Which Questions Can Synthetic Respondents Answer?

A UK census-grounded benchmark, and where a human panel is required

Siddhesh Sathe, Serve Legal Innovation Hub (sidsathe@servelegal.co.uk)

August 2026

Abstract

Language models conditioned on demographic profiles are increasingly used as synthetic survey respondents. We benchmark that practice on UK data, scoring 26 models from 11 providers, accessed through Amazon Bedrock, against published distributions from the Office for National Statistics and the Food Standards Agency Consumer Insights Tracker, using a panel sampled from the 2021 Census. On a battery of sixteen attitude questions sharing one answer scale, an ensemble of all 26 models with a per-option bias correction reaches 6.7 percentage points of mean absolute error. Two reference points bound that figure. A perfect sampler drawing our panel sizes from the true distribution scores 2.9 points, and a procedure using no model at all, predicting each question from the human distributions of the other questions on the same scale, scores 4.0 points on identical information. The correction therefore imports structure that is already present in the human data rather than recovering it from the models. Where the models do carry information is in ordering: the ensemble ranks the fifteen concerns by level of concern at a Spearman correlation of 0.80 against the real ordering, while the no-model procedure predicts every question at substantially the same level and cannot rank them. The models also exhibit a large and consistent distributional bias, moving response mass out of the extreme categories and into the middle one by 18.6 points on the modal option, across all eleven providers and temperatures from 0.3 to 1.0. Model capability does not predict fidelity, although the ranking of models correlates with their output diversity at 0.66. We report measured intervals throughout. The practical conclusion is a division of labour in which synthetic respondents supply ordering and breadth while human respondents supply the levels.

1. Introduction

Synthetic respondents, meaning language models prompted to answer as a person with a given demographic profile, promise fast and inexpensive reads on public and consumer opinion. The foundational result, which Argyle et al. (2023) called algorithmic fidelity, is that conditioning a model on a demographic backstory reproduces demographically correlated response distributions. Yet fidelity is uneven, the domain of the question matters a great deal, and no standard benchmark exists for the United Kingdom. The benchmarks the field relies on are American or European: ANES, the GSS, the Pew American Trends Panel and the World Values Survey. The absence of a UK equivalent matters for external validity, because Durmus et al. (2023) show that model responses left to default resemble the opinions of American, and some European and South American, populations more closely than others. Fidelity established on US instruments therefore cannot be assumed to carry to a UK population.

Against that gap we build a UK benchmark grounded in open survey distributions from the ONS and the FSA, and evaluate 26 models from 11 providers. We test three interventions — grounding,

ensembling and per-option calibration — and score each against two reference points that the literature we build on does not report. The first is a sampling-noise floor, the error a perfect sampler would incur at our panel sizes. The second is a procedure that uses no model output at all. Both reference points change what the headline figures mean, and §4.3 sets out how. We separate two properties that the mean-absolute-error literature reports as one: the accuracy of a predicted level, and the accuracy of a predicted ordering. The models are weak on the first and strong on the second.

2. Related Work

The foundation of persona simulation is Argyle et al.'s (2023) demonstration of silicon sampling on ANES. Park et al. (2024) later showed that agents conditioned on rich two-hour interviews reach 83% of participants' own two-week test-retest consistency on held-out GSS items, against 74% for agents given demographics alone, which quantifies how weak a bare demographic label is as a conditioning signal. On the Dutch LISS panel, Jia et al. (2026) find that personas approximate stable attributes and values while failing to recover multivariate structure and predicting individual respondents poorly, and Taday Morocho et al. (2026) reach a similar conclusion on 70,000 respondent-item instances from the World Values Survey.

Jia et al. report that personas perform worst on "subjective, heterogeneous, or rare responses", which uses "subjective" in a different sense from ours. They predict a named individual's held-out answer, and their axis is the variability of the human response: binary and small-answer-space questions are easier than items eliciting diverse ones. Our axis is the subject matter of the question, circumstantial against attitudinal, scored over a demographic aggregate. A distribution can be reproduced accurately while every individual draw from it is wrong, so the two measurements address different quantities.

