ONLINE APPENDIX

A Retest Agreement and Predictive Accuracy

Consider one person who answers one categorical question on two occasions. Assume that the two answers are independent draws from a stable distribution $p = (p_1, \ldots, p_J)$ over the J possible answers. The probability that the two answers agree is

$$R = \sum_{j=1}^{J} p_j^2. \tag{A.1}$$

A predictor that knows p and always predicts the most likely answer is correct with probability $C = \max_j p_j$. Because $p_j^2 \leq C\, p_j$ for every j, and because one term of the sum equals C^2,

$$R \leq C \leq \sqrt{R}. \tag{A.2}$$

Under these assumptions, retest agreement is therefore a lower bound on the accuracy of an informed predictor and not an upper bound. The example $p = (0.8, 0.2)$ yields $R = 0.68$ and $C = 0.8$. If people and questions differ in their distributions, the same inequalities hold for the averages, $E[R] \leq E[C] \leq \sqrt{E[R]}$, by Jensen's inequality.

Three qualifications apply. The bounds refer to a predictor that knows the distribution of each person's answers, and a predictor with less information may not attain them. Dependence between the two occasions, learning, and changes in a person's circumstances violate the assumptions. And the derivation applies to the share of exactly matching categorical answers. It does not apply to scores that give partial credit for numerical answers, such as the accuracy score in Toubia et al. (2025). For these reasons, the main text reports accuracy and retest agreement side by side and uses the range from R to $\sqrt{R}$ as a reference.

B Selected Quantitative Benchmarks

Table B.1 compares selected benchmarks across targets, estimation procedures, positive results, and stress tests. We retain each study's original metric rather than pooling different measures.

Table B.1: A structured overview of reported model performance

Study	Target level	Human test	Reported performance	What it supports
Ashokkumar et al. (2026)	Direction and size	70 survey experiments; 469 effects	Predicted and observed effects correlate at $r = 0.85$ overall and 0.90 for studies not public by the cutoff. Effects are 1.7–1.8 times as large; $r = 0.25$ for small effects.	Strong ranking of broad effects; weak calibration and weaker fine comparisons.
Manning and Horton (2025)	Individual action in a defined domain	4,249 people; 1,490 strategic games	Agents assign 1.54 times the baseline-AI probability to observed play, 1.48 times cognitive hierarchy, and 1.38 times selected Nash equilibria. Their top-choice rate is 24.3%, below 28.5% and 30.4% for the latter benchmarks.	Theory and human training data help on one metric within a defined game family; the ranking is metric-specific.
Binz et al. (2025)	Individual choice	Held-out people and new tasks after training on 10.7 million choices	Centaur beats the base model and domain-specific cognitive models in almost every held-out experiment; results are reported as task-specific likelihoods.	Broad generalization can follow extensive training on human behavior.
Meister et al. (2025)	Population distribution	Human opinion distributions across several datasets	Directly verbalized distributions generally align better than sequential samples or raw token probabilities.	Forecasting a distribution and simulating respondents are different tasks.
Park et al. (2026)	Individual report and behavior	1,052 people; June 2026 manuscript	Interview agents reach 65.7% on survey items versus 58.1% for demographics and 79.5% retest agreement. Dividing by retest agreement gives the headline of 83%. Retest agreement is not a ceiling: under a simple model the attainable accuracy lies between 79.5% and 89.2%. Richer profiles do not improve predictions in economic games.	Personal information helps for survey answers; retest agreement is a yardstick and not a ceiling.
Toubia et al. (2025)	Individual report and behavior	2,058 people; 17 held-out tasks	The main specification reaches an accuracy score of 71.72%; the best reported configuration reaches 71.92%, versus 59.17% for the paper's guessing baseline and 81.72% for retest agreement. The evaluated items are withheld from the twins. Twins miss several behavioral effects.	Personal histories add signal, but predictions stay below people's own consistency.
Suh et al. (2025)	Population and subgroup distributions	3,362 questions; 70,000 subgroup–response pairs	Fine-tuning reduces the model–human distribution gap by up to 46% and improves tests on new surveys and groups.	Survey training improves survey prediction; it does not test incentivized behavior.
Iadisernia and Camassa (2025)	Aggregate reports	Four variables over 50 quarterly survey rounds	Model and human panels have similar forecast accuracy. Adding 2,368 persona descriptions gives no measurable gain, and model forecasts are much more homogeneous.	Timely context can support aggregate forecasts; persona diversity need not produce human diversity.
Huang et al. (2026)	Aggregate reports	Monthly consumer confidence in the United States, EU27, and Japan	ConsumerSim ranks first among the paper's reported persistence, time-series, regression, and information-augmented benchmarks.	Promising aggregate reconstruction; not an individual or revealed-behavior test.
Zhang et al. (2026)	Group effect and transfer	Regional field experiment	Raw LLM: -350 basis points; full human experiment: -60 ± 8; correction trained on 10% of regions: -65 ± 10; human data from these regions alone: -57 ± 27. Intervals are 95% bootstrap intervals. A correction trained on the first week yields -41.	A correction from human data carries over across regions, but not across weeks.
Baumann et al. (2025)	Downstream inference	37 labeling tasks; 2,361 hypotheses	Plausible configurations yield an incorrect conclusion for about 31–50% of hypotheses. Deliberate search can manufacture still larger error rates.	Model and prompt choice are part of the research design.

Notes. The main metric reported by each paper is retained. Correlations, exact-answer rates, distribution gaps, likelihood ratios, and effect estimates are not on a common scale and are not pooled. The complete screened evidence map appears in Appendix D.

C The Pipeline in Detail

This appendix contains the detailed account of the pipeline that Section 3 summarizes. It follows the same five stages and the two consequences, and it documents the sources for the technical statements about how models are trained and used.

We follow the chain stage by stage: the training data (Appendix C.1), pretraining (Appendix C.2), post-training (Appendix C.3), elicitation (Appendix C.4), which also covers how the sampled answers are aggregated into the estimate, and decoding and inference-time computation (Appendix C.5). Each subsection has the same three parts. The mechanism describes how the stage works technically; it is stated as established fact or measured regularity and sourced to the method papers, technical reports and studies that document it. The hypothesized link states what the mechanism would lead us to expect when a model stands in for people. It is written as predictions that the evidence can check. The evidence reports the relevant studies, mostly in the order of the predictions. Most are behavioral-simulation studies, and results from outside behavioral tasks are labeled as such. We separate two kinds of study. Some isolate a stage, for example by comparing saved versions (checkpoints) of one model before and after post-training. Others measure only what comes out at the end of the chain; their results can fit the link without showing which stage produced them. We give each step a name so the reader can carry it: selective recording in the training data, objective mismatch in pretraining and post-training, role mismatch in elicitation and decoding, and then two consequences, reduced form rather than structure, and shared error.

We are explicit about what kind of claim each part is. The mechanisms are documented facts about how models are built and regularities measured on trained models. The links are not: model training is largely opaque to outside researchers, so each link is our hypothesis about how a documented feature maps onto what simulations show, and establishing it would in most cases require more targeted research. The claim about the record is a conjecture about corpora that cannot be inspected for the closed models most studies use; we know of no audit of the open ones for their behavioral content. The claims about training are partly logical and partly supported by evidence that does not uniquely separate one step from another. Appendix C.7 grades each and says which experiments would tell them apart.

The opacity is concrete. Closed models disclose little about how they were built: the GPT-4 technical report withholds its architecture, training compute, dataset construction, and training method (OpenAI, 2023), and GPT-4 scores zero on data size and data sources in the Foundation Model Transparency Index (Stanford CRFM, 2024). What can be read depends on what was published. At the time of the cited studies, some closed models exposed no token probabilities (Anthropic's, for example), and OpenAI's exposed them only for the most likely tokens at each

position, 20 in the cited study (Meister et al., 2025; Jang et al., 2026). Base checkpoints exist only for some open families, and not always at the largest sizes (Binz et al., 2026). A few fully open families document every training stage together with its data (Team OLMo, 2025; Project Apertus, 2025).

