Semilattice

July 31, 2026

Mind the Gaps: Mixture-of-Minds for Human Simulation

Pranav Dahiya
Semilattice
p.dahiya@semilattice.ai

Abstract

Predicting how a population will answer a new question is a long-standing goal. Statistical methods succeed at the level of the mass but falter at the level of the individual. Large language model simulators inherit this gap. They recover a population's central tendencies while flattening its heterogeneity, and they carry social biases and prompt brittleness that distort individual predictions. This paper introduces Anacreon, an audience simulation model that targets the individual level within a narrow, well-specified domain. Anacreon learns an authorship embedding that separates individuals, clusters a real qualitative corpus around seed people, and trains a dedicated adapter for each cluster, a mixture of minds, on a Gemma 4 12B base. It harvests demographics, psychological traits, and survey responses from public text, and augments each record with a chain-of-emotion. It reduces prompt brittleness by shuffling response options and reduces positive bias by balancing the training distribution. On a large, externally sourced survey, Anacreon reaches a state-of-the-art ordinal alignment of 0.775, the individual-level accuracy measure on which the field has converged, with a small residual bias. The work is a step toward drawing aggregate insight from faithfully simulated individuals.

"Dasein is a being that does not simply occur among other beings. Rather it is ontically distinguished by the fact that in its being this being is concerned about its very being."
—Heidegger[1]

1 Introduction

The ambition to render human behaviour calculable begins with death. In 1662, John Graunt, now regarded as the father of demography, distilled statistical regularity from London's weekly bills of mortality. His work was the first empirical life table. Graunt's innovation was epistemological. He saw that individual deaths are unpredictable while death in aggregate is orderly [14]. The theory became an industry in 1762 with the establishment of the Society for Equitable Assurances on Lives and Survivorships in London [22]. The foundation laid by Graunt now underpins a vast global insurance industry. Its worldwide premiums reached a record of about USD 7.1 trillion in 2023 [36].

In 1835 Adolphe Quetelet proposed the idea of *l'homme moyen*, the average man about whom human measurements cluster [29]. He imported the astronomer's law of error, the normal distribution, into the study of human beings. He found that many human traits stayed numerically stable from year to year, among them height, weight, strength, and, more controversially, rates of drunkenness. The statistics of demography reached its modern state in the first half of the 20th century with the innovation of scientific opinion polling. Polling showed that a small, well-drawn sample could outpredict a far larger but haphazard one. Leading up to the 1936 United States presidential election, the *Literary Digest* mailed out 10 million mock ballots and, from 2.4 million returns, forecast that Alfred

[1] From *Being and Time*, trans. Joan Stambaugh, revised by Dennis J. Schmidt (Albany: SUNY Press, 2010).

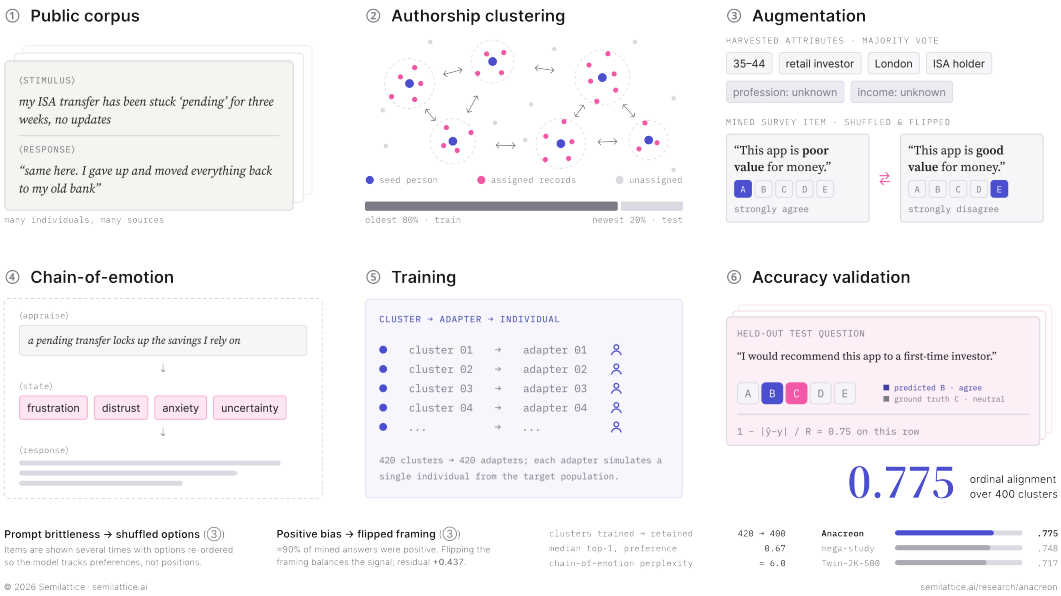


Figure 1: Overview of the Anacreon pipeline. Public text is clustered by author into a mixture of per-cluster models. Each cluster's records are augmented with harvested attributes and a chain-of-emotion, and a dedicated model is fine-tuned and selected against survey accuracy, reaching a state-of-the-art ordinal alignment of 0.775.

Landon would defeat Franklin Roosevelt [35]. Roosevelt won in a landslide. The magazine's vast sample was drawn from telephone and automobile registries, which skewed affluent and Republican, so its size did not offset its bias. George Gallup, using a quota sample of only 50,000, correctly called Roosevelt's victory and even predicted the magnitude of the *Digest*'s error [12].

