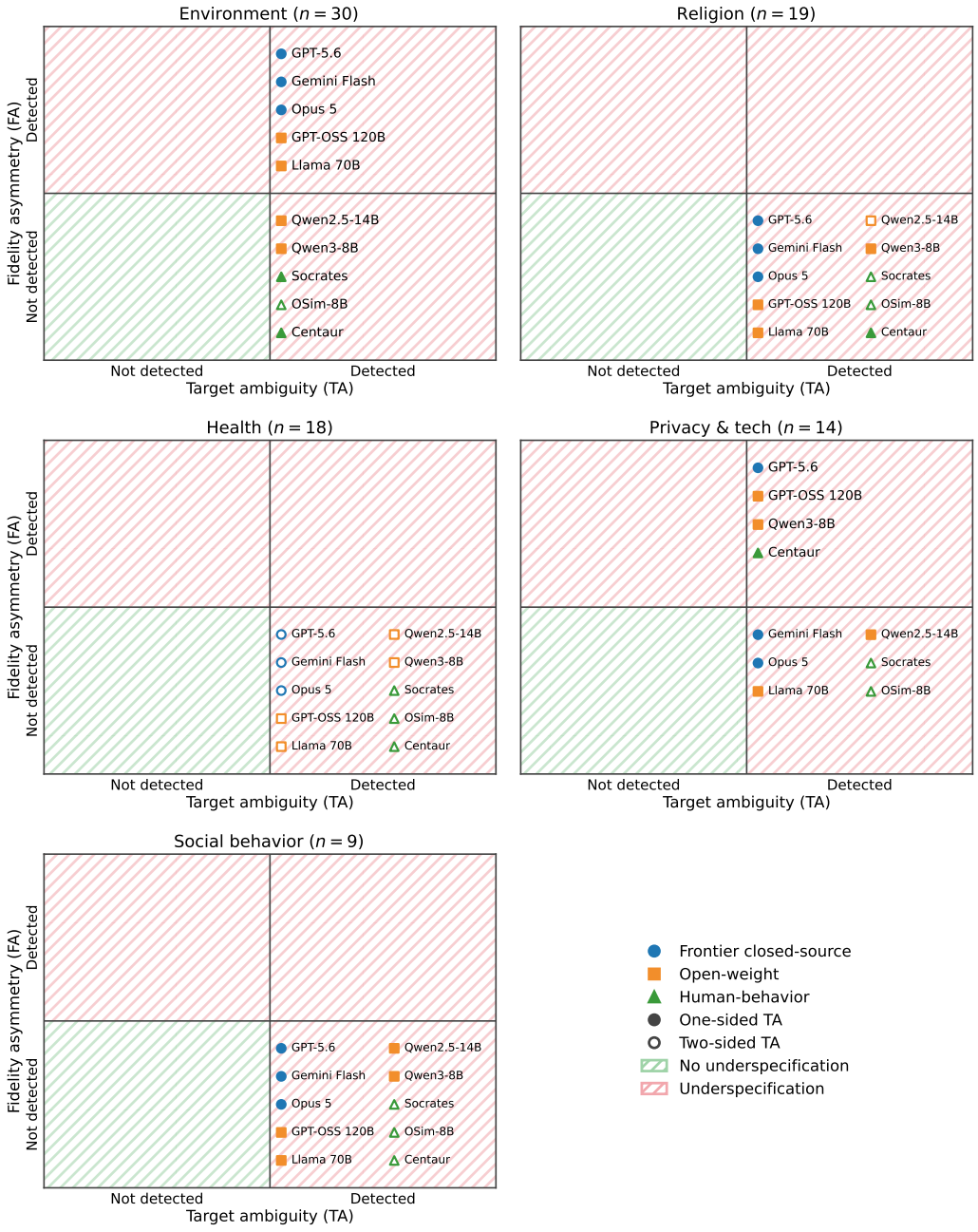


Figure 11: **Behavioral underspecification by domain.** Each panel is the quadrant chart of Figure 3 recomputed on the say–do pairs of one domain, for all ten models, and n is the number of pairs. Civic behavior (5 pairs) and education and work (3) are too small to test and are omitted.

E Demographic Differences Cancel Across Pairs

Appendix D.1 finds both forms of behavioral underspecification in every demographic group. This appendix asks a finer question, whether the preference for what people say over what they do is stronger for some groups than for others. Because a group inside one pair can be small, we measure that preference on individual respondents rather than on yes rates. The *say advantage* is the share of respondents whose say answer the model predicts correctly, minus the same share for the do answer, minus the same difference for a guesser that always gives the majority answer. The last subtraction

Property	Categories	Why we coded it
Cost to self	free, trivial, modest, substantial	Acting can cost money or effort that merely stating does not, so the do side may fall further short as the behavior gets more expensive.
Frequency	one-off, occasional, repeated, continuous	A behavior that must be kept up is harder to sustain than a single act, so recurrence may change how far doing trails saying.
Temporal separation	same survey, weeks, months to years	The longer between the statement and the recorded action, the more room an intention has to lapse before it is measured.
Social visibility	fully private, household-visible, semi-public, public, highly public	Visible behavior carries social pressure that a private survey answer does not, which can pull the stated and actual rates together or apart.
Dependency on others	independent, minor, some coordination, needs an institution, needs another person	A behavior that needs other people or an institution can fail for reasons outside the person's control, unlike a stated intention.
Time taken	under a minute, minutes, about an hour, hours, days to weeks, months, years, and not applicable for states	The time one instance of the behavior demands, which like cost could separate intending from doing.
Personal agency	low (gatekept), moderate, high, sole control	How much the outcome is in the person's own hands, since low agency caps how far doing can follow saying.
Gap category	intention, attitude, value, norm, belief, identity	The kind of statement, since an intention is a nearer commitment to acting than a general value or belief (Appendix B.1).
Relative timing	concurrent or prospective, each with a self-reported, objective or incentivized action	How the pair is designed and how the action is captured, which sets how directly the do side reflects real behavior (Appendix B.2).

Table 3: **The nine properties we coded each say-do pair on, their categories, and why each might relate to the say-do gap.** A category is a level a pair can take on that property. Both forms of behavioral underspecification are tested across these levels in Figure 12. Gap category and relative timing are the two taxonomy axes and are defined in full in Appendix B.1 and Appendix B.2.

removes the advantage that base rates alone confer. The say advantage is positive when what people say is predicted better, the opposite sign to Δ_{FA}.

Everything here covers the ten models under the independent design. The eight attributes are age, education, gender, income, politics, race, region and religion, and which of them a source records varies from survey to survey. For age, education, gender and income, every source can be reduced to the same two groups, the younger and older third, the less and more educated, men and women, and the lower and higher income third, so the direction of a difference means the same thing in every pair. For politics, race, region and religion, the groups a source records differ from survey to survey, an ideology scale in one and a party in another, so only whether the groups differ can be compared across pairs, not in which direction. One *cell* is one group of one attribute, inside one say–do pair, for one model, and there are $21{,}263$ of them. The one pair whose source records no respondent attributes is omitted. When groups are compared after averaging over pairs, a group is kept only if it appears in at least 50 of the 97 remaining pairs, which leaves 20 of the 125 groups the sources define. This keeps a difference between two groups from arising only because they were measured on different questions.

Which question is asked matters far more than which model or which person. The say advantage is not a constant a model carries with it, and so neither is the fidelity asymmetry it tracks. Across say-do pairs it runs from 25 points against the say side to 41 points in favour of it, and the

		n	GPT-5.6	Gemini Flash	Opus 5	GPT-OSS 120B	Llama 70B	Qwen2.5-14B	Qwen3-8B	Socrates	OSim-8B	Centaur
Gap category	Intention	13	TA^2	TA^2	TA^2	TA^2	TA^2	TA^2	TA^2	TA^2	TA^2	TA^2
	Attitude	44	TA	TA	FA+TA	FA+TA	TA	TA	TA	TA	TA^2	none
	Value	12	FA+TA	TA	FA+TA	TA	TA	TA^2	TA^2	TA	TA^2	TA
	Norm	6	TA	TA	TA	TA	TA	TA	TA	TA^2	TA^2	TA
	Belief	10	TA	TA	TA	TA	TA	TA	TA	TA^2	TA^2	TA^2
	Identity	13	FA+TA	TA	TA^2	TA^2	TA	TA^2	TA^2	TA^2	TA^2	TA
Frequency	one-off	16	TA	TA	TA	TA	TA	TA	TA^2	TA^2	TA^2	TA^2
	occasional	14	TA	none	none	TA^2	TA	TA	TA^2	TA	TA^2	TA
	repeated	57	FA+TA	FA+TA	FA+TA	TA	TA	TA	TA	FA+TA^2	TA^2	TA
	continuous	11	TA	TA	TA	TA	FA+TA	TA	FA+TA	TA^2	TA^2	FA+TA
Cost to self	free	53	FA+TA	FA+TA	FA+TA	FA+TA	FA+TA	TA	TA	none	TA^2	FA+TA
	trivial	25	FA+TA	FA+TA	FA+TA	TA^2	TA	TA	TA^2	FA+TA^2	TA	TA^2
	modest $	17	TA^2	TA^2	TA^2	TA^2	TA^2	TA^2	TA^2	TA^2	TA^2	TA^2
	substantial $	3	too few pairs to test									
Temporal separation	same survey (0 days)	62	FA+TA	FA	FA	FA+TA	FA+TA	none	TA	FA+TA	TA^2	FA+TA
	weeks (study-specific)	23	TA	TA	TA	TA	TA	TA	TA^2	TA^2	TA^2	TA^2
	months to years (panel)	13	FA+TA	FA+TA	TA	TA^2	TA	TA^2	TA^2	TA^2	TA^2	TA
Social visibility	fully private	20	TA	TA	FA+TA	TA	TA	TA	TA	TA^2	TA^2	TA
	household-visible	32	TA	TA	TA	TA	TA	TA	TA^2	TA^2	TA^2	TA
	semi-public	32	FA+TA	TA	FA+TA	FA+TA	FA+TA	TA	TA^2	TA^2	TA^2	TA
	public	12	TA	TA	TA	TA	TA	TA	TA^2	TA	TA^2	FA+TA
	highly public	2	too few pairs to test									
Dependency	independent	41	FA+TA	TA	FA+TA	TA	TA	TA	TA	FA+TA	TA	FA+TA
	minor	24	TA	TA	TA	TA^2	TA	TA^2	TA^2	TA^2	TA^2	TA^2
	some coordination	10	TA	TA	TA	TA	TA	TA	TA	TA^2	TA	TA
	needs institution	19	none	none	none	TA	TA	none	TA	TA	TA^2	TA
	needs a person	4	too few pairs to test									
Relative timing	concurrent, self-report	77	FA+TA	FA	FA+TA	FA+TA	FA+TA	none	none	FA+TA	TA	FA+TA
	prospective, self-report	15	TA	TA	TA^2	TA^2	TA	TA^2	TA^2	TA^2	TA^2	TA^2
	prospective, objective	3	too few pairs to test									
	concurrent, incentivized	2	too few pairs to test									
	concurrent, objective	1	too few pairs to test									
Time taken	<1 min	3	too few pairs to test									
	~mins	48	FA+TA	TA	FA+TA	TA	TA	TA	TA^2	TA^2	TA^2	FA+TA
	~1 hr	19	FA+TA	FA+TA	TA	TA	TA	TA^2	TA	TA^2	TA^2	TA^2
	hours	9	TA	TA	TA	TA	TA	TA	TA	TA^2	TA^2	TA
	days–weeks	4	too few pairs to test									
	months	1	too few pairs to test									
	years	1	too few pairs to test									
	n/a (state)	13	TA	TA	TA	TA	TA	TA	TA	TA^2	TA	TA
Personal agency	low (gatekept)	3	too few pairs to test									
	moderate	19	TA	none	none	TA	TA	TA	TA	TA	TA	TA
	high	39	TA	TA	TA	TA	TA	TA	TA	TA^2	TA^2	TA
	sole control	37	FA+TA	TA	FA+TA	FA+TA	TA	TA	FA+TA	TA	TA^2	FA+TA