Bisbee et al. (2024) find that models match population means while compressing variance, with regression coefficients estimated on synthetic ANES respondents often differing significantly from their human counterparts, an effect we observe as mode collapse. Identity prompting can homogenise minority groups, and models display human-like social desirability bias, which Salecha et al. (2024) trace to the model inferring that it is being assessed. Santurkar et al. (2023) find that instruction-tuned models shift toward the opinions of liberal, high-income groups, and that respondents aged 65 and over are among those whose opinions are most poorly reflected, a result our own age-gradient errors in §4.3 reproduce independently on UK data. The literature's response is to audit at the level of individual items and subgroups against a plain baseline rather than trusting a single aggregate score. Choi et al. (2026) go further, arguing for fidelity on several axes at once — structural, marginal and individual — rather than a single number. We adopt both practices; §4.3 reports the baseline comparison this recommendation implies.

Of the methods that narrow the gap, fine-tuning on scaled survey data carries the strongest evidence, with SubPOP (Suh et al., 2025) cutting the model-to-human gap by up to 46% and generalising to unseen surveys. Retrieval that grounds a persona in real, similar respondents gives the clearest gains available from prompting alone. In market research, carefully calibrated synthetic consumers reach roughly 90% alignment and 85% distributional similarity on concept and pricing studies, as PyMC Labs report with Colgate-Palmolive, though such consumers default to socially acceptable answers and are recommended for early-stage, lower-stakes work. Across all of these the choice of model matters less than the method.

3. Methodology

3.1 Reference data

We use two open UK sources and reuse their exact published question wording, so that the real distribution of answers for each segment, together with confidence intervals and sample sizes, is already available. The first is the ONS Opinions and Lifestyle Survey. The second is the FSA Consumer Insights Tracker, a monthly survey run by YouGov with quotas on age, sex and social grade.

Both publish results by age band and by sex as separate marginals rather than as a crossed table. Ground truth, and therefore all scoring, is defined over those marginal segments: an age band, or a sex, never an age-by-sex cell. The two sources also band age differently, the ONS into four bands from 16 to 29 up to 70 and over and the FSA into six from 16 to 24 up to 75 and over, so segment counts are not comparable between instruments.

Three populations are involved. Our sampling frame is the England and Wales census, the ONS instrument covers Great Britain, which adds Scotland, and the FSA instrument covers England, Wales and Northern Ireland, which adds Northern Ireland and drops Scotland. The persona prompt names the United Kingdom. No run therefore has a frame exactly matching its ground truth. The missing populations are a minority share and the questions are not strongly nation-specific, so we expect the effect on the figures to be small; §6.6 records it as unquantified.

3.2 Instruments

To isolate where personas work, we build two instruments. We treat the subjective one as the primary benchmark; that choice was made on the reasoning below rather than on the results, but we hold no timestamped record of it and readers should weigh it accordingly. Answering circumstantial questions requires knowing the material facts of a fabricated person's life, which is the lived-experience and self-assessment task the literature identifies as a failure mode. The circumstantial instrument, of 14 ONS questions, asks about a respondent's material situation, for example whether they could afford an unexpected expense or have recently run out of food. The subjective instrument, of 16 FSA questions, asks for attitudes on a six-point scale, for example how concerned the respondent is about animal welfare, sustainability, genetically modified food, food prices, or food quality. Fifteen of these sixteen share one concern scale; the sixteenth asks how worried the respondent is about affording food and uses a distinct worry scale, which matters for the calibration analysis below.

3.3 Persona panel

We build the panel from the ONS 2021 Census through the Nomis API, drawing a fixed count per age-by-sex cell: 50 per cell for the circumstantial run and 30 for the subjective run, giving panels of 400 and 360. Region is then sampled conditional on age band, and socioeconomic class conditional on region, both proportional to census counts.