Appendices C.1 to C.5 follow the pipeline. Appendix C.6 draws the two consequences of the chain as a whole, and Appendix C.7 places each recurring fact of Figure C.1 against the stages and grades the support for each explanation.

Figure C.1: Ten recurring facts

A. As the claim changes

1. Models often rank broad, familiar average effects correctly.
2. Their estimates of effect sizes are less reliable and often too large.
3. Accuracy falls for small effects and for comparisons between subgroups.
4. Synthetic samples understate variation within groups and distort differences between groups.
5. Predictions for individual people fall well short of how consistently people repeat their own answers, and they miss well-known behavioral patterns.
6. Matching what people report does not establish what they will do.

B. As the estimator changes or is repeated

7. With the assistant models most studies use, asking for a description of a distribution usually comes closer to the human data than simulating respondents and tallying their answers.
8. Richer personas do not reliably help. Measured personal histories help more than demographic labels, and mainly for survey answers.
9. Behavioral training data, explicit theory, and correction with human data improve performance within a defined domain.
10. A model's stated confidence, its token probabilities, and its repeated draws describe the model as it was run. More draws reduce simulation noise, not uncertainty about people.

Notes. These are recurring patterns in the literature, not laws. Section 2 gives the evidence behind each, and Table C.1 relates each to the stages of the pipeline.

C.1 From Behavior to the Record: Training Data

Mechanism. Before any model is trained, there is a record, and the record is not a sample of behavior. It is a sample of what people wrote, mixed in some training pipelines with text that other models wrote. Pretraining corpora are text. GPT-3 was trained on a filtered version of Common Crawl, an expanded WebText corpus, two internet-based book corpora, and English Wikipedia (Brown et al., 2020); GPT-4 was pre-trained to predict the next token in documents drawn from publicly available internet data and data licensed from third parties (OpenAI, 2023). Behavior therefore enters the record only as a written account of it. In some pipelines, model-written text is added on purpose: in the second training stage of OLMo 2, the mathematics component is mostly synthetic (Team OLMo, 2025).

What the model learns from is the training corpus as assembled: with near-duplicates collapsed, low-quality text dropped, domains reweighted, and a final phase of the same run, often called midtraining, in which the learning rate is annealed and the mix shifts toward higher-quality web text and specialized data. GPT-3's developers filtered Common Crawl with a classifier trained to recognize high-quality reference text, removed near-duplicate documents, and sampled the sources they judged higher in quality more often than their size alone would imply (Brown et al., 2020). In OLMo 2's midtraining pool, instruction data make up about 2 percent of the tokens, alongside filtered web text, academic papers, question-and-answer data, and mathematics (Team OLMo, 2025).

How the corpus was assembled can be checked only for fully open families (Team OLMo, 2025; Project Apertus, 2025). For GPT-4, for example, it is undisclosed (OpenAI, 2023; Stanford CRFM, 2024), so what the record contains about behavior is the part of the pipeline least open to inspection.

Behind this lies the size of the space the record would have to cover. A decision situation has indefinitely many features: who decides, what is at stake, which options are shown and how, under which rules, in which period, after which history. Behavior depends on combinations of these features, not on each in isolation. The same default works differently at different stakes, for different people, and with different alternatives on the screen. The record is a scattered set of points in this space, each from its own context, and most combinations have never been recorded. For an unrecorded combination a model has nothing to go on except what it associates with the description, which is why a setting that is familiar in description can be far in behavior.

Hypothesized link. Three kinds of information about behavior could appear in the record. There are descriptions of behavior: accounts, anecdotes, summaries of studies, people saying what they would do. There are measurements that link particular people to particular actions, the kind of data an experiment or an administrative file produces. And there are measurements that link the same features of people to their behavior under changed incentives or rules, the kind of data a policy evaluation produces. Language corpora plausibly contain a great deal of the first and far less of the second and third. The claim is not that behavior is absent from text; the internet contains transactions, reviews, and logs, as Horton (2023) rightly notes. The claim is that what appears is rarely linked. The record says a lot about what people say and about what was done, and very little that connects what a given person said, what incentives she faced, and what she then did. Two gaps result. The familiar say–do gap separates reports from actions. The do–who gap separates an observed action from the characteristics of the person and the situation that produced it.

Four features of the record matter for what follows. First, the person by situation by choice

grid is enormous, and only a small corner of it was ever written down. Even when every element of a description is common, their exact combination may be rare. Second, reports and accounts of behavior dominate. Even behavior usually reaches the record as somebody's written description of what happened. Text a model wrote about behavior is one step further removed from behavior than text a person wrote. Third, actions appear without systematic linkage to the traits, reports, and incentives of the person who took them, so the record can teach a model that attitudes and actions diverge in general (LaPiere, 1934) without teaching it where the gap falls for this person in this situation. A model could explain the sunk-cost fallacy perfectly and fail to exhibit it, because the text it learned from was written by people describing the bias from the outside. Fourth, accounts of policy changes and of how people responded exist in abundance, but linked before-and-after observations on the same people are thin.

These features lead to four predictions. Reports would be easier than incentivized actions because the record is made largely of reports. Broad averages could be recovered because broad regularities recur. Person and subgroup fidelity would be harder because the linkage that would teach it is scarce. And responses to changed rules would be poorly identified because the paired observations that would identify them are scarcest of all.

One further candidate belongs here. The record is selected toward notable effects: what gets written about is what was surprising or publishable. If a model learns effect sizes from such a record, it will learn them too large.

The record also contains the research literature itself. A model that "predicts" an experiment described in its training data is not predicting, and the contamination is partly conceptual rather than literal: the model may never have seen a particular study and yet have seen its effect described two hundred times. These are hypotheses about a record that cannot be inspected in full; Appendix C.7 says what would test them.

Evidence. No study isolates this stage: none varies what a model was pretrained on while holding everything else fixed, and the composition of closed corpora cannot be inspected. The evidence is therefore indirect.

The closest test adds the missing records rather than removing them. A model fine-tuned on 10.7 million human choices across 160 experiments predicts held-out participants and new tasks within the family it was trained on (Binz et al., 2025). On held-out tasks from a larger collection it also raises alignment with individual human responses ($d = 0.28$; Binz et al., 2026). Fine-tuning on survey data reduces the gap between model and human response distributions by up to 46 percent (Suh et al., 2025). Both results fit the claim that the record lacks linked data: the behavioral model's data link choices to individual participants, and the survey data link responses to demographic groups. Because the survey gains are on reports, they are consistent

with group-level linkage helping report prediction; they do not bear on whether actions are harder to predict than reports. Neither separates the record from the objective, however: fine-tuning adds the records and changes what the model is trained to predict at the same time, and what the behavioral model has learned is contested (Bowers et al., 2026; Schröder et al., 2025).

The interview-agent study is the only one we know of that holds the people and the granularity fixed and varies the outcome type: rich interviews that lift survey prediction from 58 to 66 percent add nothing detectable when the same people play five incentivized economic games (Park et al., 2026). This fits the say–do reading of the record but also has an elicitation reading (Appendix C.4); Appendix C.7 weighs them.

Results on subgroups are consistent with the link but do not isolate it. When models answer opinion surveys, their answers are closer to those of some demographic groups than to others, and which groups they resemble differs between models trained only on internet text and models tuned with human feedback (Santurkar et al., 2023). A model, in other words, does not speak for the population evenly. Asked to answer as a member of a particular group, models often sound more like outsiders imitating that group than like its members describing themselves (Wang et al., 2025). The authors trace this to the record: web text rarely says who wrote it, so a model has few chances to learn how members of a group actually speak. They tested only post-trained models, however, so post-training may also play a part.