FA and TA | One form only | Neither | Too few pairs

Figure 12: **Behavioral underspecification by property of the behavior.** One row per category of the nine properties of Table 3, with n the number of say–do pairs in it, and one column per model. Each cell gives the quadrant of Figure 3 that the model falls in when the tests are re-run on that category's pairs. FA+TA is both forms, FA or TA one form, and none neither. A superscript 2 marks two-sided target ambiguity. Grey rows hold five or fewer pairs, too few for the tests to reach $q < 0.05$.

spread across pairs is 11.6 points $[9.5, 13.5]$ against 3.0 $[1.4, 3.9]$ across models and 1.7 $[1.1, 2.1]$ across demographic groups (Figure 13). The interval for pairs clears both of the others, while the model and group intervals overlap each other, so we claim no order between those two. Whether a

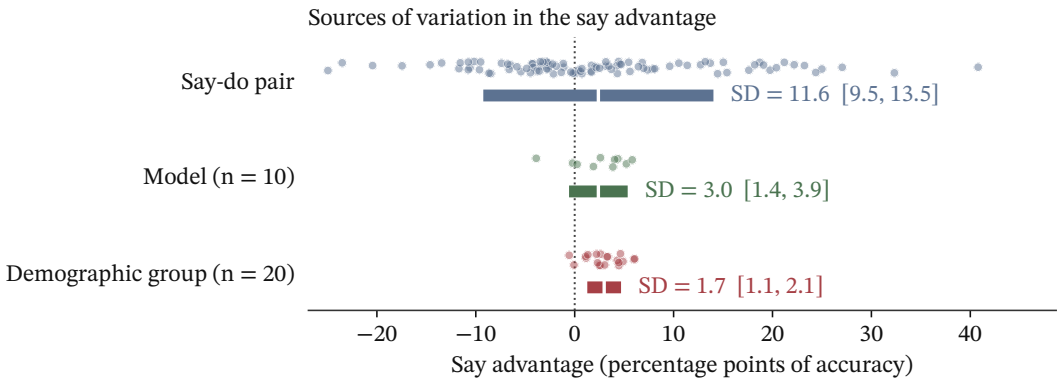


Figure 13: **Which question is asked moves the say advantage more than which model or which person.** The say advantage, say-minus-do accuracy net of a majority-answer guesser, in percentage points. Each dot is one say-do pair, one model or one demographic group, and the bar under each row is the spread of those dots with the slices' own sampling noise removed, labelled with a bootstrap interval over the units resampled. The three rows are the same quantity in the same units, so their spreads can be read against each other directly.

	Say advantage				Predicted yes rate			
	one pair		all pairs		one pair		all pairs	
Attribute	**diff**	**cells**	**diff**	**models**	**diff**	**cells**	**diff**	**models**
Age	3.7	24%	0.0	0/10	5.1	41%	2.7	0/10
Education	2.9	23%	0.4	0/10	4.2	34%	0.0	0/10
Gender	1.8	16%	0.0	0/10	2.0	19%	0.0	0/10
Income	4.1	20%	2.7	0/10	2.7	24%	1.0	0/10
Politics	7.4	39%	1.8	0/10	8.6	59%	3.6	3/10
Race	4.2	23%	2.6	0/10	4.7	35%	2.6	3/10
Region	0.0	11%	0.0	0/10	0.8	16%	0.0	0/10
Religion	5.4	32%	0.6	0/10	8.7	53%	3.3	3/10

Table 4: **How far apart two demographic groups are, attribute by attribute.** The full version of Figure 14. Above the rule are the attributes whose two groups are the same in every source, and below it those whose groups differ from survey to survey. *diff* is the typical difference between two groups of that attribute in percentage points, with the cells' own sampling noise removed, either inside a single say-do pair or after averaging each group over all pairs. *cells* is the share of pair-by-model cells in which a test tells the groups apart, and *models* how many of the ten do so after averaging. Differences shrink from left to right within each half of the table, which is what cancellation across pairs looks like.

model predicts what people say better than what they do is therefore mostly a property of what is being asked, and only slightly a property of who is asking or of whom it is asked about.

How the two levels are tested. Inside one pair, the unit of a test is the person. After averaging over pairs it is the pair, so a difference between two groups counts only if it would reappear on a new question. Using the person at that level instead would establish almost every difference, because each group average is built from tens of thousands of people. It would establish 49 of the 80 model-and-attribute combinations on accuracy and 61 of 80 on the predicted yes rate, against 0 and 9 when the pair is the unit.

Group differences are real inside a single pair and do not carry over to the next one. Take one say-do pair, one model and one attribute, and the groups do come apart. Two groups typically differ by a few points in the say advantage, and the difference is established in between a tenth and two fifths of these cells (Figure 14a, light bars). On age, education, gender and income the established ones have a typical size of 8.9 points, and a test of whether an attribute's groups differ at all, run

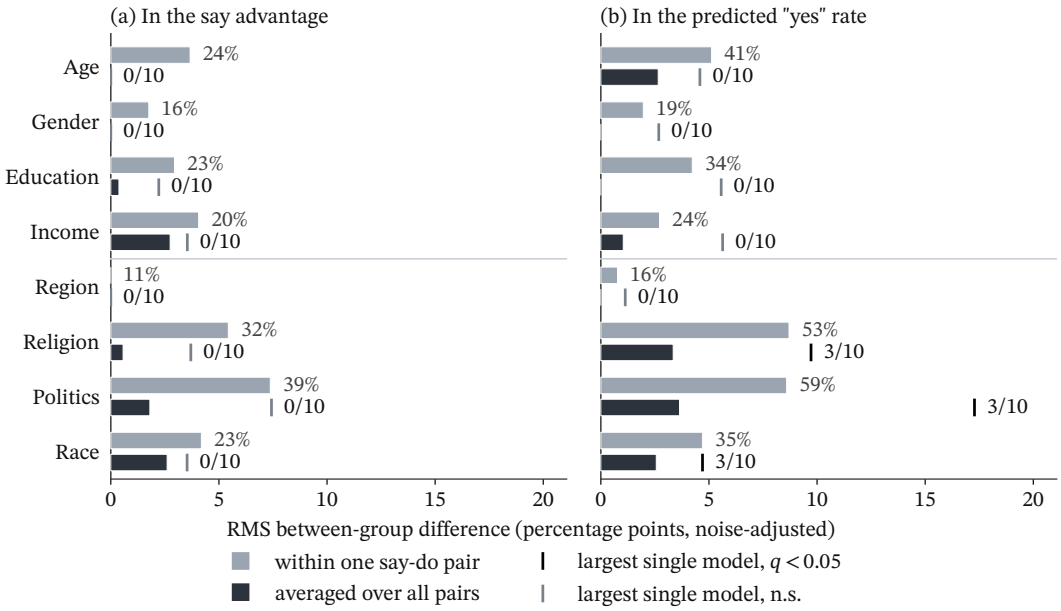


Figure 14: **Group differences are real inside one say-do pair and do not carry over to the next.** How far apart two groups of the same attribute are, in the say advantage (a) and in the predicted yes rate (b), on a shared scale. Light bars are measured inside a single say-do pair and dark bars after averaging each group over every pair it appears in, and the tick marks the model with the largest difference. The label on a light bar is the share of pair-by-model cells in which a test tells the groups apart, and the label on a dark bar is how many of the ten models it tells apart after averaging. Above the rule are the attributes whose two groups are the same in every source, and below it those whose groups differ from survey to survey.

separately in each pair, rejects in between a fifth and two fifths of them. What does not survive is the step from one pair to many. Averaging each group over all say-do pairs shrinks the differences close to zero, and no model shows an established difference in the say advantage between the groups of any of the eight attributes (dark bars). The reason is visible in the directions. Among the significant single-pair differences on age, education, gender or income the higher-scoring group is the higher-income one 78% of the time but the male one only 36% of the time, so the averages cancel. The differences are therefore not absent. They point different ways in different pairs, and what we can say is that demographic effects do not carry across pairs, which is a stronger statement than that they are small.

What is left concerns religion, politics and race, and only what the model predicts. The same comparison for the predicted yes rate leaves more standing (Figure 14b). Age, education, gender and income again come out with nothing established, but religion, politics and race each keep a difference of a few points for three of the ten models, and the largest single case reaches 17 points. A model can therefore predict that one group says yes more often than another without predicting either group's say answers better than its do answers. For the story of the main text, the conclusion is that no demographic group carries the preference for what people say. Inside a single pair groups do differ, but the direction changes from pair to pair.