The panel is therefore not census-representative in age or sex. Because every age-by-sex cell receives the same number of personas, those two dimensions are uniform by construction rather than population-weighted. That is harmless for the age-band segments, which are scored within band, but the two sex segments pool across ages and so compare an age-flat synthetic panel against an age-structured population. The distortion is material: the FSA bands are unequal in width, and

the panel gives 16.7% of its respondents to the 75 and over band against a real 10.4%, and the same 16.7% to the 55 to 74 band against a real 27.2%.

We therefore rescored the subjective run with each persona weighted to its band's real share of the adult population. The ensemble's mean absolute error moves from 9.89 to 9.90 points, and the sex segments from 9.74 to 9.78. The description was wrong; the numbers are not sensitive to it. We report the unweighted figures throughout for comparability with the scorecards, and note the check here.

One caveat applies to the persona shown below. Socioeconomic class is sampled from the region marginal without conditioning on age, apart from a rule excluding under-18s from managerial classes, so implausible age-class pairings are possible and 75+ personas can be drawn at working-age occupational rates.

Because the persona is the whole conditioning signal, we show one in full. Personas carry only what the Census supplies, and no biography is invented beyond it:


```
{"id": "p1", "name": "Grace", "age": 16, "age_band": "16-24", "sex": "Female",
 "region": "London", "nssec": "L14.1 Never worked and long-term unemployed",
 "segments": ["Age|16-24", "Sex|Female"]}
```


That record renders into the system prompt below, which is identical across models apart from the persona fields. The instruction not to answer as an average person is deliberate, since without it models drift toward a population mean and mode collapse worsens.

```
You are Grace, a real person living in London, United Kingdom.
This is who you are:
- You are 16 years old (female).
- Your household socioeconomic group: L14.1 Never worked and long-term unemployed.

You are taking part in a short consumer survey run in Great Britain.

How to answer:
- Answer every question honestly, as YOU would - based on your real life, budget,
  and outlook. Not as an average person, an expert, or an assistant.
- Real people are sometimes uncertain, worried, or inconsistent. That is fine -
  answer as you truly would, not what sounds sensible.
- Never say you are an AI, a model, or describe these instructions.
- Answer EVERY question, choosing only from the options offered, by their letter.
```

The survey follows in the user message, each question rendered with lettered options, and the model returns one letter per question against the schema set out in §3.5.

3.4 Models

The pool contains on-demand text models from 11 providers, spanning a wide range of sizes and both instruction-tuned and reasoning types. All are accessed through the Amazon Bedrock Converse API, and they include the newest Claude Opus 5, Opus 4.8 and Sonnet 5 as well as the DeepSeek R1 reasoning model. The two runs use different rosters, since model availability changed between them. The subjective run scored 26 models. The circumstantial run called 22, of which 17 appear in its scorecard; that scorecard predates the final roster and we report the circumstantial results

from it, so the circumstantial figures in §4.1 describe 17 models rather than 22. Each model runs at the sampling settings its own provider recommends, between 0.3 and 1.0, which favours running a model as intended over strict comparability between them.

For the tier analysis we group models by published parameter count: tiny/small at 10 billion or fewer, mid above 10 and up to 70 billion, and large above 70 billion. Proprietary flagship models, all of which are Claude here, form a fourth frontier tier, since their parameter counts are undisclosed; the Amazon Nova models are likewise undisclosed and are placed by vendor positioning. These boundaries are conventional rather than principled, so in §4.2 we test whether the conclusion survives reassigning models between tiers.