The selection of the record toward notable effects is one possible reading of the 1.7 to 1.8 overshoot (Ashokkumar et al., 2026). We return to a second reading in Appendix C.3 and do not claim the evidence separates them.

That models retain their training data is documented. GPT-4o reproduces 344 of 10,000 congressional bill descriptions verbatim, and instructing it to use only information available before a date barely changes this (Ludwig et al., 2026). The correlation of 0.90 for experiments not public before the training cutoff rules out literal recall for those studies (Ashokkumar et al., 2026) but not conceptual contamination. A high correlation with published effects shows, at minimum, that the model knows the literature, which for a hypothesis is useful and for a replacement of the experiment is not.

The other two predictions are taken up elsewhere: broad averages in Appendix C.2, which reports the ranking success, and responses to changed rules in Appendix C.6.1.

C.2 From the Record to the Base Model: Pretraining

Mechanism. A language model is trained to predict the next token (a word or part of a word) in a corpus (Brown et al., 2020; OpenAI, 2023). In the ideal case it learns the distribution

of text conditional on text. Pretraining minimizes cross-entropy, and at the optimum, reached only with unlimited capacity and data, it reproduces the distribution in its data: if the text says option A 40 percent of the time, so does the model. Cross-entropy is a proper scoring rule: it is minimized when the model's probability equals the conditional frequency of the continuation in its data, not when all mass sits on the mode (Gneiting and Raftery, 2007). A pretrained probability therefore estimates how often a continuation follows a context of that kind in the training corpus as assembled (Appendix C.1).

Real models fall short of that optimum in a systematic way. The model condenses an enormous corpus into a system that predicts what comes next, and patterns that recur often survive that process better than rare details. The accuracy with which a model answers a factual question rises with the number of pretraining documents relevant to it, and facts that appear rarely are learned poorly even by large models (Kandpal et al., 2023); duplicated text, conversely, is memorized more readily (Lee et al., 2022).

The model that comes out of this stage is the base model. Pretraining is the one stage whose probabilities refer to the assembled corpus. Supervised instruction tuning refers them to a curated set of example answers, and preference optimization to no target distribution at all (Appendix C.3).

Hypothesized link. A pretrained probability is the right kind of number for the wrong distribution. Corpus continuation frequencies are not human choice frequencies, for the reasons of Appendix C.1. The distribution of text about behavior is not the distribution of behavior conditional on characteristics, incentives, and institutions. What pretraining adds to this is compression, and two consequences follow.

One possibility is that the model comes to rely on recurring group-level regularities rather than on the fine joint mapping that individual prediction requires. It would know how people of a given kind tend to behave, not how a particular person with a particular combination of traits behaves. On that reading, a persona prompt conditions the model on a description and returns something close to a group regularity, not a measured person. For combinations of traits and situations that the record never covered, the model must generalize from learned regularities rather than retrieve anything about the person, and the reliability of that generalization is exactly what needs separate validation.

The second consequence is that the mapping from a person's features to that person's behavior may simply not have been learned, in the way a regression on the same features would learn it. Features that predict what will be written about a situation need not preserve the mapping that predicts what will happen in it. Telling the model more about a person would then change its answers little.

The same fact, that what survives training is recurring regularity, should make the model a good expert, one that forecasts how people will respond rather than answering as one of them. An expert forecast draws on regularities that recur, which is what the model has. One further prediction follows from the mechanism itself. Because cross-entropy rewards honest frequencies, a base model asked the same question many times should reproduce the spread of answers in its text rather than repeat the most common one.

Evidence. The most direct evidence comes from base models, which only some open families release. Binz et al. (2026) score base and post-trained checkpoints of three open families on the probability they assign to the choices individual people made across hundreds of experiments. Telling them a participant's age, nationality, education, and questionnaire answers barely changes their fit (effects between $d = -0.02$ and 0.02, essentially none), in base and instruction-tuned models alike. The failure to use person-level information is therefore present before post-training, as both consequences predict. One caveat applies: the test transcripts include each participant's earlier responses in the session, so part of what base models get right is learned from linked data supplied in the prompt rather than from the corpus.

End-to-end results on post-trained models point the same way but cannot locate the failure. A two-hour interview buys no more predictive power for economic games than four demographic variables (Park et al., 2026). Twins hold several hundred of a person's answers in their prompt and remain about ten points below the consistency with which people repeat their own answers (Toubia et al., 2025). Each observation is consistent with a feature-to-behavior mapping that was never learned; each is also consistent with a representation or elicitation failure downstream.

The ranking success of Section 2.2.1, a correlation of 0.85 across 469 effects (Ashokkumar et al., 2026), fits the expert prediction, but it was measured on a post-trained model; it shows that the recurring regularities survive post-training, not that pretraining alone produces them.

The spread predicted by the mechanism is partly there. In three matched pairs, base models emulate opinion distributions that their post-trained versions estimate. That is, when a base model and its post-trained version each answer a survey question many times as individual respondents and the answers are aggregated, the base model comes closer to how people actually answered; when each is instead asked to state the distribution directly, the post-trained version comes closer (Grief-Albert et al., 2026). The pattern fits the distinction this section draws: pretraining leaves a model that can still act as a population, and post-training makes it a better expert but a worse respondent. The same matched pairs show compression at the level of groups: base models track the pattern of opinion differences between demographic groups but compress those differences to about 70 percent of their human size (Grief-Albert

et al., 2026). This fits the claim of Wang et al. (2025) that likelihood training flattens groups: cross-entropy reproduces frequencies only at an optimum that real models do not reach.

C.3 From the Base Model to the Assistant: Post-training

Mechanism. The deployed model is not the pretrained one. It has been tuned to produce outputs that human raters, AI raters, or automatic checkers score as helpful, correct, coherent, and safe (Ouyang et al., 2022; Bai et al., 2022a; DeepSeek-AI, 2025). Post-training is where probabilities tend to be pushed toward zero and one, by more than one mechanism. The tuning proceeds in steps, and each step has a different target: instruction tuning, preference optimization and, for reasoning models, reasoning training.

Supervised instruction tuning trains the model to reproduce a fixed set of example answers (Ouyang et al., 2022), so answers unlike the examples become less likely. The set is not released for GPT-4, for example (OpenAI, 2023). Because this step is again cross-entropy training, its probabilities refer to the example answers rather than to a corpus: a likely answer is now one that resembles the examples.

Preference optimization maximizes a learned reward subject to a penalty for drifting from a reference model. The reward is learned from pairwise preferences: raters compare answers and say which one they prefer (Christiano et al., 2017; Ouyang et al., 2022). The penalty keeps the tuned model close to the model it started from: in InstructGPT the instruction-tuned checkpoint (Ouyang et al., 2022); the original formulation anchored the penalty to the pretrained model (Ziegler et al., 2019). A stylized calculation shows what that does to a response distribution (Figure C.2). The optimum is the reference distribution tilted by the exponentiated reward, and no target distribution appears anywhere in the objective (Rafailov et al., 2023): nothing in the training asks the model's answers to match how often people give them. Preference optimization reweights: the optimum keeps every option's reference probability and multiplies it by a factor for the reward it earns. As the penalty weakens, the output concentrates on the highest-reward answer. Three qualifications matter for interpreting this result. It is a stylized result, not a description of any proprietary pipeline. It shows that post-training can shift and concentrate the distribution, not that it must; "flatten" would be the wrong word, since the change makes the distribution more peaked, not more uniform. And even a post-training that left the text distribution untouched would not have made it a human behavioral distribution.

Even where the options are equally helpful, raters' preference for typical answers (typicality bias) makes the reward favor what is already likely, so post-training sharpens toward the option the reference favors. Zhang et al. (2025) model the reward as true utility plus a bonus for answers the reference model already finds likely, a multiple α of its log-likelihood, estimate α

at 0.57 to 0.65 on a human preference dataset, and show that the optimum is then the reference distribution raised to a power $\gamma = 1 + \alpha/\beta$ greater than one, times the tilt from the true reward. Raising probabilities to a power greater than one, and rescaling them to sum to one, widens the gaps between answers: the most likely answer gains and the least likely loses. When the reward prefers a different option strongly enough, the favorite can change. Inverting this needs r and β, the reward and the weight of the penalty, together with the full deployed probabilities, and none is disclosed in full, so on a closed model the deployed distribution cannot be mapped back to the reference one.