F Detailed Results under Different Prompting Paradigms

A prompting paradigm fixes what a model is told about a person and which of that person's two answers it has to produce. Every result in Section 5.1 uses one paradigm, and Section 5.2.3 adds a second. This appendix reports all four paradigms we ran. They share the demographic profile, the wording of each item and the answer format, and their prompts are given in full in Appendix H.

- **Respondent.** The model is given the person's demographic profile and asked, in the first person, either what the person says or what the person does. The two questions are asked in separate calls, so neither prediction sees the other, and neither prompt mentions that the two answers can differ. We also call this the independent paradigm. It produces Table 1 and every result in Section 5.1, and it is the baseline each paradigm below is compared with.
- **Forward** (Appendix F.1). The model is given the profile and the answer the person actually gave to the say question, and predicts the do answer. This is the TPB-inspired conditioning of Section 5.2.3. Only P'_{do} is produced differently, and P'_{say} is the Respondent prediction.
- **Reverse** (Appendix F.2). The mirror of Forward. The model is given the profile and what the person actually did, and predicts the say answer. Only P'_{say} is produced differently, and P'_{do} is the Respondent prediction.
- **Joint** (Appendix F.3). The model is given the profile alone, as in Respondent, but produces the say answer and the do answer together in a single generation. No human answer is supplied, so unlike Forward and Reverse this paradigm could be used in deployment. Socrates and Centaur were not run under it.

Unlike the Respondent prompts, the Forward, Reverse and Joint prompts also state that people sometimes do not do what they say. Each of these paradigms therefore differs from Respondent in two ways, the information or format it adds and that statement, and the comparisons below measure the two together. Each paradigm's table has the form of Table 1, with the same columns, units, tests and correction, so its last three columns are the contrasts Δ_{FA}, $\Delta_{\mathrm{TA}}^{\mathrm{say}}$ and $\Delta_{\mathrm{TA}}^{\mathrm{do}}$ of Definitions 3.1 and 3.2, and every number is in percentage points.

F.1 The Forward Paradigm, the Do Prediction Given the Say Answer

Conditioning on the say answer removes fidelity asymmetry for every model. With the person's own say answer supplied, Δ_{FA} is not significant for any of the ten models (Table 5), where seven showed fidelity asymmetry under independent calls. Only P'_{do} changes, so this comes entirely from the do side. $d(P'_{\mathrm{do}}, P_{\mathrm{do}})$ falls by 4.0 to 7.7 points for the six models not built on Qwen, for example from 19.9 to 12.2 for GPT-5.6, and for five of those six the cued do error lands within about a point of the same model's say error. For the four Qwen-based models the change is small. The do error falls by 1.1 to 2.7 points for Qwen2.5-14B, Socrates and OSim-8B, and rises by 1.0 point for Qwen3-8B.

Target ambiguity is removed for most models, but not all. Six models distinguish both targets under the cue and so exhibit neither form: GPT-5.6, Gemini Flash, Opus 5, Qwen2.5-14B, Qwen3-8B and Centaur. Four of them were underspecified under independent calls, GPT-5.6 and Centaur in both forms and Gemini Flash and Opus 5 through fidelity asymmetry alone, while the two Qwen models exhibited neither form to begin with. Of the other four, GPT-OSS 120B and Llama 70B lose their fidelity asymmetry and keep one-sided target ambiguity with the do target ambiguous, Socrates loses its fidelity asymmetry and keeps one-sided target ambiguity with the say target ambiguous, and OSim-8B keeps the two-sided target ambiguity it showed before. The cue is the person's true say answer, which no deployment would have, so this measures how much of the do-side underspecification is missing information rather than offering a method for removing it. Section 5.2.3 discusses how a model could supply that answer itself.

F.2 The Reverse Paradigm, the Say Prediction Given the Do Answer

Conditioning on the do answer does not reduce behavioral underspecification. Self-perception theory Bem (1972) runs the arrow the other way, with people reading their own attitudes partly off what they have already done, so we also gave the say prediction the person's observed do answer. The cue helps far less in this direction. It reduces $d(P'_{\mathrm{say}}, P_{\mathrm{say}})$ by at most 3.5 points (Llama 70B), against 7.7 on the do side in the forward direction, and it leaves Opus 5 and OSim-8B slightly worse (Table 6). The say side is the easier of the two to predict without any cue, which leaves a cue less error to remove there. Because the improvement lands on the side that was already better predicted, fidelity asymmetry does not shrink: the same seven models exhibit it as under independent calls, and for six of them Δ_{FA} is larger. Target distinction is lost rather than gained. The cued

		Distance in points (mean ± SD across datasets)				Behavioral underspecification contrasts		
Type	**Model**	$d(P'_{\text{say}}, P_{\text{say}})\downarrow$	$d(P'_{\text{say}}, P_{\text{do}})\uparrow$	$d(P'_{\text{do}}, P_{\text{do}})\downarrow$	$d(P'_{\text{do}}, P_{\text{say}})\uparrow$	Δ_{FA}	$\Delta_{\text{TA}}^{\text{say}}$	$\Delta_{\text{TA}}^{\text{do}}$
Frontier closed-source	**GPT-5.6**	13.1 ± 10.5	29.6 ± 20.6	12.2 ± 9.4	18.7 ± 20.4	$+1.0^{\text{n.s.}}$	-16.5^{***}	-6.5^{*}
	Gemini Flash	12.0 ± 10.3	27.9 ± 18.9	10.9 ± 10.6	20.9 ± 23.5	$+1.1^{\text{n.s.}}$	-16.0^{***}	-10.0^{***}
	Opus 5	11.5 ± 11.7	27.7 ± 20.5	14.0 ± 12.1	25.1 ± 22.4	$-2.5^{\text{n.s.}}$	-16.2^{***}	-11.0^{***}
Open-weight	**GPT-OSS 120B**	14.0 ± 12.4	26.0 ± 18.4	14.2 ± 10.2	14.2 ± 15.5	$-0.2^{\text{n.s.}}$	-12.0^{***}	$-0.1^{\text{n.s.}}$
	Llama 70B	14.0 ± 13.7	30.6 ± 20.4	15.2 ± 11.5	12.5 ± 17.0	$-1.2^{\text{n.s.}}$	-16.7^{***}	$+2.7^{\text{n.s.}}$
	Qwen2.5-14B	19.5 ± 17.6	31.6 ± 21.4	18.4 ± 15.8	24.9 ± 23.8	$+1.2^{\text{n.s.}}$	-12.0^{***}	-6.5^{**}
	Qwen3-8B	18.5 ± 15.6	24.2 ± 18.0	18.8 ± 14.6	26.7 ± 22.2	$-0.3^{\text{n.s.}}$	-5.7^{*}	-7.8^{**}
Human-behavior models	**Socrates**	15.2 ± 10.4	19.5 ± 17.8	18.9 ± 14.3	27.2 ± 21.0	$-3.7^{\text{n.s.}}$	$-4.3^{\text{n.s.}}$	-8.3^{***}
	OSim-8B	17.2 ± 12.3	18.8 ± 16.1	15.5 ± 13.5	16.1 ± 14.2	$+1.6^{\text{n.s.}}$	$-1.6^{\text{n.s.}}$	$-0.5^{\text{n.s.}}$
	Centaur	14.0 ± 12.7	25.8 ± 19.8	13.2 ± 11.3	17.3 ± 13.9	$+0.8^{\text{n.s.}}$	-11.8^{***}	-4.2^{**}

Table 5: **The Forward paradigm, where the do prediction is told what the person said.** Every column means what it means in Table 1, and only P'_{do} is produced differently, from a call that supplies the person's own say answer. P'_{say} is the same prediction as in Table 1, so the first two distance columns repeat that table to within a tenth of a point and the last two are the ones to read. Superscripts are q-values corrected across the ten models in the column, where * marks $q < 0.05$, ** marks $q < 0.01$ and *** marks $q < 0.001$. An $^{\text{n.s.}}$ means not established.

		Distance in points (mean ± SD across datasets)				Behavioral underspecification contrasts		
Type	**Model**	$d(P'_{\text{say}}, P_{\text{say}})\downarrow$	$d(P'_{\text{say}}, P_{\text{do}})\uparrow$	$d(P'_{\text{do}}, P_{\text{do}})\downarrow$	$d(P'_{\text{do}}, P_{\text{say}})\uparrow$	Δ_{FA}	$\Delta_{\text{TA}}^{\text{say}}$	$\Delta_{\text{TA}}^{\text{do}}$
Frontier closed-source	**GPT-5.6**	9.7 ± 8.1	24.5 ± 19.9	19.9 ± 14.2	22.2 ± 18.6	-10.2^{***}	-14.8^{***}	$-2.3^{\text{n.s.}}$
	Gemini Flash	9.8 ± 8.6	25.0 ± 19.1	16.8 ± 13.3	22.6 ± 19.2	-7.0^{***}	-15.3^{***}	-5.8^{*}
	Opus 5	11.7 ± 12.5	27.6 ± 20.9	18.0 ± 13.1	23.8 ± 20.6	-6.3^{***}	-15.9^{***}	-5.8^{*}
Open-weight	**GPT-OSS 120B**	13.1 ± 10.3	13.8 ± 12.9	18.7 ± 11.8	19.5 ± 16.7	-5.6^{***}	$-0.7^{\text{n.s.}}$	$-0.7^{\text{n.s.}}$
	Llama 70B	10.5 ± 8.6	21.2 ± 18.5	20.4 ± 15.3	19.5 ± 17.4	-9.9^{***}	-10.7^{***}	$+0.8^{\text{n.s.}}$
	Qwen2.5-14B	17.7 ± 16.0	19.3 ± 18.2	21.1 ± 17.1	26.9 ± 22.1	$-3.4^{\text{n.s.}}$	$-1.6^{\text{n.s.}}$	-5.9^{**}
	Qwen3-8B	17.4 ± 16.6	18.7 ± 15.1	17.8 ± 14.3	23.1 ± 19.9	$-0.4^{\text{n.s.}}$	$-1.3^{\text{n.s.}}$	-5.3^{*}
Human-behavior models	**Socrates**	14.5 ± 10.5	19.0 ± 16.9	20.0 ± 14.1	27.6 ± 21.0	-5.4^{**}	$-4.5^{\text{n.s.}}$	-7.6^{**}
	OSim-8B	19.0 ± 13.2	16.4 ± 13.2	16.7 ± 14.6	18.4 ± 14.2	$+2.3^{\text{n.s.}}$	$+2.6^{\text{n.s.}}$	$-1.7^{\text{n.s.}}$
	Centaur	10.9 ± 9.6	15.4 ± 14.6	19.5 ± 14.9	20.8 ± 14.5	-8.6^{***}	-4.5^{*}	$-1.3^{\text{n.s.}}$

Table 6: **The Reverse paradigm, where the say prediction is told what the person did.** The mirror of Table 5. Here the cue goes to P'_{say} and P'_{do} is the same prediction as in Table 1, so the last two distance columns repeat that table to within a tenth of a point and the first two are the ones to read. Every column otherwise means what it means in Table 1. Superscripts are q-values corrected across the ten models in the column, where * marks $q < 0.05$, ** marks $q < 0.01$ and *** marks $q < 0.001$. An $^{\text{n.s.}}$ means not established.

say prediction moves toward what people do, so GPT-OSS 120B, Qwen2.5-14B and Qwen3-8B no longer distinguish the say target. That leaves GPT-OSS 120B with two-sided target ambiguity, and the two Qwen models, which exhibited neither form under independent calls, with one-sided target ambiguity on the say side.