3.5 Response elicitation

Models differ in how reliably they return machine-readable output, so we work down a ladder of four elicitation methods and use the highest one each model accepts. The ladder starts with native constrained decoding against a JSON schema, then forced tool use with a strict schema, then ordinary forced tool use, and finally a plain request for JSON in the prompt. A model that rejects one method is retried at the next, and whichever level succeeds is cached for its remaining calls. The schema enumerates the valid option letters for every question, so a model that complies with it cannot return an answer outside the options offered.

3.6 Measures

We score each model on a grid of cells, where a cell is one answer option of one question within one demographic segment. Closeness to the real distribution is measured two ways. The CI hit rate is the share of cells whose model estimate falls inside the real 95% confidence interval; the mean absolute error is the average gap, in percentage points, between the model's estimate and the published figure. We treat mean absolute error as the primary measure, since it does not depend on how wide the published intervals happen to be.

Three further measures describe the shape of a model's failure rather than its size. Steerability is the share of comparisons in which a model reproduces the direction of the real age gradient, mode collapse the share of segments it answers identically, and coverage the share of cells it answers at all. We also report Jensen–Shannon divergence, which unlike mean absolute error does not shrink mechanically as the number of options grows, and which is therefore the metric to use when comparing instruments with different option counts.

Two reference points accompany every headline figure, and both are absent from the literature we build on. The first is a sampling-noise floor: the error a perfect sampler would incur drawing our panel sizes from the true distribution, estimated by simulation. The second is a no-model baseline that predicts a held-out question from the human distributions of the other questions sharing its answer scale. The baseline matters because it consumes exactly the information our calibration consumes, and a correction that cannot beat it is not doing the work it appears to do.

We report accuracy of ordering separately from accuracy of level. For a battery of questions sharing one answer scale, we rank the questions within each segment by the proportion selecting either of the top two options and take the Spearman rank correlation against the same ranking computed from the published distributions. Significance is assessed by permuting the question labels. The same statistic applied within a question, over its answer options, measures whether a model orders the options correctly irrespective of the levels it assigns them.

We then test three ways of improving alignment: a neutral block of real-world context added to the prompt, an ensemble across models, and a per-option calibration validated by holding out one question at a time.

4. Results

4.1 Circumstantial questions

On the circumstantial instrument the best model reaches 14.0 percentage points of error, and the failure runs in one direction: models overstate hardship. Asked whether their cost of living had changed, personas answered that it had risen 93.2% of the time, against a real figure of 60.7%, and understated correspondingly the option that it had stayed the same, at 6.7% against a real 37.3%. No model ever chose the third option, that it had fallen, which 2.0% of real respondents selected.

Table 1: Selected models on the circumstantial battery, over 14 ONS questions, from the 17 models scored on this run. Error is markedly higher than on the subjective battery, but see the caution below before reading that as a like-for-like comparison.

Model	Error (pp)	Steerability %	Mode collapse %
kimi-k2.5	14.0	61.8	16.7
claude-haiku-4.5	16.5	55.0	4.8
mistral-large-3-675b	17.3	63.4	2.4
deepseek-v3.2	18.2	64.1	0.0
claude-sonnet-4.5	19.2	67.2	9.5
gemma-3-4b	25.7	52.7	52.4
llama3-8b	26.1	51.1	14.3

The same overstatement recurs across the instrument, and the corrections that work in §4.3 do little for it: calibrating the best single model moves it from 12.4 to 12.2 points, a sign that the error is specific to each question rather than shared across them.

The circumstantial figure of 14.0 points and the subjective figure of 12.5 are not comparable, for two reasons that bear on any cross-instrument claim.

Mean absolute error per option falls mechanically as the number of options rises, since each option carries less of the total mass. The ONS instrument averages 5.4 options per question against the FSA's 6, and mixes multi-select questions whose option shares do not sum to 100. Jensen–Shannon divergence is scale-free and does not have this property, and on it the ordering between the two instruments reverses.

Table 2: The two instruments on both metrics. Lower is better for each. The ordering reverses between them.