Reasoning training adds a stage that sharpens harder and toward a different target: a checker that marks answers right or wrong, rather than annotators expressing a preference (DeepSeek-AI, 2025). Performance on such tasks is bought with entropy, the spread of the model's probabilities across answers, since the model is trained to converge on one answer (Cui et al., 2025). The term "reasoning" can suggest that these models deliberate as people do, but the training objective does not establish that correspondence. "Reasoning models" are post-trained toward verifiable answers, such as the result of a math problem, with more tokens available before the answer is given: the model writes out intermediate steps, called a chain of thought, before answering. A chain of thought can also be elicited by prompting alone, for example by including worked examples that spell out their steps (Wei et al., 2022). The name thus covers two different things, a training stage that rewards correct answers and a prompt that asks for steps, and neither is designed to reproduce how a person deliberates.

Hypothesized link. This step would help explain a cluster of facts that compression does not. If it does, post-trained models should match the choices of individual people less well than their base models do, for two reasons.

First, synthetic samples would be too homogeneous in part because the deployed distribution has been concentrated, on top of whatever homogeneity pretraining already leaves. Post-trained models should therefore give less varied answers than their base versions.

Second, simulated respondents would give the normatively tidy answer, the one a textbook would call correct, such as avoiding the sunk-cost fallacy and always maximizing, because the tuned model has been rewarded for answers that are correct rather than typical. In Horton's analogy, the good student knows the Nash equilibrium and plays it; most people do not; a good student is a bad simulation of the median subject.

Reasoning training should push further in the same direction, since its reward is correctness itself. On an opinion item there is no right answer to converge on, so a model trained this way should reproduce disagreement between people less well. If the effect comes from the reward rather than from writing out steps, a chain of thought obtained by prompting alone need not

Figure C.2: What pretraining and post-training do to a response distribution: a numerical example

Four answer options; the population answers them with probabilities $p = (0.40, 0.30, 0.20, 0.10)$. Suppose, generously, that the corpus reproduces these population frequencies. *Pretraining* minimizes cross-entropy, $-\sum_k p_k \ln q_k$, which is minimized at $q = p$: a model that has learned its data reproduces the frequencies in its data. *Post-training* maximizes $\sum_k q_k r_k - \beta \sum_k q_k \ln(q_k/\pi_k)$, a reward r learned from pairwise preferences minus a penalty for drifting from a reference model π; the optimum is $q_k \propto \pi_k \exp(r_k/\beta)$. No target distribution appears in the objective. With $r = (1.0, 0.6, 0.3, 0.0)$ and $\pi = p$, the deployed distribution is $(0.54, 0.27, 0.13, 0.05)$ at $\beta = 1$, $(0.80, 0.16, 0.04, 0.01)$ at $\beta = 0.3$, and a point mass on the first option as $\beta \to 0$. The example gives the highest reward to the option the reference already favors; with a flat reward the optimum would equal π. Typicality bias in preference data produces exactly this alignment (Zhang et al., 2025). *Temperature* T then raises every probability to the power $1/T$ and renormalizes; it flattens or sharpens the deployed distribution but cannot change which option wins or, in this example, recover the pretrained frequencies, and sampling many times at a given temperature recovers a number one could have computed from the probabilities directly. *Truncation* (top-k or top-p sampling) instead sets the probabilities below a cutoff to exactly zero before sampling; unlike temperature it cannot be undone, and providers do not always document their default cutoffs. Two caveats: the calculation is stylized, and the temperature identity is exact for a single answer token, not for a multi-token response or for an answer that follows a reasoning chain.

Notes. Our calculation. The reward-plus-penalty objective is the standard form of reinforcement learning from human feedback (Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022b; Casper et al., 2023); Rafailov et al. (2023) derive its closed-form optimum. In InstructGPT the reference model is the instruction-tuned checkpoint rather than the pretrained one (Ouyang et al., 2022); setting $\pi = p$ is a simplification.

show it. Neither kind of chain is a simulation of how a person deliberates, and a model that reasons is not thereby reasoning the way the described person would.

And magnitudes that are too large have a second candidate explanation here, mode-seeking toward the textbook effect, alongside the selection of the record in Appendix C.1. For both the tidy answers and the inflated magnitudes, the record and post-training are competing explanations; Appendix C.7 sets out what would separate them.

Evidence. This is the stage with the most direct evidence, because checkpoints of the same model before and after post-training exist for open families. Post-training makes models less human-like on behavioral tasks (Binz et al., 2026). Across three open families, instruction-tuned and reasoning-trained models assign less probability to the choices individual people made than their base models do (average $d = 0.11$ and 0.14); the gap has widened with each Qwen generation, from $d = 0.02$ to 0.16; it is largest in reasoning and psycholinguistic tasks; and more deterministic output does not explain it alone (Binz et al., 2026). Post-training can also change which option is favored (Binz et al., 2026). Post-trained models also exaggerate the opinion differences between demographic groups about twofold, where base models compress them (Appendix C.2; Grief-Albert et al., 2026). Llama 2's chat models, compared with their base versions, respond less like people to the wording changes that shift human survey answers

and more to changes that do not (Tjuatja et al., 2024). Models tuned with human feedback represent the public less well than models that were not, and one such model voiced a single dominant view on some items (Santurkar et al., 2023).

Outside behavioral tasks, the loss of variety can be measured directly. On the Tulu 3 checkpoints of Llama 3.1 70B, in a poem-writing task, output diversity already falls after instruction tuning and falls again after preference optimization (Zhang et al., 2025), and reinforcement learning from human feedback reduces output diversity relative to supervised tuning alone (Kirk et al., 2024). Alignment reduces conceptual diversity, the variety across simulated individuals in how they understand concepts (Murthy et al., 2025). That comparison is against instruction-tuned versions of the same models, which makes it the one behavioral result here that separates preference optimization from supervised tuning; no model, however, reaches human levels of diversity. All models in that study have 7 billion parameters, and four of the five aligned versions were tuned with AI rather than human feedback: the narrowing is not specific to human raters, though it is untested at larger scale. Instruction tuning makes models collapse to a single answer while leaving them able to describe the distribution they fail to sample; their base models sample imperfectly too, but fail far less often (Jang et al., 2026). Post-training therefore amplifies homogeneity rather than creating it.

Several intuitive biases that earlier models display disappear in later chat versions (Hagendorff et al., 2023), though the comparison spans model generations that differ in size and data as well as tuning, and the earlier models were themselves already tuned with human feedback. The twins also give the normatively tidy answer, reversing the sunk-cost fallacy and always maximizing (Toubia et al., 2025), but they were never run on a base model, so they are consistent with this stage without isolating it.

In the paired comparison above, reasoning-trained models depart further from people than instruction-tuned ones (Binz et al., 2026). Models trained toward verifiable answers have been reported to reproduce disagreement less well: such training degrades the prediction of how human annotators disagree, whereas prompting for a chain of thought improves it (Ni et al., 2026).

The inflated magnitudes (Ashokkumar et al., 2026) likewise come from a post-trained model without a base-model comparison, so they are compatible with this stage but do not test it.

The paired evidence also shows that post-training as such is not the problem: post-training on human choices, as in Appendix C.1, raises alignment instead of lowering it (Binz et al., 2026). Because that training changes the data and the target together, it does not say which of the two matters (Appendix C.1). Section 2.4.1 reviews the parallel evidence that post-training reshapes stated confidence.