The 20th century did not stop at sampling. The Cold War defence-research complex recast strategic human behaviour as formal mathematics [38]. It also reimagined society as something that could be simulated rather than merely sampled. Founded in 1959, the Simulmatics Corporation built "the People-Machine" that sorted the electorate into voter types and predicted their responses [20]. It advised the 1960 Kennedy campaign and then claimed credit for the win. Its confidence eventually outran its evidence. Lepore [20] shows that many of its predictions were common sense dressed in the authority of the machine. When the New York Times hired it to model the 1962 midterm elections, the effort collapsed into a disorganised shambles. So did its contract with DARPA, where it was hired to win over the South Vietnamese to the American cause. Its overreach prefigures a failure of modern LLM-based simulation that section 2 reviews. Simulmatics promised to model everything from buying a dishwasher to countering an insurgency [20], a breadth so vast that its predictions could rarely be grounded or tested.

Agent-based models (ABMs), brought to prominence with the Schelling segregation model [33], flipped the script by modelling populations from the bottom up instead of the top down. Schelling demonstrated how simple hand-set rules about individual preferences could lead to observed collective patterns that no one intended. Joshua Epstein and Robert Axtell generalised this insight into artificial societies whose inequality, migration, and trade emerged from local rules alone [9]. Such models show how aggregate emergence can be simulated even with narrow insights into individual behaviour.

Across this arc, aggregate trends have proved predictable only when they are stable. A stable trend is not driven by any one person, so statistics captures it well. Insurance risk and public-health rates stay orderly from year to year, without any model of an individual. Transient trends resist this. Election results and approval ratings turn on particular people at a particular moment, so they cannot be deduced from stable regularities. They have to be measured, which is why scientific opinion polling was developed. Agent-based modelling appears to bridge the two, but it delivers less than it promises [40]. As Smaldino [34] argues, an agent-based model shows only that a mechanism is sufficient to produce an observed pattern. It explains how a pattern could have arisen, not what an unobserved population would actually do. The individual–aggregate gap therefore persists. It would

close only if individual behaviour could be predicted well enough across a representatively sampled population, and this becomes tractable when the behaviour is restricted to a narrow, well-specified domain.

The wider the target, the less any prediction can be grounded against real data. Existing LLM simulators run headlong into this limit. They repeat the overreach of "the People-Machine" [20] when they ask one model to stand in for any population on any question. They recover a population's central tendencies while collapsing the heterogeneity, uncertainty, and minority positions that distinguish real individuals within it. This flattening is not incidental. To become useful assistants, these models undergo extensive reinforcement learning from human feedback, which sharply reduces the diversity of their outputs [17]. A better assistant is a worse human simulator, since it answers as one agreeable average rather than as the many distinct people a population contains. The base LLM also carries measurable social biases [32, 11], and these systems rarely correct them. *Anacreon's goal is to simulate independent, heterogeneous individuals well enough to draw meaningful aggregate insights.* It simulates narrow, well-specified domains, and it works to remove the model's bias rather than trust it. A machine trusted to know more than it does is more dangerous than one that admits its limits.

People reveal a great deal about themselves through the records they leave in public. From a dataset of more than 58,000 volunteers, Kosinski et al. [19] showed that Facebook "likes" alone predict many private attributes. The likes recover a person's Big Five (OCEAN) personality traits, and they distinguish Democrats from Republicans in 85% of cases. Public behaviour is therefore a strong signal for private disposition. Anacreon relies on this signal. It collects what a population makes public and infers the private traits that shape its answers. Rather than a single dense model, it trains a *mixture of minds*. It addresses two failure modes that the literature has repeatedly documented in naive LLM simulation (section 2). The first is *prompt brittleness*: simulated responses shift with the exact wording, framing, and option ordering of the prompt rather than tracking stable preferences. The second is *bias*: models over-represent dominant groups and caricature others through demographic stereotyping. On a large, held-out, survey test set, Anacreon achieves state-of-the-art ordinal alignment.

2 Background

This section traces how LLM simulators moved from matching populations to modelling individuals. It outlines why faithful individual prediction remains hard, how bias distorts the groups being simulated, and how the field has learned to measure what remains.

2.1 From aggregate to individual fidelity

The earliest LLM simulators targeted populations, not persons. Prompted models could reproduce classic findings in aggregate. Argyle et al. [3] coined the term *silicon samples*. They conditioned a model on real socio-demographic backstories and recovered the opinion distributions of many subgroups, a property they named *algorithmic fidelity*. Their evaluation was explicit about its scope. It matched the aggregate distribution but did not assess individual-level correspondence. Population-level prototyping followed the same logic, expanding a few seed personas into a community to surface plausible interactions [24].

Attention then turned to identifiable individuals. *Generative Agents* gave the template, producing coherent longitudinal behaviour for sandbox characters [25]. Park et al. [26] scaled this to 1,052 real people, grounding an agent for each in a two-hour interview and structured surveys. Their agents recovered roughly 85% of a participant's own two-week test–retest consistency, and richer self-report data narrowed accuracy gaps across subgroups relative to demographic prompting, which is known to induce stereotyping [32].

Rich grounding, however, buys aggregate plausibility more cheaply than individual heterogeneity. The *Twin-2K-500* twins reached a 0.717 individual-level accuracy score, 87.7% of the human ceiling, yet expressed markedly less response variance than humans [37]. A large follow-up study on the same individuals was more sobering. On novel stimuli, twin responses correlated only weakly with their human counterparts (average $r \approx 0.20$) and fell back toward the population mean [28]. They compress variation, and a large share of regression coefficients estimated on them differ in size

or even flip in sign [6]. A good aggregate fit does not protect the individual-level relationships an analysis depends on.

These distortions motivate a shift from prompting frozen models to training grounded ones. Finetuning on large corpora of human responses improves distributional alignment, though it exposes a tension. Better distribution matching can reduce single-response accuracy [18]. The same approach may scale to a single trained model of behaviour across many experiments [5].