F.3 The Joint Paradigm, Both Answers in One Call

Asking for both sides at once removes target ambiguity but not fidelity asymmetry. Here both answers come from a single call, which, like the forward and reverse prompts, tells the model that people sometimes do not do what they say. No human answer is supplied, so unlike the Forward paradigm this is something a deployment could do. The single call and the statement change together, so what follows is their joint effect. Eight of the ten models take part (Figure 15). Under independent calls four of these eight exhibit target ambiguity, GPT-5.6, GPT-OSS 120B and Llama 70B one-sided and OSim-8B two-sided. In the joint call every model distinguishes the do target, at $q < 0.01$ or better and by 5.9 to 12.3 points, and seven of the eight distinguish both targets, so only OSim-8B still exhibits target ambiguity, now one-sided with the say target ambiguous. Fidelity asymmetry, by contrast, mostly persists or appears. It is detected for five of the eight, including Qwen2.5-14B and Qwen3-8B, which showed none under independent calls. It is no longer detected for GPT-5.6 (-2.9) or Gemini Flash, although Gemini Flash's Δ_{FA} of -4.6 points is nearly its -4.8 under independent calls. Table 7 gives this paradigm the full form of Table 1, including the four distances the figure does not show.

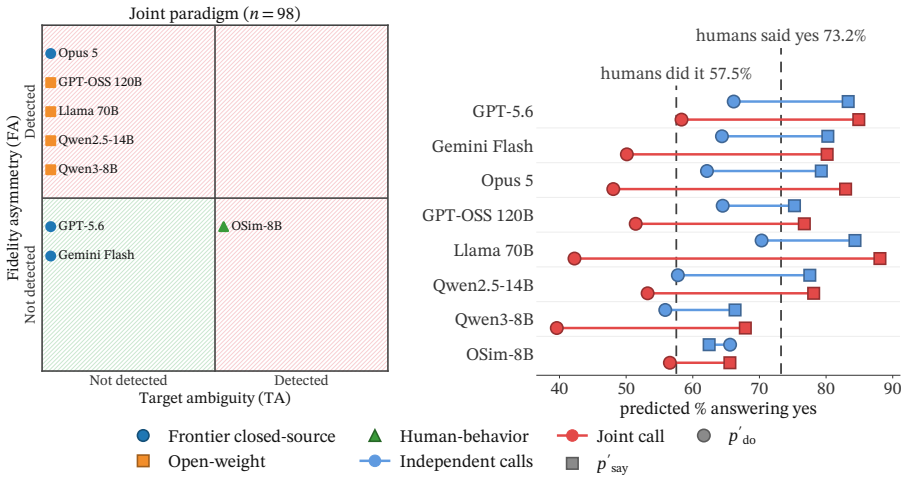


Figure 15: **What changes when both answers are asked for in one call.** On the left, behavioral underspecification under the Joint paradigm, drawn as in Figure 3. Each model is placed by whether fidelity asymmetry and target ambiguity are detected in the tests of Table 7, with q-values corrected across the eight models that have a Joint run. No model shows two-sided target ambiguity, so every marker is filled. On the right, where each paradigm puts a model's two mean predicted yes rates, the upper blue bar for the Respondent paradigm's two independent calls and the lower red bar for the single Joint call. A square is the predicted say rate p'_{say}, a circle the predicted do rate p'_{do}, and the dashed lines are the human rates p_{do} and p_{say}. A long bar means the model has pulled the two sides apart, whether or not either end of it is in the right place. Socrates and Centaur have no Joint run.

The separation is not an improvement in accuracy. The ladder on the right of Figure 15 shows how the do target becomes distinguished. Between the two paradigms the mean predicted say rate moves by at most 3.7 points for any model, while the mean predicted do rate falls for every model, by between 4.5 and 28.1 points. Measured within each say–do pair and then averaged, the size of the gap between a model's predicted say and do rates widens for seven of the eight, at the extreme from 19.7 to 47.5 points for Llama 70B. Measured the same way, people show a gap of 20.3 points, and seven of the eight models now predict a wider one. P'_{do} thus ends up further from P_{say} without ending up closer to P_{do}. Measured against the Respondent paradigm, the do-side error improves for no model at $q < 0.05$ and grows significantly for two of the eight, Llama 70B (from 20.4 to 26.9 points) and Qwen3-8B (from 17.8 to 26.0). For Qwen3-8B, this larger do error is what produces the fidelity asymmetry it did not have before. Qwen2.5-14B acquires one the other way, with its do error unchanged at 21.1 points and its say error falling from 19.5 to 16.2. The joint call therefore trades target ambiguity for an exaggerated say–do gap, a further case of the point made in Section 5.1 that the absence of one form of behavioral underspecification does not imply accurate simulation.

G Details of Statistical Tests

Every test in the paper has the same shape, so we give it once here and note the few places that depart from it.

One say-do pair is one observation. A distance is a property of a whole sample rather than of a person, so a single say-do pair yields one number per quantity and not N of them. Every quantity we test is a difference between two of those distances taken on the same pair, such as Δ_{FA} for one model, and the 98 pairs supply 98 such differences. Those differences are the sample every test below runs on, and the number of respondents inside a pair enters only through how precisely that pair's four rates are measured.

The tests are paired, and the target-distinction tests are one-sided. The two distances in a difference are computed on the same pair from the same responses, and in both target-distinction contrasts they share a term, since $d(P'_{\text{say}}, P_{\text{say}})$ and $d(P'_{\text{say}}, P_{\text{do}})$ are both distances from the same

Type	Model	Distance in points (mean ± SD across datasets)				Behavioral underspecification contrasts		
		$d(P'_{\text{say}}, P_{\text{say}})\downarrow$	$d(P'_{\text{say}}, P_{\text{do}})\uparrow$	$d(P'_{\text{do}}, P_{\text{do}})\downarrow$	$d(P'_{\text{do}}, P_{\text{say}})\uparrow$	Δ_{FA}	$\Delta_{\text{TA}}^{\text{say}}$	$\Delta_{\text{TA}}^{\text{do}}$
Frontier closed-source	**GPT-5.6**	14.1 ± 11.7	30.7 ± 21.0	17.0 ± 12.9	26.2 ± 23.6	$-2.9^{\text{n.s.}}$	-16.6^{***}	-9.2^{***}
	Gemini Flash	11.9 ± 10.6	27.7 ± 19.0	16.4 ± 14.7	28.6 ± 26.4	$-4.6^{\text{n.s.}}$	-15.8^{***}	-12.2^{***}
	Opus 5	13.0 ± 13.9	29.4 ± 21.3	22.6 ± 18.6	34.9 ± 28.2	-9.7^{***}	-16.4^{***}	-12.3^{***}
Open-weight	**GPT-OSS 120B**	12.7 ± 13.2	25.1 ± 18.5	20.2 ± 15.2	26.5 ± 22.7	-7.6^{***}	-12.5^{***}	-6.3^{**}
	Llama 70B	16.6 ± 14.3	33.2 ± 20.0	26.9 ± 18.7	36.2 ± 26.7	-10.3^{***}	-16.6^{***}	-9.3^{***}
	Qwen2.5-14B	16.2 ± 15.5	28.2 ± 20.8	21.1 ± 17.6	27.0 ± 22.6	-4.9^{*}	-12.1^{***}	-5.9^{**}
	Qwen3-8B	16.5 ± 14.2	22.6 ± 17.4	26.0 ± 18.4	38.0 ± 23.9	-9.5^{***}	-6.1^{**}	-12.0^{***}
Human-behavior models	**OSim-8B**	14.7 ± 11.6	18.0 ± 16.2	17.4 ± 11.8	24.0 ± 16.7	$-2.7^{\text{n.s.}}$	$-3.2^{\text{n.s.}}$	-6.6^{***}

Table 7: **The Joint paradigm, where both sides come from one call.** Both predictions are read off a single generation that asks for the say answer and the do answer together and states that the two can diverge, so no column repeats Table 1 and no human answer is supplied. Socrates and Centaur have no joint run, the second because it answers with one token and cannot carry a two-part answer. Every column otherwise means what it means in Table 1. Superscripts are q-values corrected across the eight models in the column, where * marks $q < 0.05$, ** marks $q < 0.01$ and *** marks $q < 0.001$. An $^{\text{n.s.}}$ means not established.

prediction (and likewise for P'_{do}). Treating them as two independent samples would assume away a dependence the construction creates, so we test the per-pair difference. The fidelity-asymmetry hypothesis of Section 4.3 names no direction, so its test is two-sided. Each target-distinction hypothesis names one, so its test is one-sided in that direction. We compute the opposite tail as its own test rather than as one minus the first, because the signed-rank distribution is discrete and ties place mass at neither tail, and we need to be able to report that a difference runs significantly the wrong way instead of folding that into "not significant", which is what separates positive evidence of target ambiguity from a mere failure to detect target distinction.

We use a signed-rank test rather than a t test. The per-pair differences vary widely in size from pair to pair, so a few large ones can dominate a mean. We therefore use the Wilcoxon signed-rank test on the 98 differences, which asks whether they are consistently of one sign rather than whether their mean clears zero by enough.