C.4 From the Assistant to the Prompted Model: Elicitation

Mechanism. The last step is the researcher's. A deployed model becomes a particular estimator through the prompt and persona, the reasoning protocol, the response format, the temperature and decoding rule (Appendix C.5), the number of repeated generations, and the way answers are parsed and tallied. Parsing and tallying come after decoding in the pipeline; we treat them here because, like the prompt, they are the researcher's choices. Each of these choices matters: changes to prompt formatting that leave its meaning intact can move measured accuracy by up to 76 percentage points, and which format works best correlates only weakly across models (Sclar et al., 2024). This is where a conditional text generator is asked to be a person. A persona is a description placed in the prompt, which the model conditions on before predicting what comes next, in the same way as task instructions and worked examples (Brown et al., 2020). Repeated synthetic respondents share one model, one prompt, and one set of blind spots; they are not independent people.

Further facts concern the measurement rather than the model. Option order, question wording, how an answer is parsed, and how refusals are coded change the recorded number (Pezeshkpour and Hruschka, 2024; Zheng et al., 2024; Tjuatja et al., 2024; Dominguez-Olmedo et al., 2024; Wang et al., 2024). Models prefer particular option labels, a bias traced to the prior they place on the label tokens themselves, whatever option a label marks (Zheng et al., 2024), and for instruction-tuned models the answer implied by the first token's probability, which studies often read as the model's answer, and the answer they actually write can disagree in over 60 percent of cases (Wang et al., 2024). Averaging over option orders reduces order effects, and scoring the generated answer avoids a documented mismatch between first-token probabilities and the text the model actually gives; on current models, reasoning models included, these steps do not remove the effects (Zheng et al., 2024; Wang et al., 2024, 2025; Wang, 2026).

Hypothesized link. A conditional text generator asked to be a person does not thereby become a draw from the population of that kind of person. A persona makes the model's answer more coherent with the description; it does not make the answer a sample from the people the description fits. Supplied measurements can go unused.

This step would help explain the rest of Panel B of Figure C.1. If it does, describing a distribution should beat sampling it, because the description uses the model as an expert while the sample asks it to be a population it does not represent. The reading predicts a larger gap for post-trained models, whose sampling has been concentrated (Appendix C.3), than for base models. A direct forecast and a tally of simulated respondents would differ because they are different estimators, not because one of them found the truth. Richer personas need not help,

because coherence is not accuracy. And because synthetic respondents differ only in what their personas say, a synthetic sample should vary less than the people it stands for. The same logic reaches outcome types. For the model, an incentivized game arrives as a text vignette, and its answer is a report either way; it never faces the stakes, so matching what people report may be all that elicitation can deliver, whatever the human outcome type (Section 2.2.2).

There is a broader observation here. Free-form generation is what these models are distinctively good at; respondent simulation asks instead for calibrated human frequencies in a fixed response space. The most credible prediction results in Section 2 concern bounded estimands or domains, such as the ranking of effects, and population-faithful respondent generation demands a far richer match than any of them. Conclusions drawn from a simulation should also shift with wording and readout choices that leave the question's meaning unchanged.

Evidence. Holding the model fixed and changing only how it is asked isolates this stage, and such comparisons are common. Asked to describe a population's distribution of answers directly, a model usually comes closer to the human distribution than when it generates many individual answers that are then tallied (Meister et al., 2025), although the sampling arm in that comparison generates a sequence of answers in one output rather than independent calls. Instruction-tuned models fail to sample from distributions they can correctly describe, and describing the distribution in one call more than halves the error against human survey data relative to persona aggregation, averaging answers across many personas (Jang et al., 2026). A prompt that asks for a verbalized distribution rather than a single answer recovers part of the lost diversity (Zhang et al., 2025). Asked to forecast 12 effects directly, GPT-4 gets the sign right for 10 of them, judged against the effects produced by the same paper's simulated respondents, but its forecasts are on average 13.2 times those effects, or 5.3 times after removing the largest discrepancy (Manning et al., 2024). Consistent with the predicted interaction, the advantage of describing over sampling appears in post-trained models and reverses in matched base models (Appendix C.2).

Richer personas do not reliably help. Interviews help on survey items but not on games (see below). Twins that hold several hundred of a person's answers remain well below that person's own consistency (Toubia et al., 2025), and adding 2,368 persona descriptions to a macroeconomic forecasting exercise gives no measurable gain while the model's forecasts stay far more homogeneous than the human panel's (Iadisernia and Camassa, 2025). Across hundreds of experiments, adding participant information to the prompt fails to improve individual-level prediction, in base and instruction-tuned models alike (Binz et al., 2026); since base models fail as well, this fits the elicitation reading but cannot be told apart from a mapping never learned in pretraining (Appendix C.2).

For outcome types the evidence is thinner. Interviews beat demographics on survey items and not on games (Park et al., 2026), which fits the stakes reading, but no study yet separates the elicitation reading of the report–action gap from the record-based one (Appendices C.1 and C.7).

How the model is asked and read out changes the conclusions. A prompt that fails to say the treatment was randomly assigned lets the model invent pre-existing differences between the arms (Gui and Toubia, 2025). Across 37 labeling tasks, ordinary choices of model, wording, answer order, and temperature yield a wrong qualitative conclusion for about 31 percent of hypotheses with frontier models and 50 percent with smaller ones (Baumann et al., 2025). Models' survey answers are governed by ordering and labeling biases, differences between models largely vanish once these are adjusted for, and which groups a model appears to represent is predicted by how close each group's answers are to uniform, irrespective of the model's training (Dominguez-Olmedo et al., 2024). Applying a model's default chat template, the standard formatting a chat model expects around a conversation, lowers its fit to human responses (Binz et al., 2026). Because the recorded number also depends on how answers are read out, validation attaches to the whole configuration rather than to a model name (Section 2.2.3).

C.5 From the Prompted Model to a Synthetic Population: Decoding and Inference-Time Computation

Mechanism. Decoding is the step that turns the model's probabilities into an actual answer. Temperature is an invertible rescaling of the deployed distribution: it divides the logits by T before the softmax[10] (Ackley et al., 1985; Holtzman et al., 2020). In effect, it raises every probability to the power $1/T$ and rescales them to sum to one (Figure C.2): a temperature above one flattens the distribution, and one below one sharpens it. With T known, draws and probabilities are interconvertible, and the ranking and the mode are preserved at every T. Truncation is not invertible. Top-k and top-p sampling[11] discard the unlikely options before drawing and renormalize the rest (Fan et al., 2018; Holtzman et al., 2020), so an option below the cutoff appears with frequency exactly zero however long one samples, while its probability remains a positive number. Which answers are possible at all may then be set by decoding settings, and providers do not always document their default cutoffs. No decoding setting is guaranteed to undo post-training: supervised tuning and shifts in the favored answer are not power transforms, and even where post-training acts like temperature scaling (Zhang et al., 2025), the exponent that would undo it is unknown on a closed model.

[10] Logits are the model's raw scores for each possible next token; the softmax turns them into probabilities.

[11] Top-k sampling keeps only the k most likely options; top-p sampling keeps the smallest set of options whose probabilities add up to at least p.

When the model generates a hidden chain of reasoning before its answer, the probability one can read on the answer token, the model's probability for its final answer, is $p(A \mid c, \text{prompt})$ for the one chain c that was sampled. The quantity of interest is the mixture over all the chains the model could have written, each weighted by how likely it is,

$$p(A \mid \text{prompt}) = \sum_{c} p(c \mid \text{prompt})\, p(A \mid c, \text{prompt}),$$

and a single call does not provide it. Repeated generation marginalizes over chains and is therefore the right estimator of the deployed answer distribution, even where probabilities are available (Wang et al., 2023; see also Wei et al., 2022).

Hypothesized link. Raising the temperature transforms the model's output distribution, but it cannot establish that the added spread corresponds to differences across humans. Because temperature preserves the ranking and the mode, a biased mode stays the mode and no setting recovers a between-person distribution the model does not hold. A higher temperature may bring the overall spread closer to people's, but it should not correct a biased mode or make the added variation correspond to differences between people.