2.2 Bias and group representation

A simulated population can misrepresent the very groups it stands in for. Wang et al. [39] identify two failure modes. The first is *misportrayal*. Prompted with a demographic identity, a model more often reflects what outsiders think of a group than what its members say about themselves, and this falls hardest on marginalised identities. The second is *flattening*. The model renders each group as homogeneous and ignores within-group heterogeneity. These effects compound the demographic misalignment of Santurkar et al. [32] and the variance compression noted above.

Hu et al. [15] treat bias as a property to be corrected before simulation. They generate narrative personas from human corpora and reweight them by importance sampling so that the persona set's Big Five distribution matches a global human reference. This lowers the distributional distance to that reference by about 32% relative to the strongest prior persona set. However, this still fails to address the failure modes highlighted above.

2.3 Measuring the gap

Comparing simulators requires a alignment metric for the ordinal and Likert scales that are pervasive in surveys, where binary exact-match is uninformative. Toubia et al. [37] and Kolluri et al. [18] adopt what this paper terms *ordinal alignment*. Given a predicted response $\hat{y}$ and a ground truth y on a scale of range R (for an n-point scale, $R = n - 1$), accuracy is one minus the normalised mean absolute deviation, averaged over questions,

$$\mathrm{A}_{\mathrm{ord}} = 1 - \frac{\mathrm{MAD}}{R} = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i - y_i|}{R}. \tag{1}$$

The measure lies in $[0, 1]$. It equals 1 for an exact match and 0 when the prediction is maximally wrong, and it reduces to ordinary accuracy on binary items. For example, on a 1-to-7 item, a predicted 6 against a true 5 scores 0.833. It has become the community's settled measure of individual-level fidelity, adopted by the Twin-2K-500 dataset, its mega-study follow-up, and the normalised-deviation accuracy of Kolluri et al. [18]. On this common scale, the state of the art still sits well short of human reliability [26, 37].

Ordinal alignment is necessary but not sufficient. A model can score well on eq. (1) while collapsing the population's variance, so the field often pairs it with distribution-level checks on the spread and shape of predicted responses.

3 Method

Anacreon is trained on unstructured text drawn from many sources. Each record is cast as a *(stimulus, response)* pair, where the stimulus, s, is a condition and the response, r, is the outcome an individual produced under it. The stimulus space is defined by the target population to be modelled. Rather than map the stimulus straight to the response, Anacreon first constructs a *chain-of-emotion* (c), a short trace of the emotional states and appraisals the stimulus evokes, then produces the response conditioned on that trace (section 3.2.1). Fine-tuning minimises the expected negative log-likelihood of the target,

$$\mathcal{L}(\theta) = -\,\mathbb{E}_{(s,c,r)\sim\mathcal{D}}\big[\log p_\theta(c, r \mid s)\big], \qquad p_\theta(c, r \mid s) = p_\theta(c \mid s)\, p_\theta(r \mid s, c), \tag{2}$$

where $\mathcal{D}$ is the training set and θ the model parameters.

3.1 Clustering

A subset of individuals with rich data is chosen as a seed population. Inclusion requires a minimum quantity of data per individual. A transformer is then trained to embed individuals so that the seed people are maximally separated in the embedding space. Training uses the contrastive authorship-representation objective of Rivera-Soto et al. [31]. Each individual in a batch contributes two disjoint samples of their text, which the encoder maps to embeddings u_i and v_i. The objective pulls the two samples of the same individual together and pushes samples of different individuals apart. For a batch of B individuals, with cosine similarity $s(x, y) = x^\top y/(\|x\| \, \|y\|)$ and temperature τ, the loss is the symmetric contrastive (InfoNCE) term

$$\mathcal{L} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp\big(s(u_i, v_i)/\tau\big)}{\sum_{j=1}^{B} \exp\big(s(u_i, v_j)/\tau\big)} + \log \frac{\exp\big(s(u_i, v_i)/\tau\big)}{\sum_{j=1}^{B} \exp\big(s(u_j, v_i)/\tau\big)} \right]. \tag{3}$$

The other individuals in the batch act as negatives. Minimising eq. (3) separates individuals in the embedding space, which is what the clustering step below relies on.

The trained model is then applied to the full corpus. Every record is assigned to its nearest seed individual, which partitions the corpus into clusters. Each cluster is then split by time. The newest 20% of records is held out for testing, and the remaining 80% is used for training. The temporal split prevents later information from leaking into training.

3.2 Data Augmentation

An LLM-as-judge extracts attributes from each cluster's text. These include demographics and a set of psychological traits. A proprietary harness reduces hallucination, and a majority vote across multiple runs stabilises each attribute. The harness favours accuracy over coverage. When the runs disagree too much for a majority to form, the attribute is left undefined rather than guessed. This is why several demographic categories in section 4 carry an unknown bucket. Survey questions are also mined from the text, since some text forms reveal answers to standard instruments. Figure 2 shows one example, where user generated free-text is turned into a multiple-choice question whose answer is known. Survey questions are mined separately on the train and test splits, so that test answers cannot leak into training. The mined survey set spans two formats: ordinal, Likert-style and categorical multiple choice questions.

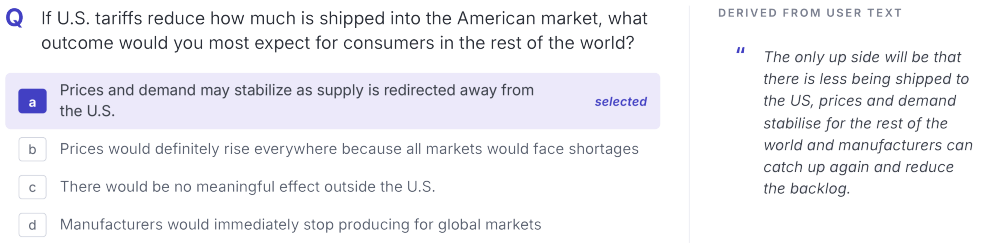


Figure 2: Generation of a mined survey question.