Stars are Benjamini-Hochberg q-values inside a family of models. Each column of a results table applies one comparison to every model, so the models in a column are the family we correct across. We report Benjamini-Hochberg q-values controlling the false discovery rate within that family, ten models wide in the Respondent, Forward and Reverse paradigms and eight in the Joint paradigm, where two models have no run. Throughout, * marks $q < 0.05$, ** marks $q < 0.01$ and *** marks $q < 0.001$, and q never means anything else in this paper, since no raw p-value is reported. Where the main text and Table 1 write $p < 0.05$, $p < 0.01$ or $p < 0.001$, the value is this q. A small q says a difference is consistent across say-do pairs. It says nothing about whether the difference is large, which is what the point estimate beside it is for.

Intervals are computed two ways, one per kind of unit. The slope and spread intervals of Appendix C are percentile bootstrap intervals resampling say-do pairs, 5,000 draws, because those quantities are fitted over the pairs and have no closed-form standard error. The spread intervals of Figure 13 resample the units of each row, 4,000 draws, which is what makes rows built from 10 models and from 97 pairs comparable. Wherever a p comes from a bootstrap it cannot fall below one over the number of draws, so a value at that floor means below the resolution of the resampling and not exactly that number.

The 98 pairs are not fully independent, and we measured how far from it they are. They come from 31 source surveys, and 79 of the 98 come from the same survey as at least one other pair, with one panel wave contributing 15 pairs on its own. Treating all 98 as independent is therefore mildly anti-conservative. Collapsing each survey to one number is not the correction it looks like either, because those 15 pairs cover religious practice, environmental action and firearm safety, among other behaviors, and averaging them discards that variation and turns "holds across behaviors" into "holds across surveys". We settled it by measuring the intraclass correlation of the tested differences within surveys. For Δ_{FA} it runs from 0.15 to 0.57, which puts the effective number of independent pairs between 44 and 73 rather than at 31 or at 98. For the two target-distinction contrasts it is below

0.10 in 15 of the 20 model-by-comparison cells and never above 0.34. The reported tests treat each pair as one observation. As a check we also compute a cluster bootstrap over the 31 surveys and a collapse to the 31 survey means, both with 20,000 resamples.

Which verdicts survive the more conservative readings, and which do not. The three contrast columns of Table 1 report 20 verdicts at $q < 0.05$ across the ten models. Fifteen of them hold under the cluster bootstrap that keeps every pair's weight, and 7 hold under the collapse to 31 survey means. Say-side target distinction ($\Delta^{\text{say}}_{\text{TA}}$) is the robust one, and the six models with the largest effects on it hold at $q < 0.001$ under all three readings. Fidelity asymmetry keeps its size under the collapse for six of its seven models, between 2.7 and 7.2 points against 4.7 to 6.8 before, but with 31 units it no longer clears 0.05, and for Centaur it changes sign, which says that verdict is carried by particular waves rather than holding survey by survey. Of the five do-side target-distinction ($\Delta^{\text{do}}_{\text{TA}}$) verdicts, two hold under the cluster bootstrap and none under the collapse. Because target ambiguity is detected when target distinction is not, the collapse would therefore widen target ambiguity rather than remove it. We report the pair-level tests because the intraclass correlations above show the collapse over-corrects.

Two things a small q here does not license. These are distances between population rates, so a model can score zero on a pair while being wrong about every individual in it, and no test in this paper supports a claim about individual accuracy. Second, a target-distinction contrast on a pair can be no larger in size than the human say–do gap on that pair, $|p_{\text{say}} - p_{\text{do}}|$, so a pair where people show almost no gap contributes almost nothing to it. Eighteen of the 98 pairs have a human gap below 5 points, against a median of about 15, which dilutes the target-distinction effects we report. Fidelity asymmetry has no such bound.

The demographic tests. Appendix E compares the groups of one attribute with a fixed-effect homogeneity test over the group means, corrected separately within the four attributes whose two groups are the same in every source and within the four whose groups differ by survey. The unit is the person inside a single pair and the pair after averaging over pairs, for the reason given there.

Reproducibility. Every resampling runs from a fixed seed, so the tables and figures reproduce run to run, and every table is generated by a script from the stored model responses.

H Prompt Construction

Every call to a chat model is a **system message**, a fixed instruction that sets the model's role and the required answer format, followed by a **user message** carrying the specific case, which is a short demographic profile, the relevant item, and its response scale. The model returns a single **JSON object** and nothing else. Centaur, a completion model, receives the same content in its own native format (Appendix H.9). A call that asks one question uses the form `{"answer": <value>}`, a call that asks several uses one numbered field per question, and the joint design keys its two sides by name. Each answer key mirrors the key of its question, so `do-question-3` is answered by `do-answer-3`.

Following the persona-conditioned, key-value formatting of Kolluri et al. (2025), each user message renders the individual as a flat list of `- Attribute: value` lines and quotes each survey item in its original first-person, source-faithful wording, in the tense in which it was actually asked. Intentions stay in the future ("*Do you intend to* ..."), a completed act in the past ("*Did you vote* ..."), and recurring behavior in the habitual present ("*In a typical week, how many days do you exercise?*"). Items translated into English from another language are quoted in translation with no note about the original language, which covers 5.1% of all say–do pairs. Where a source item was written in the third person, or where the scored measure is a composite over several acts, the Respondent paradigm carries a first-person restatement declared in the pair's own configuration rather than being reworded at render time. That affects 16.3% of pairs, almost all of them on the "do" side, and every restatement preserves the item's content and tense.

The central quantity of interest is the gap between what people *say* and what they *actually do*, so each individual carries two measurements, a statement and an observed action. The four paradigms differ in which of the two the model is given and which it must produce.

- **Respondent.** The model is told to simulate the individual and answer in the first person, given only the demographic profile. The statement and the action are asked in two separate calls, and nothing in either prompt mentions that the two can diverge. This is the independent design and the paper's default.
- **Forward.** The model is given the profile *and* what the individual said, and produces what they did.
- **Reverse.** The model is given the profile *and* what the individual did, and produces what they said. Forward and Reverse together are the conditioned designs.
- **Joint.** From the profile alone, the model produces both sides in one generation, as parts (1) and (2) of a single answer.

All four system messages come from one family. Each is the sentence "*You are simulating a single individual* . . . ", then a task sentence, then the closing "*Remain consistent with the individual's perspective* . . . ". The task sentence is assembled from the same two predicates throughout, which are what the individual *says* about their stated values, attitudes, beliefs or intentions, and the action they would actually be *observed or recorded* as performing. The Respondent message inserts no task sentence at all, Forward asks for the action given the statement, Reverse is that sentence with the two swapped, and Joint asks for both as (i) and (ii).

Forward, Reverse and Joint additionally name the **say–do gap** outright, in the project's own vocabulary rather than the literature's. That wording is used because it is the only one accurate across the whole roster. Only 12.2% of pairs put a behavioral intention on the "say" side (Appendix B), and the rest measure a stated attitude, value, belief, norm or self-description, so "stated intention" would over-claim on the large majority. The Respondent message names neither side and never mentions the gap, by design, because a divergence between its two blind calls is meant to reveal whatever implicit sense of the gap the model already has.

Deliberation. Every call in the paper runs at the default setting, in which the model may deliberate internally but returns only the final answer.

Pairs without demographics. For the 1.0% of pairs whose source records no respondent attributes, the persona block is replaced by a single neutral opening. Respondent and Joint use `You are simulating an individual. No demographic details are provided.` and the conditioned designs use `Consider an individual.`

Reading the boxes. Each paradigm gets one box holding its whole prompt, in the style of Kolluri et al. (2025). Inside a box the prompt is divided into its parts by commented headers, and those headers are the only lines that are not prompt text, so `# ===== SYSTEM MESSAGE =====` and `# --- Persona ---` are annotations rather than bytes the model receives. No prompt in the roster contains a hash, so nothing in a box can be mistaken for a header. Deleting every commented line leaves exactly what was sent, which is the system message, and then the user message as the **persona** carrying the demographics, the **given** stimulus that is the already-known half in the conditioned designs, the **question** that is asked or produced, the **measurement-separation** statement where the pair has one, and the **response** format. In a template box the parts that vary by pair and by participant are set in *italic* between doubled angle brackets, as in `<<answer placeholder>>`, and a fully substituted worked example follows each template. Parts are separated by one blank line. The one exception is the conditioned designs, in which the answer format follows the ask directly with no blank line unless a measurement-separation statement intervenes. System messages do not vary by pair.

H.1 What varies across pairs, and what does not

The four paradigm skeletons below are fixed. Everything a pair can change is one of the slots in Table 8, each documented with a worked case in §H.6 through §H.8. Two facts about the dataset are worth stating first.

- **Every pair contributes exactly one "say" item.** No say side is a battery and none is averaged into a scale mean. Where the source instrument was a multi-item scale, the pair

names the single item it scores, so the organic-farming pair uses only "*I am willing to engage in organic agriculture*" and the COVID physical-activity pair only "*In the next 4 weeks, I plan to do more than 20 minutes of physical exercise at least 3 times a week*".

- **Multi-item asking survives only on the "do" side**, in the 14.3% of pairs whose action was measured as a set of separate acts and scored as a count or weighted sum (§H.7).

Table 8: The pair-varying slots, as percentages of all say–do pairs. One pair is 1.0%. Shares are of pairs rather than of prompts, and where a slot can be exercised on one side only the per-side shares are given in the right-hand column.

Slot	%	What changes
Persona lines	99.0	`- Attribute: value` lines drawn from that pair's own demographic columns. The remaining 1.0% ships no demographics and renders the neutral opening instead.
Item wording and tense	100	Source item text quoted verbatim, in the tense in which it was asked.
Respondent-voice restatement	16.3	A first-person restatement of the item, used only in the Respondent design (2.0 say side, 14.3 do side).
Response-option rendering	100	Explicit option list, inline anchored scale, bare numeric bounds, or nothing at all (§H.6).
Drop and renumber	30.6	An option the binarization excludes is hidden and the survivors are renumbered (19.4 say side, 11.2 do side, never both).
Answer-format placeholder	100	`<integer from lo to hi>` everywhere except the 1.0% whose legal values are a discrete set and are therefore named.
Multi-item behavior battery	14.3	The do side is asked act by act inside one call, with one JSON field per act (§H.7).
Measurement-separation statement	34.7	19.4 state an ordered interval plus a study note, 15.3 the note alone, and the rest say nothing (§H.8).
Per-occasion persona or age clause	15.3	The persona is re-rendered as of the occasion being asked about (14.3), and an explicit age is stated (3.1).