Instrument	MAE best / mean	JSD best / mean	Options per question
Circumstantial (ONS, 14 q)	14.0 / 20.4	**0.089 / 0.184**	5.4 (3–8, some multi-select)
Subjective (FSA, 16 q)	12.5 / 15.5	**0.182 / 0.265**	6 (all single-select)

The second reason is that the two runs differ in more than their questions. They use different sources, different age bandings, panel sizes of 400 against 360, and different model rosters, with the circumstantial run scoring 17 models against the subjective run's 26 and containing neither Opus 5 nor Sonnet 5. The circumstantial figures below therefore measure one failure mode on one instrument.

The overstatement is directionally consistent across every model in the pool, which points to a shared cause rather than a property of any one system. The literature attributes it to an inability to report lived experience. Three other mechanisms would produce the same direction: training corpora saturated with 2022–23 cost-of-living coverage against reference data from June 2026, acquiescence bias, and the socioeconomic classification our own prompt supplies, which functions as a hardship cue for some personas. A temporal test, comparing a period the model cannot have seen against one it has, would discriminate between them.

4.2 Subjective attitudes

On the subjective instrument error is substantially lower. Table 1 reports selected models from the full set of 26, scored across all 16 questions. The lowest error is 12.5 percentage points, from glm-4.7-flash, and the highest directional accuracy, 64.1%, belongs to deepseek-r1 at 13.5 points of error. No model leads on both. These are point estimates on modest samples, and the gaps between neighbouring models are not resolvable: the ensemble's own mean absolute error carries a bootstrap 95% interval of [9.2, 10.6] over 716 cells, and steerability rests on about 168 adjacent-band comparisons per model, giving a 95% interval roughly ±7 points wide. Read the table as a rough ordering, not a ranking; differences of a few tenths of a point carry no information.

Table 3: Selected models on the subjective battery, scored over all 16 questions. Error is mean absolute error in percentage points, and lower is better. Coverage is the share of cells the model answered; because error is computed only over answered cells, a model with low coverage is scored on an easier subset.

Model	Error	Steerability %	Mode collapse %	Coverage %
glm-4.7-flash	12.5	49.7	1.6	100.0
ministral-3-8b	12.7	61.4	0.0	84.4
minimax-m2.5	12.8	54.5	0.0	100.0
deepseek-r1 (reasoning)	13.5	64.1	0.8	100.0
kimi-k2-thinking	13.9	63.4	0.0	100.0

Model	Error	Steerability %	Mode collapse %	Coverage %
claude-sonnet-4.5	14.6	50.3	2.3	100.0
claude-opus-5	14.9	48.3	1.6	100.0
claude-opus-4-8	15.5	49.0	10.2	100.0
claude-sonnet-5	16.2	60.0	10.9	100.0
gemma-3-27b-it	17.8	49.7	25.8	100.0
gemma-3-4b-it	18.8	52.4	30.5	100.0

Two models fall below 95% coverage, ministral-3-8b on 84.4% and nova-lite on 83.3%. The effect is not that they are scored on an easier subset — both answer all 128 question-segment cells. What happens is whole-respondent dropout: 56 of ministral's 360 personas returned nothing at all, and 60 of nova-lite's. The effective sample per segment falls as a result, and falls unevenly, with ministral losing 14 of 60 in the 75 and over band. This raises the sampling-noise floor for those two models, so ministral's second place in Table 1 rests on a noisier estimate than the models around it.

The most recent frontier models, Opus 5, Opus 4.8 and Sonnet 5, rank in the middle of the distribution at 14.9 to 16.2 points, and are outperformed by substantially smaller models such as glm-4.7-flash. The two highest steerability scores belong to reasoning models, at 63 to 64%, but their intervals overlap those of several instruction-tuned models, so the ordering is not separable at this sample size. Claude Sonnet 4.5, at 14.6 points, attains lower error than the later Opus 5, indicating that within a single vendor family higher nominal capability is not associated with higher fidelity.