Due to the positivity bias in LLM responses [10]. most of the ordinal questions generated by the mining agent had positive responses, skewing towards "agree"/"strongly agree" in 90% of cases. To fix this problem, every question with a non-neutral response was included twice in the training and test sets, once framed positively and once negatively. This not only aims to address the sycophancy problem but also judges how susceptible the trained models will be to prompt brittleness.

3.2.1 Chain-of-Emotion

Each record is also augmented with a *chain-of-emotion*, in the sense of the appraisal-based construction of Croissant et al. [7] and the latent emotional-state alignment of Wu et al. [41]. An agent harness takes the stimulus and selects the harvested attributes relevant to it. Only a few attributes matter for any given stimulus, not all. This is consistent with the constructive-choice view, in which

people weight a context-dependent subset of attributes [4]. The chain then maps the evolution of emotional state from stimulus to response. It is constructed for both unstructured text and mined survey questions.

3.3 Training and Checkpoint Selection

A separate high-rank LoRA adapter is trained for each cluster on its augmented data, on top of a Gemma 4 12B base model [13], using QLoRA [8] for compute efficiency. Together the per-cluster adapters form Anacreon's mixture of minds, one model specialised to each cluster. At rank 64, the 12B base yields about 260M trainable parameters per cluster. Summed over the 420 trained clusters, Anacreon learnt about 110B parameters in total. During each training epoch, the order of responses for each survey question in the training set is randomly shuffled. This acts as a sort of dropout to ensure the model is not learning to predict responses purely based on their position in the list of options.

The clusters vary in quality, as expected. Coherent clusters, whose members are genuinely similar, train well. Incoherent clusters train poorly. In the worst case a cluster contains conflicting statements, so the model cannot tell whether a response should be positive or negative. A long tail of clusters failed to converge and was pruned. At inference the 400 retained per-cluster models unpack to 4.9T parameters.

The best checkpoint is chosen by balancing three quantities on held-out data. The first is top-1 accuracy $A_{\mathrm{mc}}(\theta)$ on the multiple-choice questions, the true positive rate. The second is the ordinal alignment $A_{\mathrm{ord}}(\theta)$ of eq. (1) on the ordinal questions. The third is the perplexity of the chain-of-emotion. For the target tokens $w_{1:N}$ of a held-out example, only the chain-of-emotion, perplexity is the exponentiated mean per-token negative log-likelihood,

$$\mathrm{PPL}(w_{1:N}) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p_\theta(w_i \mid w_{<i}, s)\right), \tag{4}$$

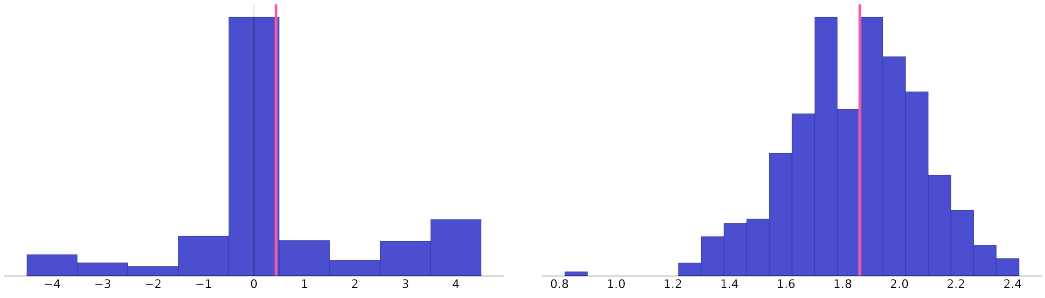
the exponentiated cross-entropy between the model and the target [16]. A low perplexity means the model has converged well on next-token prediction of the chain-of-emotion. Writing $\mathrm{PPL}(\theta)$ for its mean over the held-out pairs, the three signals are combined into a selection score, and the checkpoint that maximises it is kept,

$$\theta^\star = \arg\max_\theta \ \alpha\, A_{\mathrm{mc}}(\theta) + \beta\, A_{\mathrm{ord}}(\theta) - \gamma\, \mathrm{PPL}(\theta), \tag{5}$$

where the nonnegative weights α, β and γ balance accuracy, ordinal alignment, and convergence.