H.2 Respondent, simulating the individual

One system message serves both calls. Only the quoted item and its response block differ between them.

Respondent, the whole prompt template

```
# ===== SYSTEM MESSAGE =====
You are simulating a single individual. Answer each question as that
individual would, using the demographic profile and any other
information provided. Remain consistent with the individual's
perspective throughout the response. Follow the requested output
format exactly and do not include any explanation unless explicitly
requested.

# ===== USER MESSAGE =====
# --- Persona ---
You are simulating an individual with the following demographic
profile:
<<demographic profile, one "- Attribute: value" line each>>

# --- Question, answered as the individual ---
Answer the question below exactly as this individual would<< (answer
as this individual would have at age N)>>. Follow the response
instructions precisely.

Question:
"<<item, in its original first-person wording>>"

Response options:
<<one "code = label" line per surviving option>>
```

```
# --- Response format ---
Respond with a single JSON object and nothing else, in exactly this
form: {"answer": <<answer placeholder>>}
```

The age clause appears only for the pairs that record an age at that measurement (§H.8), and the `Response options:` block is omitted entirely for an unlabelled numeric scale (§H.6).

Worked example, voter turnout, statement call.

Respondent, worked example, statement call, user message

```
# --- Persona ---
You are simulating an individual with the following demographic
profile:
- Election year: 2022
- Age: 38
- Gender: Female
- Race: White
- Education: 2-Year
- Party identification: Republican
- Family income: 100k - 120k
- State: Kentucky

# --- Question, answered as the individual ---
Answer the question below exactly as this individual would. Follow the
 response instructions precisely.

Question:
"Do you intend to vote in the general election?"

Response options:
1 = Yes, definitely
2 = Probably
3 = I already voted (early or absentee)
4 = Plan to vote early
5 = I plan to vote before Election Day
6 = No

# --- Response format ---
Respond with a single JSON object and nothing else, in exactly this
form: {"answer": <integer from 1 to 6>}
```

Same participant, action call. Note the past-tense item and the different option set.

Respondent, worked example, action call, same participant

```
# --- Question, answered as the individual ---
Answer the question below exactly as this individual would. Follow the
 response instructions precisely.

Question:
"Did you vote in the general election?"

Response options:
1 = Voted
2 = No record of voting

# --- Response format ---
Respond with a single JSON object and nothing else, in exactly this
form: {"answer": <integer from 1 to 2>}
```

H.3 FORWARD, STATEMENT GIVEN, ACTION PRODUCED

The measurement-separation statement sits after both sides, immediately before the answer format, so that both of the measurements it relates are already on the page when the reader or the model reaches it.

Forward, the whole prompt template

```
# ===== SYSTEM MESSAGE =====
You are simulating a single individual. Answer each question as that
individual would, using the demographic profile and any other
information provided. For this specific task, simulate the action the
individual would actually be observed or recorded as performing, given
 what this individual says about their stated values, attitudes,
beliefs, or intentions. Keep in mind that humans sometimes demonstrate
 a gap between what they say and what they do (say-do gap). Remain
consistent with the individual's perspective throughout the response.
Follow the requested output format exactly and do not include any
explanation unless explicitly requested.

# ===== USER MESSAGE =====
# --- Persona, as of the action occasion ---
Below is a profile of an individual:
<<demographic profile, as of the ACTION occasion>>

# --- Given, the stated side ---
This individual <<was asked the following question | rated how much
they agreed with the following statement>><< (this was asked when they
 were N years old)>>, and gave the answer shown:
"<<say item, original wording>>"
<<scale clause>>
They <<answered | reported>> <<stated value, on the rendered scale>>.

# --- Produce, the action side ---
Based on this information, indicate how this person would answer the
following<< (this was measured when they were N years old)>>:
"<<do item, original wording>>"
<<scale clause>>

# --- Measurement separation, only if the two halves were not recorded
 together ---
<<fixed interval sentence, if the design is statement-first>>
<<study note>>

# --- Response format ---
Respond with a single JSON object and nothing else, in exactly this
form: {"answer": <<answer placeholder>>}
```

The context verb tracks the instrument. An agreement-anchored Likert item is introduced as "*rated how much they agreed with the following statement*", which covers 4.1% of all say–do pairs, and everything else as "*was asked the following question*".

Worked example, voter turnout.

Forward, worked example, voter turnout, user message

```
# --- Persona, as of the action occasion ---
Below is a profile of an individual:
- Election year: 2022
- Age: 38
- Gender: Female
- Race: White
- Education: 2-Year
- Party identification: Republican
```

```
- Family income: 100k - 120k
- State: Kentucky

# --- Given, the stated side ---
This individual was asked the following question, and gave the answer
shown:
"Do you intend to vote in the general election?"
using the following options:
1 = Yes, definitely
2 = Probably
3 = I already voted (early or absentee)
4 = Plan to vote early
5 = I plan to vote before Election Day
6 = No
They answered 1.

# --- Produce, the action side ---
Based on this information, indicate how this person would answer the
following:
"Did you vote in the general election?"
using the following options:
1 = Voted
2 = No record of voting

# --- Measurement separation ---
What this person said and what they actually did were not recorded at
the same time: what they said was recorded first, and what they
actually did was recorded a few weeks later.
The turnout intention was stated BEFORE election day. What they
actually did was established after the election from the official
record of who voted -- it is not their own later account of whether
they voted.

# --- Response format ---
Respond with a single JSON object and nothing else, in exactly this
form: {"answer": <integer from 1 to 2>}
```

H.4 Reverse, action given, statement produced

Reverse, the whole prompt template

```
# ===== SYSTEM MESSAGE =====
You are simulating a single individual. Answer each question as that
individual would, using the demographic profile and any other
information provided. For this specific task, simulate what this
individual says about their stated values, attitudes, beliefs, or
intentions, given the action the individual would actually be observed
 or recorded as performing. Keep in mind that humans sometimes
demonstrate a gap between what they say and what they do (say-do gap).
 Remain consistent with the individual's perspective throughout the
response. Follow the requested output format exactly and do not
include any explanation unless explicitly requested.

# ===== USER MESSAGE =====
# --- Persona, as of the statement occasion ---
Below is a profile of an individual:
<<demographic profile, as of the STATEMENT occasion>>

# --- Given, the action side ---
This individual was asked the following, and answered as shown<< (this
 was measured when they were N years old)>>:
"<<do item, original wording>>"
```

```
<<scale clause>>
They <<answered | reported>> <<observed value, on the rendered scale>>
.

# --- Produce, the stated side ---
Based on this information, indicate <<how this person would answer the
 following question | how strongly this person would agree with the
following statement>><< (this was asked when they were N years old)>>:
"<<say item, original wording>>"
<<scale clause>>

# --- Measurement separation, only if the two halves were not recorded
 together ---
<<fixed interval sentence, if the design is statement-first>>
<<study note>>

# --- Response format ---
Respond with a single JSON object and nothing else, in exactly this
form: {"answer": <<answer placeholder>>}
```

Reverse has the same shape as Forward with the two halves exchanged. The persona is rendered as of the statement occasion, the given part reports the action, and the ask is for the statement.

Worked example, voter turnout, the two parts that swap. The persona, the separation statement and the answer format are as in the Forward example above.

Reverse, worked example, voter turnout, the two halves that swap

```
# --- Given, the action side ---
This individual was asked the following, and answered as shown:
"Did you vote in the general election?"
using the following options:
1 = Voted
2 = No record of voting
They answered 1.

# --- Produce, the stated side ---
Based on this information, indicate how this person would answer the
following question:
"Do you intend to vote in the general election?"
using the following options:
1 = Yes, definitely
2 = Probably
3 = I already voted (early or absentee)
4 = Plan to vote early
5 = I plan to vote before Election Day
6 = No
```

Two asymmetries specific to Reverse. First, it is the only paradigm that shows a count-scored action as the *count* it was stored as, because here the action is a fact being reported rather than something being asked act by act. The climate-actions battery, asked as five separate acts everywhere else (§H.7), appears in Reverse as one composite item on its aggregate 0–5 scale.

Given, the action side, composite form

```
This individual was asked the following, and answered as shown:
"In your everyday life, how many of the following five actions do you
take in order to help protect the environment? Count = the number of
these five you do (0-5): (1) eat less meat; (2) drive less or carpool;
 (3) reduce food waste; (4) use fewer plastics that cannot be reused (
e.g. plastic bags, straws, cups); (5) reduce the amount of water you
use."
Measured as a number from 0 to 5 actions, they reported 4 actions.
```

Second, where the stored action is a per-item record rather than a single value, the given part lists every measured item beside its answer. One pair does this, and the same box shows two further variations at once, which are the agreement framing of the ask and the neutral opening standing in for a persona in the one pair that ships no demographics, here in the demographics-on arm.