It helps to keep two kinds of human variation apart. Heterogeneity is stable variation between people. Noise is variation within a person, from changing states, trembles, inattention, and measurement error. A model's stochasticity comes from decoding, from the reasoning chain, and from arithmetic on shared servers.[12] Nothing in the training objective supplies a mapping from these sources to either kind of human variation, so the spread in a synthetic sample need not match either the differences between people or the noise within them.

Truncation sets a harder limit. If the cutoff removes the options that people choose rarely, a synthetic sample will never contain the rare answers that make up a population's tails, however large the sample.

We expect the chain written before the answer to change which answer comes out on a given call more than how varied the answers are across calls, with the range of answers set by the model. Appendix C.3 explains why neither a trained nor a prompted chain is a person's deliberation.

Evidence. Holding the model fixed and changing only the decoding setting or the chain isolates this stage. At the default temperature, a sizable share of classic experimental paradigms return nearly identical answers across a thousand draws (Park et al., 2024), but the model

[12] Each output is a sum of thousands of products accumulated in chunks, the grouping of the chunks depends on the traffic sharing the server, and rounding makes the result depend on the grouping (He and Thinking Machines Lab, 2025).

tested was already tuned with human feedback, so decoding and post-training are confounded. At temperature 1.0, GPT-4o gives the same answer in at least 98 percent of calls when asked to sample from a stated distribution, and the failure persists across temperature and top-p settings (Jang et al., 2026). In a poem-writing task, that is, outside behavioral simulation, raising the temperature trades quality for diversity, and asking for a verbalized distribution improves that trade-off at every temperature tested (Zhang et al., 2025). Populations generated by varying temperature and prompts still fall short of human conceptual diversity (Murthy et al., 2025). Post-trained models are also worse than their base models at predicting the single most likely human choice, so their misalignment is not a matter of determinism that a higher temperature could undo (Binz et al., 2026).

We know of no study that maps a model's sources of randomness onto either kind of human variation. For truncation's effect on rare answers there is no behavioral-simulation evidence, and the link remains untested.

For reasoning at inference, the evidence runs both ways, and no study tests directly whether the chain changes which answer comes out more than the range of answers. Switching the chain off does not restore human-like errors: chat models told to answer without explaining their reasoning stayed largely accurate on tasks built to elicit intuitive mistakes (Hagendorff et al., 2023), consistent with the weights, not the chain, setting the answer; the study used a temperature of zero and measured accuracy, not the spread of answers. On multiple-choice questions, position bias (a pull toward options in particular places on the list) grows with the length of the reasoning chain, and chain-of-thought replaces the position bias of direct answers rather than removing it (Wang, 2026). Appendix C.3 reports how prompted and trained chains differ in predicting human disagreement.

C.6 Consequences of the Pipeline

Two consequences follow from the pipeline as a whole rather than from any one stage. The first concerns what a match between model and people establishes (Appendix C.6.1); the second concerns what repeated model output can and cannot reveal about the error (Appendix C.6.2).

C.6.1 Why Local Matches May Not Travel: Reduced Form, Not Structure

The first consequence of the chain concerns the distance between the validated estimand and the target (Section 2.1). It is not a further step in the production of an answer; it follows from what the steps produce. If a model's local accuracy comes from familiar textual regularities, from memorized study patterns, or from reduced-form correlations in the record, then a match between model and people in one setting does not show that the behavioral relationship pro-

ducing the match has been recovered. And if it has not been recovered, nothing guarantees the match will hold somewhere else. Local outcomes validate an estimator locally. Only recovered structure or demonstrated stability warrants extrapolation. In the terms of the pipeline, the record supplies the regularities (Appendix C.1), pretraining learns them as text given text (Appendix C.2), and no later stage of the standard pipeline adds that structure; training on linked behavioral records adds some, but only within the family those records cover (Appendix C.1 and Sections 2.3 and 2.2.3).

Two pieces of evidence make this concrete, alongside the conceptual contamination discussed in Appendix C.1. The first comes from settings where the underlying world can be inspected. Vafa et al. (2024) train transformers on routes through Manhattan. The models produce valid routes 96 to 99 percent of the time, yet maps reconstructed from their predictions contain streets that do not exist, and a small random detour drops one model's valid-route rate from 99 to 8 percent. A transformer trained on orbital trajectories predicts the next position with $R^2 > 0.9999$ and yields a nonsensical force law when asked to generalize the physics (Vafa et al., 2025). These are existence proofs from checkable domains, not behavioral evidence, and we use them as an argument: excellent in-domain prediction does not establish that the relationship needed after a change has been learned. The operative question for any application is not whether the model contains a world model in some general sense. It is whether the particular relationship responsible for its match at the validated estimand remains stable under the proposed change of setting.

The second is the Lucas critique itself. Rules that people respond to change the relationship between what can be observed about people and what they do (Lucas, 1976). A model built from the record carries the literature's story about how people adapt, and where people deviate from the literature, which is precisely where experiments earn their keep, the model is likely to reproduce the textbook rather than the people. Horton's reply that language agents can reason flexibly about a changed environment is correct as far as it goes; reasoning about a change is not the same as responding to it the way people will. Dell and Rambachan (2026) state the conditions under which a correction learned in one setting can be carried to another: outcome stability, under which people with the same model output and observed characteristics have the same expected outcome in both settings, and measurement stability, under which the model produces the same distribution of outputs for a given human outcome in both. Either justifies a transport; neither can be checked from model output alone. Behavior under endogenous interaction, where what one person does depends on what others do, raises the same problem.

What does travel is consistent with this account. The successes of Section 2.3 are cases in which the validation evidence spans many settings of one kind and the target is one more setting of that kind: held-out experiments from a familiar family, expectations quarter after

quarter, corrections across regions. These are cases of transfer over a short distance, and they are uninformative about settings that the evidence never covered. Setting means the same thing here as in Section 2. This appendix asks only whether the record and the learned model contain variation informative about crossing it, and the answer is that they do where the record varied the condition and do not where it never did or where the condition is a rule that changes responses.

C.6.2 Why Repetition Does Not Reveal the Error: What the Numbers Are

The second consequence concerns Section 2.4, and it is a matter of research design as much as a property of the model. Every step above produces one deployed model. Repeating its answer reduces Monte Carlo noise about that model's own output distribution and leaves its systematic error untouched. A thousand draws are not a thousand people, because the errors are shared across draws: same model, same prompt, same blind spot. That much follows from Equation (1) alone. What the chain adds is an account of what the numbers read off a model are, stage by stage, and therefore of why none of them can be a standard error for people.

Appendices C.2 to C.5 give that account, and we collect it here. A pretrained probability estimates how often a continuation follows a context of that kind in the corpus as assembled (Appendix C.2). A post-trained probability refers to a set of example answers or to no target distribution at all, and on a closed model it cannot be mapped back to the reference model (Appendix C.3). The recorded number also depends on how the answer is elicited and read out (Appendix C.4). A draw at a known temperature is an invertible transform of the deployed distribution, a truncated draw hides the options below the cutoff for good, and with a hidden reasoning chain a single call conditions on one chain that nobody sees (Appendix C.5). Even identical inputs need not return identical probabilities on a hosted service (He and Thinking Machines Lab, 2025).

Consequence. Every number read off a model describes the specific thing that was run: this checkpoint, this prompt, these decoding settings, this chain. Draws at a fixed temperature estimate the deployment's own distribution consistently, and that distribution is a real object worth knowing. Post-training removes the guarantee that it matches the pretrained one, and the strength of the penalty that determined how much was removed is unknown and, on a closed model, unrecoverable. That opacity complicates diagnosis. It is not the fundamental reason inference fails, since even a fully transparent model would need a design linking its predictions to the human target. The distribution's mean carries the systematic error b, its spread is the spread of one procedure, and the model's stated confidence is one more of its outputs. None of these measures b or the uncertainty about it. Where there are no human outcomes

there is no estimate of the error and no measure of its uncertainty, and no amount of model output changes either.