4 Results

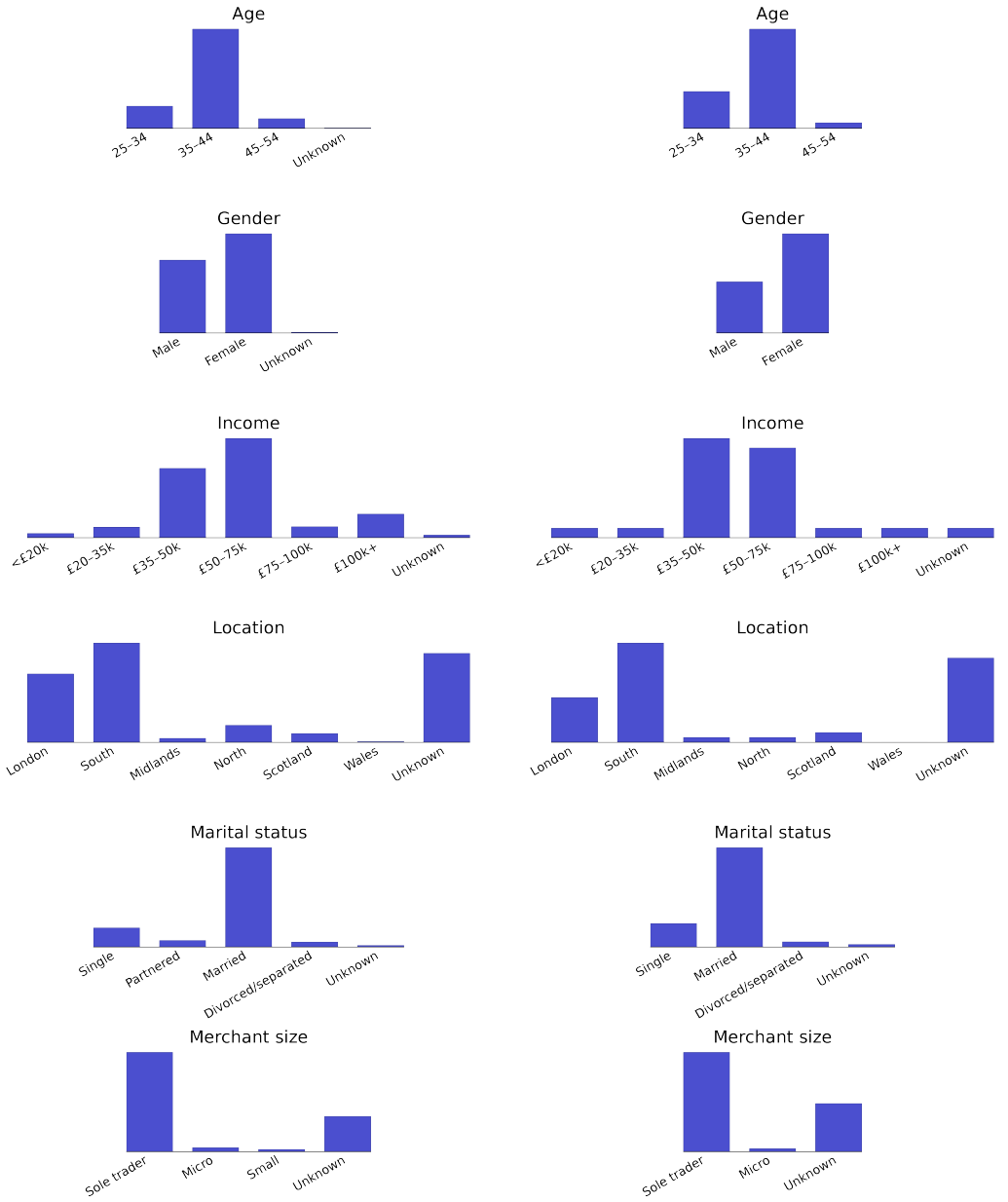
This section evaluates Anacreon on a B2B audience model built for a Semilattice customer. The modelled individuals are SME merchants who take physical, in-person card payments.



(a) Signed error, pooled over all rows.

(b) Per-cluster standard deviation σ.

Figure 3: Signed prediction error, in semantic scale positions. (a) pooled over all survey rows, peaked at zero with mean $+0.437$. (b) per-cluster standard deviation σ of the signed error, with a median near 1.85.



(a) Full modelled population.

(b) The 50 lowest-scoring clusters

Figure 4: Demographic composition over age, gender, income, location, marital status, and merchant size. (a) the full modelled population; (b) the 50 clusters with the lowest ordinal alignment. The two distributions are similar, so the weakest clusters are not concentrated in any single demographic group.

4.1 Bias

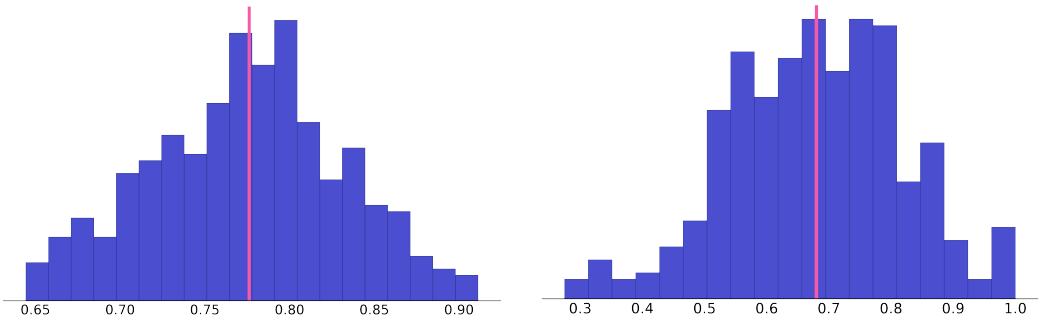
A well-calibrated simulator should place its predictions symmetrically around the true responses. Base LLMs instead tend toward a positive, sycophantic bias [10], which Anacreon is designed to reduce. Figure 3 reports the signed prediction error, measured in semantic scale positions. Figure 3a pools all survey rows. Each model was prompted with each row in its test survey set 5 times with shuffled responses. The distribution concentrates at zero but retains a small positive mean of $+0.437$. A positivity bias therefore remains, but it is much reduced relative to a base LLM.

The errors vary in spread. Figure 3b shows the standard deviation of the signed error within each cluster. The distribution is unimodal, with a median near 1.85 scale positions. This spread, rather than the small systematic bias, accounts for most of a cluster's deviation from the true responses.

Figure 4a shows the demographic composition of the modelled population across age, gender, income, location, marital status, and merchant size. The spread is broad and is not dominated by any single group. This indicates that the attribute-mining method (section 3.2) did not introduce a strong sampling bias.