Reverse, per-item action record, in the pair that ships no demographics

```
# --- Persona and given, the action side, per-item form ---
Consider an individual.

This individual was asked the following, and answered as shown:
"What is the intensity of your physical exercise?"
They answered:
- "What is the intensity of your physical exercise?": A low-intensity,
 less intense exercise
- "How many minutes do you spend on the above intensive sports
activities?": 11 to 20 minutes
- "How many times do you do the above sports activities in a month?":
Three to five times a week

# --- Produce, the stated side, agreement framing ---
Based on this information, indicate how strongly this person would
agree with the following statement:
"In the next 4 weeks, I plan to do more than 20 minutes of physical
exercise at least 3 times a week"
on a 1-6 scale (1 = Completely disagree; 2 = Strongly disagree; 3 =
Somewhat disagree; 4 = Somewhat agree; 5 = Strongly agree; 6 =
Completely agree)
```

H.5 JOINT, BOTH SIDES IN ONE GENERATION

Joint, the whole prompt template

```
# ===== SYSTEM MESSAGE =====
You are simulating a single individual. Answer each question as that
individual would, using the demographic profile and any other
information provided. For this specific task, simulate both, (i) what
this individual says about their stated values, attitudes, beliefs, or
 intentions, and (ii) the action the individual would actually be
observed or recorded as performing. Keep in mind that humans sometimes
 demonstrate a gap between what they say and what they do (say-do gap)
. Remain consistent with the individual's perspective throughout the
response. Follow the requested output format exactly and do not
include any explanation unless explicitly requested.

# ===== USER MESSAGE =====
# --- Persona, as of the statement occasion ---
You are simulating an individual with the following demographic
profile:
<<demographic profile, as of the STATEMENT occasion>>

# --- Produce (1), the stated side ---
(1) Answer the question below exactly as this individual would<< (
answer as this individual would have at age N)>>, reflecting on their
stated values, attitudes, beliefs, or intentions. Follow the response
instructions precisely.

    Say-question:
    "<<say item, original wording>>"

    Response options:
    <<one "code = label" line per surviving option>>
```

```
# --- Produce (2), the action side ---
(2) Answer the question below exactly as this individual would<< (
answer as this individual would have at age N)>>, reflecting the
action the individual would actually be observed or recorded as
performing. Follow the response instructions precisely.

    Do-question:
    "<<do item, original wording>>"

    Response options:
    <<one "code = label" line per surviving option>>

# --- Measurement separation, after both sides ---
<<fixed interval sentence, if the design is statement-first>>
<<study note>>
<<demographic-change block, panel pairs only>>

# --- Response format, both sides in one object ---
Respond with a single JSON object and nothing else, in exactly this
form: {"say-answer": <<answer placeholder>>, "do-answer": <<answer
placeholder>>}
```

Joint is the one paradigm that produces both halves in a single generation, so one persona cannot be correct for both occasions. The profile is rendered as of the statement occasion and each half carries its own age clause where the pair records one. For the same reason the measurement-separation statement matters most here, because without it the two halves would read as one simultaneous measurement. The two sides are labelled `Say-question:` and `Do-question:` and the answer object keys them by name rather than by position, so a model that emits an extra value cannot silently shift one side into the other's slot.

Worked example, degree expectation against attainment, chosen because it exercises both age clauses and the widest measurement interval in the roster.

Joint, worked example, degree expectation against attainment, user message

```
# --- Persona, as of the statement occasion ---
You are simulating an individual with the following demographic
profile:
- Sex: Female
- Race/ethnicity: Non-Black/Non-Hispanic
- Parent's highest grade completed (higher of mother/father): 13

# --- Produce (1), the stated side ---
(1) Answer the question below exactly as this individual would (answer
 as this individual would have at age 16), reflecting on their stated
values, attitudes, beliefs, or intentions. Follow the response
instructions precisely.

    Say-question:
    "What is the percent chance that you will have a four-year college
 degree by age 30?"

    Response options:
    1 = A 0-25 percent chance
    2 = A 75-100 percent chance

# --- Produce (2), the action side ---
(2) Answer the question below exactly as this individual would (answer
 as this individual would have at age 42), reflecting the action the
individual would actually be observed or recorded as performing.
Follow the response instructions precisely.
```

```
    Do-question:
    "Did you actually attain a four-year college degree (a Bachelor's
degree or higher)?"

    Response options:
    1 = Attained a Bachelor's degree or higher
    2 = Did not attain a Bachelor's degree

# --- Measurement separation, after both sides ---
What this person said and what they actually did were not recorded at
the same time: what they said was recorded first, and what they
actually did was recorded about 26 years later.
The expectation was stated in 1997, when the individual was a teenager
, and asked about their chances of holding a degree by age 30. Their
actual highest degree was recorded later, in adulthood, through 2023.

# --- Response format, both sides in one object ---
Respond with a single JSON object and nothing else, in exactly this
form: {"say-answer": <integer from 1 to 2>, "do-answer": <integer from
 1 to 2>}
```

Panel pairs and the demographic-change block. For the 12.2% of pairs drawn from two waves of the same panel, the single Joint persona is up to 28 months old by the time the model answers side (2). A block after the timing sentence says so and reports how the profile differed by the time the action was recorded, one line per attribute. Participants whose profile was stable get the same sentence with "none of it had changed" in place of the list, so silence never has to mean two different things. The lines are generated from the two snapshots each pair already declares, and a difference that is not a difference in the person is suppressed rather than printed. Three kinds are suppressed, which are a pair of columns that do not share a vocabulary, an attribute that cannot change at all such as gender or birthplace, and an attribute that can only move one way such as an age band or an earned degree. Unfiltered, the block would report gender changing for 78% of people.

Timing sentence plus demographic-change block, a panel pair

```
What this person said and what they actually did were not recorded at
the same time: what they said was recorded first, and what they
actually did was recorded about four years later.
The profile above describes this person as of when they answered (1).
By the time what they actually did was recorded, part of it had
changed:
- Age group: 50-64 (previously 30-49)
- Religious service attendance: Seldom (previously A few times a year)
```

H.6 Response options, what the model is actually offered

This is the slot with the most pair-to-pair variation and the one where a mistake is least visible, so it is specified exhaustively. Two independent choices are made, which options are shown and how they are rendered.

(a) Which options, drop and renumber. The analysis binarizes each side into say-yes against say-no and do-yes against do-no. Any option the binarization excludes, such as a neutral midpoint or an ambiguous partial category, is **not offered to the model at all**. It is removed and the survivors are renumbered $1..N$ with no gap. This happens on the say side of 19.4% of all say–do pairs and the do side of 11.2%, never on both sides of the same pair, and roughly two thirds of those sides end up with a displayed code that differs from the survey's native code. Because the displayed code space then differs from the native one, the runner maps every answer back to the native code before storing or scoring, and a participant's own already-known answer is mapped forward into the displayed space before it is shown as context in the conditioned designs.

Two consequences a reader must not miss. First, *the model never sees the source survey's native scale*. A three-way concerned-or-excited item about artificial intelligence is shown as a two-option choice, and a 1–10 justifiability ladder is shown as 1–6. Second, *a displayed code need not equal the native code*. The firearm-storage item drops the middle "some" category, so displayed `2 = None` is native code 3, and the food-waste intention item's displayed `3 = Agree` is native 4.

Drop and renumber, firearm storage, "some" is not offered

```
Question:
"Do you keep all, some, or none of your own guns in a locked place?"

Response options:
1 = All
2 = None
```

Drop and renumber, food-waste intention, the neutral midpoint is gone

```
Question:
"I intend to reduce food waste in the future."

Response options:
1 = Strongly disagree
2 = Disagree
3 = Agree
4 = Strongly agree
```

(b) How they are rendered. The Respondent design always uses an explicit `Response options:` list. The other three use a compact inline *scale clause* whose form follows the scale type, so the same pair can look different across paradigms. A labelled categorical renders as an options list everywhere, whereas a fully labelled Likert item renders as a list in the Respondent design and as `on a 1-4 scale (1 = Strongly disagree; 2 = Disagree; 3 = Agree; 4 = Strongly agree)` in the other three. Five forms occur.

1. **Fully labelled Likert**, where every point carries a label. Inline as `on a 1-4 scale (1 = A lot; 2 = Quite a lot; 3 = A little; 4 = Not at all)`, answered as `<integer from 1 to 4>`.
2. **Binary or labelled categorical**, inline as `using the following options:` plus the `code = label` list, answered as `<integer from lo to hi>`. Codes are whatever the survey used and are not necessarily 1-based, so one privacy pair renders `0 = No` and `1 = Yes` and asks for `<integer from 0 to 1>`.
3. **Anchored Likert**, where only the endpoints are labelled in the source and the interior points are genuinely answerable. The Respondent design lists the whole range and leaves the interior codes as bare numbers rather than inventing labels for them, and the other three show the endpoints inline as `on a 1-6 scale (1 = Can always be justified, 6 = Can never be justified)`. One pair.
4. **Sparse enumerated**, where the legal values are a discrete set rather than a contiguous range, so the answer instruction *names* them instead of stating bounds. One pair, a trust game with fixed token amounts.
5. **Unlabelled numeric**, where no options list exists because the source item is a bare count. No options block is shown in the Respondent design at all, and the other three show `on a 0-7 scale` inline. The aggregate scale of a count-scored battery is the related case, rendered as `Measured as a number from 0 to 5 actions` in the Reverse given part above.

Anchored Likert in the Respondent design, whole range listed, interior codes bare

```
Question:
```

```
"Overall, how strongly do you believe that dishonest acts like
cheating on taxes, accepting a bribe, or wrongly claiming benefits can
 never be justified?"

Response options:
1 = Can always be justified
2
3
4
5
6 = Can never be justified

Respond with a single JSON object and nothing else, in exactly this
form: {"answer": <integer from 1 to 6>}
```

Sparse enumerated, values named rather than bounded

```
Question:
"In a real-money trust game, you are given an endowment of 100 tokens
and may transfer any of them to an anonymous other person. How many of
 the 100 tokens do you send (0, 25, 50, 75, or 100 tokens)?"

Response options:
0 = 0 tokens
25 = 25 tokens
50 = 50 tokens
75 = 75 tokens
100 = 100 tokens

Respond with a single JSON object and nothing else, in exactly this
form: {"answer": <one of 0, 25, 50, 75, 100>}
```

Unlabelled numeric in the Respondent design, no options block at all

```
Question:
"In a typical week, how many days do you exercise?"

Respond with a single JSON object and nothing else, in exactly this
form: {"answer": <integer from 0 to 7>}
```

The design rule behind forms 3 to 5 is that **no offered value may be unanswerable and no answerable value may be unoffered**, so an option list must agree with the answer instruction printed directly beneath it. Both directions were live bugs. Listing only the justifiability ladder's two labelled endpoints under an instruction asking for `<integer from 1 to 6>` implied that 2 to 5 were invalid, and conversely describing the trust game's five legal token amounts as `<integer from 0 to 100>` invited the off-grid answers that models did in fact give and that no binarization rule can score. The same principle motivates part (a), because an option the analysis cannot score must not be on the menu or a model will pick it.

One arithmetically sparse composite. The organic-farming action is a weighted sum, at 5 for the CAS Organic label plus 3 for an environmentally friendly certification plus 1 for a traceability label, so its 0–8 range has two unreachable values. It renders as a labelled sparse scale that simply omits them, and the label of each reachable code names the certification combination it stands for.