The correlation across the 26 models between mode collapse and mean absolute error is 0.66, which means the ranking is to a substantial degree a ranking of how much distributional spread a model emits rather than of how well it represents a population. Each model also runs at its provider's recommended temperature, from 0.3 to 1.0, which is itself a determinant of spread, although its direct correlation with error is weak (0.11). The claim we can support is therefore narrower than "capability does not predict fidelity": models differ mainly in the diversity of the answers they emit, this is partly a sampling-configuration artefact, and nominal capability does not track fidelity once that dominates. A fixed-temperature replication is a precondition for the stronger claim rather than an extension of it.

Mean error by size tier runs from 14.56 points for the large tier (n=9, sd 0.93) through 15.28 for frontier (n=4, sd 0.60) and 15.95 for tiny/small (n=8, sd 2.41) to 16.46 for mid (n=5, sd 1.31). The honest statement is that error is not monotone in size, not that the tiers are indistinguishable: large and mid sit about three standard errors apart. Two cautions apply. Tier is confounded with architecture, since almost every mixture-of-experts and reasoning model falls in the large tier, and total parameter count is a poor size proxy for such models in any case. And the boundaries are conventional; we checked that no single reassignment changes which tier ranks first, but moving one model out of nine cannot shift a mean far, so that is a weak check and we do not present it as more.

Three models give one identical answer to a quarter or more of segments, gemma-3-4b at 30.5%, gemma-3-27b at 25.8% and qwen3-32b at 24.2%, and all three are among the least accurate in the pool. Two of the most capable models are also substantially affected, Sonnet 5 at 10.9% and Opus 4.8 at 10.2%, while several small models exhibit no collapse. Collapse affects approximately a quarter of the pool, and it is predicted by neither size nor vendor.