Figure 4b shows the same composition for the 50 lowest-scoring clusters, ranked by ordinal alignment. Its shape tracks the full population (fig. 4a). The weakest clusters are therefore not concentrated in any single demographic group. This suggests that poor performance reflects the data quality within a cluster rather than the demographic it represents.

4.2 Survey Accuracy



(a) Ordinal alignment distribution

(b) Top-1 accuracy distribution

Figure 5: Frequency distribution of per-cluster survey performance. (a) ordinal alignment ($1 - \mathrm{MAD}/R$) on the ordinal questions; the median cluster (pink line) scores 0.775 (b) top-1 accuracy on the multiple-choice preference questions; the median cluster reaches 0.679

The mined survey set spans two formats, ordinal Likert-style items and multiple-choice preference questions (section 3.2). Figure 5 reports both. Figure 5a shows ordinal alignment across the evaluation clusters. The distribution is centred at 0.775, and its bulk sits above the shuffle baseline band. Table 1 places this score against prior work. On the shared ordinal-alignment measure ($1 - \mathrm{MAD}/R$), Anacreon at 0.775 exceeds every prior system that reports it, namely Twin-2K-500 at 0.717, Socrates-Qwen-14B at 0.740, and the digital-twin mega-study at 0.748. This is the state of the art on this measure.

Figure 5b reports top-1 accuracy on the multiple-choice preference questions, that is, whether the predicted choice is exactly correct. The median cluster reaches 0.679. This is comparable to the 0.657 exact-match accuracy of the interview-grounded agents of Park et al. [26], the closest prior system that reports an exact-match measure (table 1). The digital-twin literature otherwise reports only ordinal questions.

4.3 Chain-of-Emotion Perplexity

Anacreon produces a chain-of-emotion before each answer (section 3.2.1). Its perplexity (eq. (4)) measures how well the model predicts these sequences. A value of k means the model is, on average, as uncertain as a uniform choice among k tokens. A value that's too low indicates the model has become more deterministic than desired; a value that's too high indicates the model has failed to converge well.

Perplexity has no universal scale. It depends on the tokenizer, the vocabulary, and the domain, so values compare only within a fixed setup. As a rough guide, strong open-domain language models report word-level perplexities in the high teens on standard English benchmarks, for instance 17.5 on WikiText-103 [30]. Single-digit values usually signal a narrow or highly predictable domain.

Table 1: Individual-level accuracy of Anacreon and prior work, grouped by the measure used; higher is better. The scores come from related but not identical measures on different datasets, so they are indicative rather than a controlled comparison. Top-1 (exact-match) accuracy compares Anacreon with the interview-grounded agents of Park et al. [26]. Ordinal alignment ($1 - \mathrm{MAD}/R$, eq. (1)) compares it with the digital-twin systems that report it. The *Approach* column summarises each system's modelling technique. The final column gives the human two-week test–retest (self-consistency) baseline where the paper reports one.

Metric	Method	Approach	Score	Test–retest
Top-1 accuracy	Park et al. [26]	interview-based prompts	0.657	0.795
	Anacreon	per-cluster fine-tuning	**0.679**	–
Ordinal alignment	Toubia et al. [37]	context-injected twins	0.717	0.817
	Kolluri et al. [18]	supervised fine-tuning	0.740	–
	Peng et al. [28]	twins, novel stimuli	0.748	–
	Anacreon	per-cluster fine-tuning	**0.775**	–

The chain-of-emotion is such a narrow domain. It is not free text. At each step the continuation is drawn from a narrow, structured space of emotional states and appraisals (section 3.2.1), so few tokens are ever plausible. This makes perplexity a sensitive convergence signal. When the valid continuations are few, a converged model concentrates almost all of its mass on them and reaches a low perplexity near the entropy of that space. A high perplexity would show the opposite, that the model still spreads mass over tokens the format does not permit. Figure 6 reports this quantity per cluster at the best epoch. The distribution is unimodal and concentrated, with a median near 6.0. The models mostly sit in the single digits, well inside the range that such a constrained target allows. This indicates that they reliably reconstruct the emotional chain that leads to each answer.

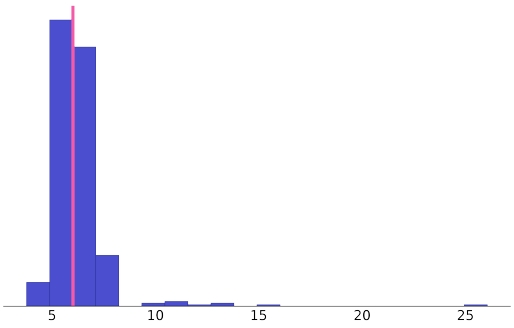
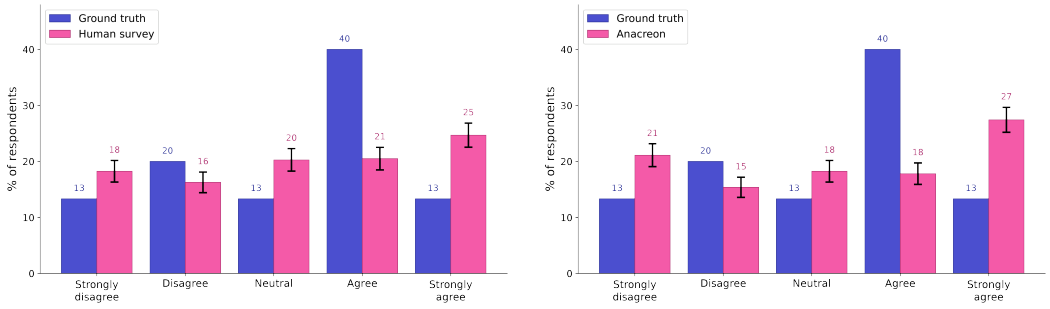


Figure 6: Chain-of-emotion perplexity frequency distribution per cluster. The vertical line marks the median, near 6.0.