Sparse composite scale, shown as Reverse context

```
This individual was asked the following, and answered as shown:
"Which organic-related farming certifications do you actually hold for
 your farm? Consider the CAS Organic Product Label, an environmentally
```

```
 friendly farming certification, and a traceability label. Select the
option that matches the certifications you currently hold."
On a 0-8 scale (0 = No certifications obtained; 1 = Traceability label
 only; 3 = Environmentally friendly farming certification only; 4 =
Environmentally friendly farming certification + traceability label; 5
 = CAS Organic Product Label only; 6 = CAS Organic Product Label +
traceability label; 8 = CAS Organic Product Label + environmentally
friendly farming certification), they answered 0.
```

H.7 Multi-item behavior batteries, asked act by act in one call

In 14.3% of all say–do pairs the action was measured as a set of separate acts and scored as a count or a weighted sum over them. Those are asked *act by act*, on the per-act scale, and never as a single "how many of these do you do" question, which is not how the behavior was measured. The per-act answers reduce back to the stored aggregate for scoring, so the ground truth is unchanged. Where a pair's action genuinely is a count question, it is asked as a count.

Every act of a battery goes in **one prompt** and is answered in one JSON object with one field per act. Each item is *named*, as `question-1:` through `question-N:`, carrying the side prefix `do-question-1:` on a labelled Joint side. The name is the model's hook from a question to the key that answers it, which a bare list index only implies, and it makes each line self-identifying. An earlier version of the pipeline asked one act per conversational turn. That mode was removed and the responses it had produced were archived rather than reused, so no battery in the data behind this paper is spread over several turns.

Respondent, an eight-act battery in one call

```
# --- Question, every act in one call ---
Answer each question below exactly as this individual would. Follow
the response instructions precisely.

Questions:
question-1: "In the last 12 months, did you contact a politician or
public official?"
question-2: "In the last 12 months, did you collaborate with a party
or citizen platform?"
question-3: "In the last 12 months, did you collaborate with another
organization or association?"
question-4: "In the last 12 months, did you wear or display campaign
badges, flags or stickers?"
question-5: "In the last 12 months, did you sign a petition?"
question-6: "In the last 12 months, did you take part in an authorized
 demonstration?"
question-7: "In the last 12 months, did you boycott or stop buying
certain products?"
question-8: "In the last 12 months, did you post or share something
about politics online?"

Response options (same scale for every question):
1 = Yes
2 = No

# --- Response format, one field per act ---
Respond with a single JSON object and nothing else, in exactly this
form: {"answer-1": <integer from 1 to 2>, "answer-2": <integer from 1
to 2>, "answer-3": <integer from 1 to 2>, "answer-4": <integer from 1
to 2>, "answer-5": <integer from 1 to 2>, "answer-6": <integer from 1
to 2>, "answer-7": <integer from 1 to 2>, "answer-8": <integer from 1
to 2>}
Give one answer per correspondingly-numbered question above, in the
same order.
```

In Forward the same battery becomes the ask, introduced by a plural lead line and followed by the shared option block.

Forward, a five-act battery as the ask

```
Based on this information, indicate how this person would answer each
of the following:
question-1: "In your everyday life, do you eat less meat in order to
help protect the environment?"
question-2: "In your everyday life, do you drive less or carpool in
order to help protect the environment?"
question-3: "In your everyday life, do you reduce your food waste in
order to help protect the environment?"
question-4: "In your everyday life, do you use fewer plastics that
cannot be reused (e.g. plastic bags, straws, cups) in order to help
protect the environment?"
question-5: "In your everyday life, do you reduce the amount of water
you use in order to help protect the environment?"
using the following options:
1 = Yes, I do this
2 = No, I do not do this
```

In Joint the battery is side (2), with each item carrying the `do-` prefix, and the answer object numbers only the side that has more than one question.

Joint, side (2) as a five-act battery

```
# --- Produce (2), the action side, act by act ---
(2) Answer each question below exactly as this individual would,
reflecting the action the individual would actually be observed or
recorded as performing. Follow the response instructions precisely.

    Do-questions:
    do-question-1: "In your everyday life, do you eat less meat in
order to help protect the environment?"
    do-question-2: "In your everyday life, do you drive less or
carpool in order to help protect the environment?"
    do-question-3: "In your everyday life, do you reduce your food
waste in order to help protect the environment?"
    do-question-4: "In your everyday life, do you use fewer plastics
that cannot be reused (e.g. plastic bags, straws, cups) in order to
help protect the environment?"
    do-question-5: "In your everyday life, do you reduce the amount of
 water you use in order to help protect the environment?"

    Response options (same scale for every question):
    1 = Yes, I do this
    2 = No, I do not do this

# --- Response format, one statement and five acts ---
Respond with a single JSON object and nothing else, in exactly this
form: {"say-answer": <integer from 1 to 4>, "do-answer-1": <integer
from 1 to 2>, "do-answer-2": <integer from 1 to 2>, "do-answer-3": <
integer from 1 to 2>, "do-answer-4": <integer from 1 to 2>, "do-answer
-5": <integer from 1 to 2>}
Give one answer per correspondingly-numbered question above, in the
same order.
```

So one battery has three renderings across the paradigms, which are a named per-act list asked in the model's own voice (Respondent), a named per-act list asked about a third party (Forward and Joint side (2)), and a single composite count question shown as already-known evidence (Reverse, §H.4). All three ask about the same acts and all three reduce to the same stored aggregate.

H.8 Measurement separation, ages, and per-occasion personas

For a prospective pair the delay between saying and doing is part of the quantity under study, so it is stated outright rather than left for the model to infer from two ages. Every paradigm that references both halves, which is Forward, Reverse and Joint, carries the statement once, after both sides and immediately before the answer format. The Respondent design carries none, because each of its prompts mentions a single measurement and so contains no interval to describe. The statement never names a survey, a wave or a session.

Just over a third of all say–do pairs say something, and the wording takes two shapes, because not every separated design is statement-first.

- **Ordered interval plus a study note**, in 19.4% of pairs. A fixed sentence asserts that the statement came first and names the interval, followed by a pair-specific note about how each half was obtained.
- **Study note alone**, in 15.3% of pairs. These are action-first designs, or same-sitting designs whose two halves refer to different periods. An organ-donor registration precedes the willingness item, a set of honesty norm items is a post-task questionnaire, and farm certifications took years to obtain before the survey, so asserting a statement-then-action ordering would misdescribe the study.
- **Nothing**, in the remaining 65.3%. Both halves were measured in one sitting and about the same period, so asserting a separation would be false.

Ordered interval plus study note, a 26-year interval

```
What this person said and what they actually did were not recorded at
the same time: what they said was recorded first, and what they
actually did was recorded about 26 years later.
The expectation was stated in 1997, when the individual was a teenager
, and asked about their chances of holding a degree by age 30. Their
actual highest degree was recorded later, in adulthood, through 2023.
```

Study note alone, an action-first design

```
What this person said and what they actually did happened close
together, in this order: they first faced a real organ-donor
registration choice and their decision was carried out in the state
registry, and only AFTERWARDS were they asked the willingness question
.
```

Study note alone, one sitting but two different periods

```
What this farmer said and what they actually did were recorded at the
same time, but they refer to different periods: the willingness
question is about what this farmer intends to do going forward,
whereas the certification question records what they had ALREADY
obtained by then -- certifications take years to obtain, so that
behavior precedes the stated intention.
```

Age clauses. In 3.1% of pairs the prompt also states the individual's age at a measurement. The Respondent and Joint designs splice `(answer as this individual would have at age 16)` into the instruction line, and the conditioned designs use `(this was asked when they were 16 years old)` on the statement side and `(this was measured when they were 42 years old)` on the action side.

Per-occasion personas. The demographic profile shows the time-varying attributes as they stood at the occasion being asked about, so the persona cannot contradict the age clause beside it. This affects 14.3% of pairs. Forward renders the persona as of the action occasion, Reverse as of the statement occasion, and Joint as of the statement occasion with each half's age clause and the demographic-change block (§H.5) supplying the rest. In the Respondent design each of the two calls renders the

persona as of its own side, so the same participant can appear one year older in the action call than in the statement call.

Respondent design, an age clause together with a reshaped option set

```
Answer the question below exactly as this individual would (answer as
this individual would have at age 16). Follow the response
instructions precisely.

Question:
"What is the percent chance that you will have a four-year college
degree by age 30?"

Response options:
1 = A 0-25 percent chance
2 = A 75-100 percent chance

Respond with a single JSON object and nothing else, in exactly this
form: {"answer": <integer from 1 to 2>}
```

That box also carries the roster's most heavily reshaped option set. A free 0 to 100 percent-chance item is offered to the model as a two-band choice, because those are the bands the binarization scores. Asking for a raw percentage and then scoring only its tails would have put most of the model's answer range outside the measure.

H.9 Querying each model in its native format

A fair comparison holds the *content* of every prompt fixed and lets each model receive that content in the interface it was trained on. Forcing a fine-tuned model into a format it never saw would measure its unfamiliarity with the format rather than its simulation of the individual, and would bias the comparison against exactly the models built for this task. Two of the human-behavior models are therefore queried in their own native formats, with the persona, the item wording and the option set left unchanged.

Centaur is trained on Psych-101 transcripts, in which a participant's choice is written as a keypress inside the running text. It is a completion model with no chat template and no JSON output, so its prompt is one plain-text body in the Psych-101 style. The system message opens the body as a leading paragraph, the persona and the item follow in the same wording every other model sees, each option is prefixed by a letter key instead of a numeric code, and the prompt closes with the cue `You press <<`. Centaur generates the next token, sampled at temperature 1.0, and that token is read as the chosen key and mapped back to the native survey code. It is offered the same option set as every other model, with the same options dropped, and a multi-item battery is asked one item per prompt.

Socrates is trained to return one short answer per turn, so on the 13 pairs whose "do" side is a multi-item battery it is asked one item per turn within a single conversation, with its earlier answers in context. Every other pair reaches it exactly as shown above. This is the same content as the batched question, delivered in the turn-by-turn shape the model was trained to answer.

Both models are trained to give one answer per generation, so the Joint design, which asks for two answers in a single generation, falls outside the task they were built for and is not run for them. The other eight models, OSim-8B included, receive the prompts exactly as shown in the boxes above.