C.7 Which Stage Explains Which Fact

Table C.1 closes the appendix by placing each of the ten facts of Figure C.1 against the stages of the pipeline and its two consequences, with more than one candidate allowed per fact and a grade for the kind of support each has. We distinguish five grades. A claim is *logical* when it follows from what the training objective does and does not target. It is *direct empirical* when a benchmark shows it, as with persona failures, direct-versus-respondent gaps, and synthetic homogeneity. It is a *mechanistic conjecture* when it concerns the composition of corpora or what compression keeps. It is an *identification consequence* when it follows from the structure of the evaluation problem, as with repetition not estimating human-relative error. And it is a *transport hypothesis* when it predicts where failures should be largest, namely where rules change the response relationship.

Each cell of the table also records how the evidence bears on a stage. Evidence is isolating when it separates that stage from the others: paired base and post-trained checkpoints of one model, one model under two estimators (a different prompt, readout or decoding setting, or with the chain switched on or off), or fine-tuning with the architecture held fixed. Because base models may already have seen some instruction data (in OLMo 2's midtraining pool, about 2 percent of the tokens; Appendix C.1), paired checkpoints may understate what post-training does. It is consistent when it fits the stage end to end without separating it. A candidate is a stage the argument points to without evidence that separates it. We say an effect is amplified by a stage when paired evidence shows it already present before that stage, and located in a stage only when it is absent before.

A few cases show how the attribution works in practice. Synthetic homogeneity (fact 4) has four candidate stages: compression in pretraining, concentration in post-training, respondent-mode elicitation, and default decoding. Pretraining already produces some of it, since base models also fail to sample distributions, though far less often (Appendix C.3); the paired evidence gives post-training an amplifying role rather than an originating one (Appendix C.3); elicitation fits without a test that separates it (Appendix C.4); and changing the decoding settings does not remove it (Appendix C.5).

Two stages compete for the normatively tidy answers of the twins (Appendix C.3). Post-training rewards answers that are correct rather than typical (Appendix C.3), and the record describes the bias from the outside (Appendix C.1). The paired evidence shows that post-trained models depart most from people in reasoning and psycholinguistic tasks, and reasoning-trained

Table C.1: What explains each recurring fact, stage by stage, and how well

#	Fact (coordinate or test)	C.1 Data	C.2 Pretraining	C.3 Post-training	C.4 Elicitation	C.5 Decoding	Non-stage	Grade
1	Broad rankings often right (granularity)		Consistent	Consistent			Comparisons difference out shared error	Direct empirical; logical
2	Magnitudes too large (granularity)	Candidate		Candidate				Conjecture; the evidence does not separate the two
3	Small effects and subgroups fail (granularity)	Consistent	Consistent: present in base models (compressed)	Isolating: exaggerates group differences	Consistent			Direct empirical for the failure; isolating for post-training's effect on group differences; conjecture for the rest of the cause
4	Synthetic samples too homogeneous (granularity, response distributions)		Consistent	Isolating: amplifies	Consistent	Isolating: settings do not remove it		Direct empirical; amplified by post-training
5	Individual scores fall against the right benchmarks (granularity, units)		Consistent: present in base models		Candidate			Direct empirical for the symptom; underidentified cause
6	Reports do not establish actions (outcome type)	Candidate			Candidate			Conjecture; one within-sample test
7	Describing beats sampling (estimator)			Isolating: see interaction row	Isolating	Candidate		Direct empirical, for post-trained models
8	Richer personas do not reliably help (estimator)		Candidate		Isolating			Direct empirical; underidentified cause
9	Behavioral data, theory, correction help (estimator)	Consistent			Consistent		Each supplies what the chain removed	Direct empirical within a domain
10	Confidence, token probabilities, and draws describe the deployment (inference)			Isolating, for confidence		Logical	Identification consequence (Appendix C.6.2)	Identification consequence; logical
×	Interaction: post-training × elicitation (fact 7)			Interacts	Interacts			Direct empirical

Notes. Grades: logical, direct empirical, mechanistic conjecture, identification consequence, transport hypothesis. The transport hypothesis applies to every fact that is carried beyond the settings that the evidence covers and is not a row of its own. Cells: isolating, consistent, or candidate, as defined above; an empty cell means the stage is not implicated. Fact 9 rests on the evidence of Section 2.3 (Manning and Horton, 2025; Zhang et al., 2026).

models most of all (Appendix C.3), but it does not yet show the direction of that departure, and the twins were never run on a base model. Both stages therefore remain candidates, and the evidence does not separate them.

The overshoot of 1.7 to 1.8 (fact 2) also has two candidates: the selection of the record toward notable effects (Appendix C.1) and mode-seeking toward the textbook effect in post-training (Appendix C.3). We know of no study that forecasts effect sizes from a base model, so the evidence does not separate them.

Describing a distribution beats sampling it in post-trained models (fact 7), but the advantage reverses in matched base models, and instruction tuning amplifies the failure to sample (Appendices C.3 and C.4). The fact is therefore an interaction between post-training and elicitation rather than a property of either stage. The failure to use supplied measurements (fact 5), by contrast, is already present in base models (Appendices C.2 and C.4), which rules out post-training as its only source; whether the mapping was never learned in pretraining or is learned but not used under elicitation remains open.

Paired checkpoints help distinguish training stages, and behavioral fine-tuning tests added data while also changing the objective; both comparisons exist in some settings (Binz et al., 2026; Grief-Albert et al., 2026; Binz et al., 2025). Supplying identical information through prompts and training would test elicitation against learning, but remains untested to our knowledge. Comparing forecasts and respondent simulations tests the two estimators. Preregistered tests at increasing distance from validation assess transport; coverage against human outcomes assesses inference. Some comparisons require open models.

D Evidence Map and Search Protocol

This appendix records the search and screening procedure used to map the evidence. Its purpose is to make the review reproducible and to show exactly what the count of 108 reports can support. The search was frozen on August 31, 2026. It had no lower publication-date limit. The protocol was not prospectively registered. One screening pass was conducted with computational assistance, so this process does not provide the error protection of two independent reviewers. We therefore use the search to describe the coverage of the field. Quantitative claims in the main text come from separate checks of the cited original papers.

D.1 Review question and eligibility rules

The review question was: *When a generative language model is used to produce, predict, or reproduce a human response, how closely does its output match observed human reports, choices, behavior, or treatment effects?*

A report was included if it met all of the following rules:

1. It studied a generative or large language model, including a language model used inside an agent, persona, or specialized behavioral model.
2. The target was a response by a person or population: a report, opinion, belief, judgment, choice, action, distribution, or treatment effect.
3. It compared model output with observed human data. The human comparison could come from a survey, experiment, field record, or held-out behavioral dataset. Data used only to build a persona or train the model did not count as an independent test unless the paper also reported a held-out human comparison.
4. It reported an empirical result in a complete article, conference paper, or working paper available in English by the freeze date.

We excluded records that studied a non-generative prediction method; simulated agents without an observed human benchmark; model behavior with no human target; text classification or economic sentiment measures that did not claim to simulate respondents; clinical or technical performance against experts rather than a human behavioral target; reviews, editorials, protocols, and method proposals without an eligible empirical test; abstracts or data deposits without a complete report; non-English reports; and superseded versions of the same study. These exclusions concern the screened evidence set, not whether a paper can be cited as background elsewhere in the review.