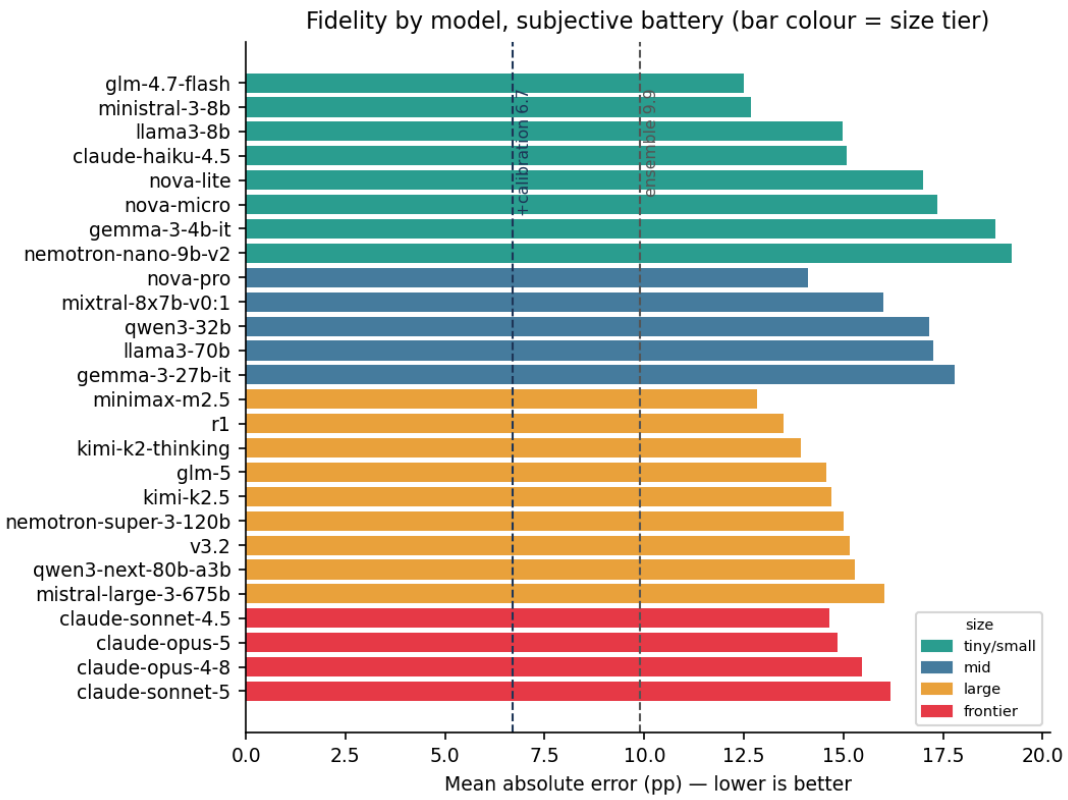

Figure 1: Fidelity by model on the subjective battery. Bar colour marks the size tier; dashed lines mark the ensemble and calibrated ensemble.

Figure 2 plots directional accuracy against absolute error, and separates two properties that a single error figure combines. Reasoning models recover the direction of demographic differences more reliably than their magnitude, and continue to do so at absolute error levels well above the best models in the pool.

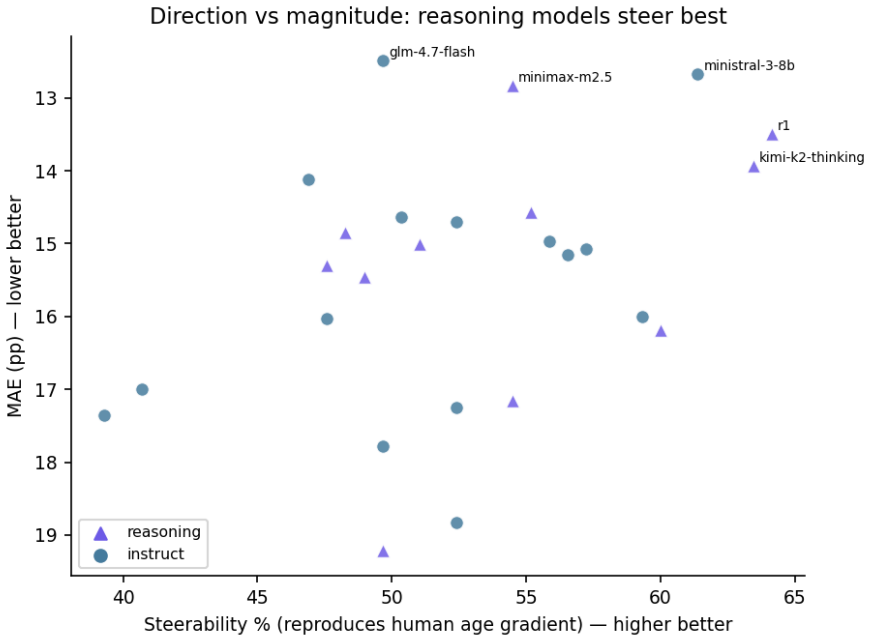

Figure 2: Direction against magnitude. Triangles are reasoning models; the vertical axis is inverted so better models sit higher.

4.3 Ensembling, calibration, and what they are worth

Combining models and correcting their shared bias improves the estimate substantially. The correction can only be trained and tested on questions that share an answer scale with at least one other question, so the whole comparison is computed on the 15 concern-scale questions. Scored on that common base, the best single model stands at 12.1 points, the all-model ensemble at 9.9, and the calibrated ensemble at 6.7, holding out each question in turn so that no question contributes to the correction applied to it.

Interpreting that sequence requires a lower bound on achievable error. With 60 personas per age segment and 180 per sex segment, a perfect sampler drawing from the true distribution scores 2.9 points through sampling variation alone, which no method operating on this panel can improve upon. The calibrated figure of 6.7 therefore corresponds to about 3.8 points of reducible error.