5 Discussion

Narrow Domain. Anacreon specifically targets narrowly defined populations, addressing the overreach that section 1 traced from Simulmatics to today's simulators. The beliefs encoded in each cluster's training set therefore act like a belief network [27]. That network is well supported near the training domain and sparse away from it. Conclusions extend to topics adjacent to the training data, where the network still carries evidence. They cannot be trusted far beyond it, where the evidence thins to guesswork. At inference a query can be checked for support in the training data, so an answer is offered only when the records provide some basis for it. Anacreon can therefore recognise the edge of its own competence and decline to speak past it.

Bias and Variance. A positivity bias, though small, has not been completely eliminated (section 4.1). More importantly, the evaluation so far concerns individual-level accuracy, and despite state-of-the-art ordinal alignment and top-1 accuracy, the models display high variance (fig. 3b). Humans are not deterministic either. Single survey items measure attitudes with low reliability,



(a) Human test-retest

(b) Anacreon

Figure 7: Illustrative response distributions on a five-point Likert item, each against the same ground truth. The blue bars indicate a theoretical ground truth response distribution that's being measured with (a) a human survey, and (b) an Anacreon simulation. Both charts indicate a worst case skew that could arise from the natural variance in human responses and the measured median variance of Anacreon. Anacreon's skew is only moderately worse than human test-retest.

around 0.5; re-surveyed respondents can give substantially different answers [2, 1]. By classical test theory, a reliability of 0.5 corresponds to a test-retest standard deviation near 1.27 scale positions on a five-point item [21]. Figure 7 sets this human baseline against Anacreon. Anacreon's per-cluster median standard deviation is about 1.85, above the human level but only moderately so. A faithful simulator should reproduce this human spread rather than collapse to a point, so the target is the human level, not zero variance. However, while the median model converges well, models at the tail end are only marginally better than a random guesser. Since the positivity bias was mostly eliminated, this should not stop Anacreon from being useful for aggregate, population-level predictions.

From individuals to aggregates. The value of a faithful individual simulator lies in what it enables at the aggregate. A population's spread, its disagreement, and its tails often carry the signal that matters. The next step is to study how better individual simulation translates into better aggregate prediction, by composing many well-simulated individuals into an estimate of a population's distribution. The template for doing so can come from Opinion Polling. A poll rarely draws a perfectly representative sample. Survey researchers correct the imbalance after the fact. They weight the respondents by post-stratification, and by its multilevel-regression extension, so that the reweighted sample matches known population margins [23]. The same correction can be applied to Anacreon's mixture of minds, reweighting the per-cluster models so that their pooled prediction matches a target population. However, this calibration is only sound when the individual predictions are sound. Individual fidelity is therefore the primary objective, and distributional alignment follows from it, not the other way around.

Ethics and data privacy. Anacreon learns from public text, and its design keeps the focus on groups rather than named persons. The clustering step (section 3.1) is central to this. Records are pooled into clusters, and every model is trained on a cluster rather than on any single individual. The seed individuals only anchor the clusters. The trained models capture behavioural trends shared across a cluster's members, not a profile of any one person. Anacreon is therefore built to describe how a segment tends to respond, not to identify, track, or manipulate a specific individual. Anacreon infers general patterns from simulation, and its outputs are population-level tendencies rather than predictions about private individuals.

6 Conclusion

This report set out to narrow the persistent gap between prediction at the level of the aggregate and prediction at the level of the individual (section 1). Statistical prediction has long paid out at the level of the mass, in insurance or public health. It has disappointed when pushed toward the specific person or act, in election forecasting. LLM-based simulators inherit this boundary. They tend to recover a population's central tendencies while flattening the heterogeneity and minority positions that define the individuals within it. Anacreon grounds a language model in real qualitative data for a narrow,

well-specified population, then predicts how that population would answer questions it has never been asked.

The method makes several contributions (section 3). It builds a seed population and an authorship embedding that separates individuals, clusters the corpus around them, and trains a mixture of minds, one dedicated model per cluster. It harvests demographics, psychological traits, and survey responses from public text, following the principle that public behaviour predicts private disposition. It also confronts the two failure modes that the literature documents in naive LLM simulation. Prompt brittleness is reduced by asking each question many times with shuffled options. Bias is reduced by flipping questions between positive and negative framing, which balances the responses the model sees during training.

These choices yield measurable gains. On a large, externally sourced survey instrument, Anacreon reaches an ordinal alignment of 0.775 (section 4.2). This is the state of the art among the systems that report the measure. The residual positive bias is small and much reduced relative to a base model, though it is not eliminated (section 4.1). The weakest clusters are not concentrated in any demographic group, which points to data quality within a cluster, rather than the demographic it represents, as the limit on performance (section 4.1).

For three and a half centuries, prediction has paid out at the level of the mass and faltered at the level of the person. Anacreon does not dissolve that boundary, but it does move it. By grounding a mixture of minds in what individuals reveal about themselves, it simulates a narrow population well enough to set the state of the art on individual-level accuracy. Bias and variance remain, and population-level fidelity is still to be shown.

References

[1] Christopher H. Achen. Mass political attitudes and the survey response. *American Political Science Review*, 69(4):1218–1231, 1975. doi:10.2307/1955282.

[2] Duane F. Alwin and Jon A. Krosnick. The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research*, 20(1): 139–181, 1991. doi:10.1177/0049124191020001005.

[3] 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. doi:10.1017/pan.2023.2.