D.2 Databases, date, and exact search strings

We searched OpenAlex and arXiv on August 31, 2026. OpenAlex searches were run through the Works API in title and abstract fields. For each term block below, the exact filter was

`title_and_abstract.search:`t_1`,..., title_and_abstract.search:`t_k`,`
`from_publication_date:1900-01-01, to_publication_date:2026-08-31`

with `per-page=200` and cursor pagination until no records remained. The exact arXiv query was `all:`t_1 `AND ... AND all:`t_k. The requests used `start=0` and `max_results=200`. Results were ordered using `sortBy=submittedDate` and `sortOrder=descending`, and pagination continued until the reported total was reached. Quotation marks shown below were passed to the APIs. Thus the table, together with these two field templates, gives every exact query sent to both databases. The accompanying files `search_summary.json` and `search_and_screen.py` retain the literal filter strings and query parameters.

Table D.1: Exact term blocks used in both databases

ID	Terms joined by `AND`
Q01	`"large language model" AND "simulate human behavior"`
Q02	`"large language model" AND "behavioral simulation"`
Q03	`"large language model" AND survey AND respondent AND simulation`
Q04	`"large language model" AND "synthetic respondents"`
Q05	`"large language model" AND "silicon samples"`
Q06	`"large language model" AND "digital twins" AND survey`
Q07	`"large language model" AND "social science experiments" AND predict`
Q08	`"large language model" AND "economic agents" AND behavior`
Q09	`"large language model" AND consumer AND behavior AND simulation`
Q10	`"large language model" AND "public opinion" AND simulation`
Q11	`"generative AI" AND "human behavior" AND simulation`
Q12	`"foundation model" AND "human cognition"`
Q13	`"large language model" AND replace AND "human participants"`

We then searched the references and citing papers of the central empirical and methods

papers and searched for later papers by their authors. Records already present in the authors' working library were screened under the same rules. This route added 53 records after exact-title matching against the database results.

D.3 Duplicate removal and screening

Within each database, multiple query hits were first collapsed by the database record identifier. The combined set was then matched by normalized DOI, arXiv identifier, and normalized title, removing duplicates both within and across databases. Title normalization converted text to lowercase, removed punctuation and formatting, and collapsed whitespace. A journal article and its working-paper version were treated as one study. The retained record preferred a DOI-bearing version and the record with the more complete abstract. Exact-title matching was then used to prevent records from the prior library and citation or author searches from being added twice.

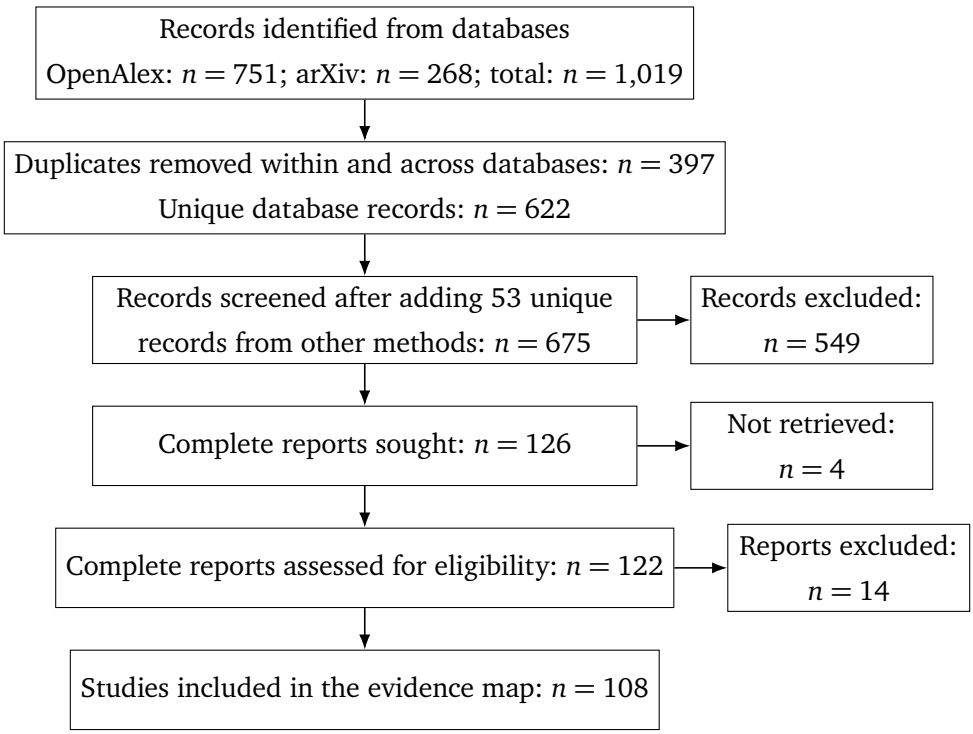
Title and abstract screening used the eligibility rules above. A deterministic keyword pass identified records that contained all four concepts: a generative language model, a human target, simulation or prediction, and an empirical comparison. It was used to remove clearly irrelevant records and order the remaining records. Complete reports passing this stage were checked against the independent-human-benchmark rule. Every decision, reason, source identifier, query identifier, DOI, and URL is retained in `final_screening_ledger.csv`; the included set is separately retained in `included_studies.csv`.

D.4 Flow counts

The thirteen OpenAlex searches produced 751 unique records after overlaps among the queries within OpenAlex were collapsed. The same procedure produced 268 unique arXiv records. Matching within and across the two databases removed 397 duplicates and left 622 database records. The prior library and citation or author searches added 53 unique records, so 675 titles and abstracts were screened. We excluded 549 at that stage. We sought 126 complete reports; four could not be located. We assessed 122 complete reports, excluded 14, and included 108 studies.

The 14 full-text exclusions were: methods or perspective papers without an eligible empirical test ($n = 4$); no independent human benchmark ($n = 4$); a research-result forecast rather than human behavior ($n = 2$); no observed human target ($n = 1$); a non-generative model ($n = 1$); a model proposal without an independent human test ($n = 1$); and a review or methods discussion ($n = 1$). Figure D.1 reports the complete flow.

Figure D.1: PRISMA flow diagram for the screened evidence set



Notes. The diagram follows the logic of PRISMA 2020. "Other methods" comprise the prior working library and backward, forward, and author searches. The review protocol was not prospectively registered.

D.5 Descriptive map of the included reports

We assigned each report one primary target using its title and abstract where available. The categories follow the granularity distinctions in the main text. They are descriptive codes, not judgments about whether a study's result is positive or negative. A paper that tests several target levels remains in the mixed category rather than being counted several times.

Table D.2: Primary targets in the screened evidence map

Primary target	Reports
Aggregate effect or index	15
Population report or distribution	32
Subgroup report or distribution	14
Individual report or trait	10
Individual behavior or choice	28
Mixed or multiple target levels	9
Total	108

Notes. Counts come from `evidence_map_v2.csv`, generated by `build_evidence_map.py`. Coding uses the primary target named in the title and abstract where available. "Behavior or choice" includes laboratory, field, strategic, and cognitive-choice tasks; it does not imply that every task uses material stakes.

The map contains 71 reports whose primary target is a report, rating, distribution, aggregate effect, or index; 28 whose primary target is a choice or behavioral outcome; and nine that span several outcome types. This count supports one conclusion about the literature's composition: most tests concern what people report or aggregate quantities built from reports. It does not show that the 71 studies agree with one another, or that all 28 behavioral studies provide equally strong tests of real-world behavior.

The structured table in Appendix B uses a narrower selection rule. A study must provide a clear quantitative human benchmark and help span the broad-to-specific comparison, the forecaster/respondent distinction, positive results, and stress tests. This rule produces an interpretable comparison while the complete map prevents the narrative from being mistaken for the full search result.

D.6 Interpretation and reproducibility

The included papers use different targets, samples, models, and measures of success. The flow count describes the breadth of the evidence base; it does not make the studies statistically exchangeable. For the same reason, the main text reports a structured comparison rather than a pooled average. The replication files preserve the raw result pages, within-source query union, deduplicated screening sheet, final decision ledger, included-study list, descriptive evidence map, search summary, and scripts that regenerate every number above.