[4] James R. Bettman, Mary Frances Luce, and John W. Payne. Constructive consumer choice processes. *Journal of Consumer Research*, 25(3):187–217, 1998. doi:10.1086/209535.

[5] Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Élteto, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marcel Mathony, Marcelo Mattar, and Eric Schulz. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025. doi:10.1038/s41586-025-09215-4.

[6] 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. doi:10.1017/pan.2024.5.

[7] Maximilian Croissant, Madeleine Frister, Guy Schofield, and Cade McCall. An appraisal-based chain-of-emotion architecture for affective language model game agents. *PLoS ONE*, 19(5): e0301033, 2024. doi:10.1371/journal.pone.0301033.

[8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html`.

[9] Joshua M. Epstein and Robert Axtell. *Growing Artificial Societies: Social Science from the Bottom Up*. Complex Adaptive Systems. Brookings Institution Press and MIT Press, Washington, DC, 1996. doi:10.7551/mitpress/3374.001.0001.

[10] Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. SycEval: Evaluating LLM sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, volume 8, pages 893–900. AAAI Press, 2025. doi:10.1609/aies.v8i1.36598.

[11] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024. doi:10.1162/coli_a_00524.

[12] George H. Gallup and Saul Forbes Rae. *The Pulse of Democracy: The Public-Opinion Poll and How It Works*. Simon & Schuster, New York, 1940.

[13] Gemma Team. Gemma 4 technical report. Technical report, Google DeepMind, 2026. URL `https://arxiv.org/abs/2607.02770`.

[14] John Graunt. *Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality*. Printed by Tho. Roycroft, for John Martin, James Allestry, and Tho. Dicas, London, 1662.

[15] Zhengyu Hu, Jianxun Lian, Zheyuan Xiao, Max Xiong, Yuxuan Lei, Tianfu Wang, Kaize Ding, Ziang Xiao, Nicholas Jing Yuan, and Xing Xie. Population-aligned persona generation for LLM-based social simulation. *arXiv preprint arXiv:2509.10127*, 2025. URL `https://arxiv.org/abs/2509.10127`.

[16] Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and James K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62 (S1):S63, 1977. URL `https://pubs.aip.org/asa/jasa/article/62/S1/S63/642598`.

[17] Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL `https://openreview.net/forum?id=PXD3FAVHJT`.

[18] 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 (EMNLP)*, pages 30096–30111, Suzhou, China, 2025. Association for Computational Linguistics. doi:10.18653/v1/2025.emnlp-main.1530.

[19] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110 (15):5802–5805, 2013. doi:10.1073/pnas.1218772110.

[20] Jill Lepore. *If Then: How the Simulmatics Corporation Invented the Future*. Liveright, New York, 2020.

[21] Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA, 1968.

[22] Maurice Edward Ogborn. *Equitable Assurances: The Story of Life Assurance in the Experience of the Equitable Life Assurance Society, 1762–1962*. George Allen & Unwin, London, 1962.

[23] David K. Park, Andrew Gelman, and Joseph Bafumi. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12(4):375–385, 2004. doi:10.1093/pan/mph024.

[24] Joon Sung Park, Lindsay Popowski, Carrie J. 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)*. ACM, 2022. doi:10.1145/3526113.3545616.

[25] Joon Sung Park, Joseph C. O'Brien, Carrie J. 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 (UIST)*, pages 1–22. ACM, 2023. doi:10.1145/3586183.3606763.

[26] 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. URL `https://arxiv.org/abs/2411.10109`.

[27] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[28] Tianyi Peng, George Gui, Daniel J. Merlau, Grace Jiarui Fan, Malek Ben Sliman, Melanie Brucks, Eric J. Johnson, Vicki Morwitz, et al. A mega-study of digital twins reveals strengths, weaknesses and opportunities for further improvement. *arXiv preprint arXiv:2509.19088*, 2025. URL `https://arxiv.org/abs/2509.19088`.

[29] Adolphe Quételet. *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale*. Bachelier, Paris, 1835. Published in 2 volumes.

[30] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI technical report*, 2019. URL `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`.

[31] Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 913–919. Association for Computational Linguistics, 2021. doi:10.18653/v1/2021.emnlp-main.70.

[32] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR, 2023.

[33] Thomas C. Schelling. Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1(2):143–186, 1971. doi:10.1080/0022250X.1971.9989794.

[34] Paul E. Smaldino. Models are stupid, and we need more of them. In Robin R. Vallacher, Andrzej Nowak, and Stephen J. Read, editors, *Computational Social Psychology*. Psychology Press, New York, 2017. URL `https://smaldino.com/wp/wp-content/uploads/2017/01/Smaldino2017-ModelsAreStupid.pdf`.

[35] Peverill Squire. Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly*, 52(1): 125–133, 1988. doi:10.1086/269085.

[36] Swiss Re Institute. sigma 3/2024: World insurance: Strengthening global resilience with a new lease of life. Technical report, Swiss Re, 2024. URL `https://www.swissre.com/institute/research/sigma-research/sigma-2024-03-world-insurance-global-resilience.html`.

[37] Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. Twin-2K-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*, 44(6):1446–1455, 2025. doi:10.1287/mksc.2025.0262.

[38] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.

[39] 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*, 7:400–411, 2025. doi:10.1038/s42256-025-00986-z.

[40] Paul Windrum, Giorgio Fagiolo, and Alessio Moneta. Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, 10(2): 8, 2007. URL https://www.jasss.org/10/2/8.html.

[41] Shirley Wu, Evelyn Choi, Arpandeep Khatua, Zhanghan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. HumanLM: Simulating users with state alignment beats response imitation. *arXiv preprint arXiv:2603.03303*, 2026. URL https://arxiv.org/abs/2603.03303.