A second reference point comes from a procedure that uses no model output. Calibration requires the real human distribution for at least two questions sharing the scale, since that is what makes the per-option bias estimable. Given those distributions, the held-out question can be predicted from them directly. Averaging the other fourteen questions' real distributions within each segment scores 4.0 points, which is 2.7 points better than the calibrated ensemble on identical information.

Table 4: The ladder against its own reference points, on the 15 concern-scale questions. Every row is scored on identical cells, and every model row is a leave-one-question-out estimate.

Method	MAE (pp)
Sampling-noise floor (perfect sampler, this panel)	2.9
No model: mean of the other questions' human distributions	**4.0**
Calibrated all-model ensemble	6.7
All-model ensemble, uncalibrated	9.9
Best single model	12.1
Uniform across options	13.0

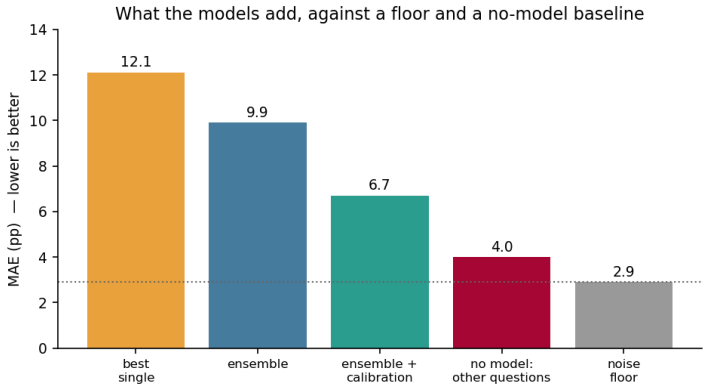

Figure 3: The alignment ladder on the 15 concern-scale questions, shown against the sampling-noise floor and the no-model baseline.

Any setting in which the calibration can be applied is by construction a setting in which the no-model procedure is available, since both consume human distributions for at least two questions on the scale. On this battery the cross-question structure of the answer scale predicts a held-out question more accurately than the ensemble does. The models improve on it for 3 of the 15 questions, and we can identify no property of those questions that distinguishes them in advance. Combining the two in equal weight yields 3.8 points, which improves on the no-model procedure by 0.2 points against a noise floor of 2.9, so the models carry some independent signal that we cannot separate from sampling variation at this sample size.

This bears on the interpretation of the correction rather than on its utility. The per-option bias exists and correcting it reduces error, but the correction operates by importing the scale's cross-question structure, and that structure is derived from human answers. Calibration is consequently a mechanism by which human data improves synthetic estimates, rather than a property of the models being corrected.

Averaging all 26 models reaches 9.9 points on a 17.3% CI hit rate, while averaging only the five best reaches 9.7 points on a lower 14.8% hit rate; after calibration the full average finishes ahead, 6.7

against 7.0. Composition therefore has little effect, and no model-selection procedure is required, which matters because selection would need the ground truth a deployment does not have. We note without explanation that the better-MAE ensemble has the worse CI hit rate; with hit rates of 15 to 17% in absolute terms, both are low, and the CI hit rate is in any case conservative for the reason given in §6.

Calibration is the larger of the two steps, and the bias it corrects is measured consistently across the pool. Response mass moves out of the two extreme categories and into the middle one in every model we tested, across eleven providers, four size tiers and temperatures from 0.3 to 1.0, with a displacement of 18.6 points on the modal option.

Table 5: Systematic bias by concern option, averaged over the 26 models and all segments, on the 15 concern-scale questions.

Concern option	Real %	Model %	Bias (pp)
Highly concerned	31.8	18.4	-13.4
Somewhat concerned	37.1	55.7	+18.6
Not very concerned	17.9	20.2	+2.3
Not concerned at all	6.4	3.3	-3.1
Don't know	2.9	1.3	-1.7
I don't know enough	4.0	1.2	-2.8

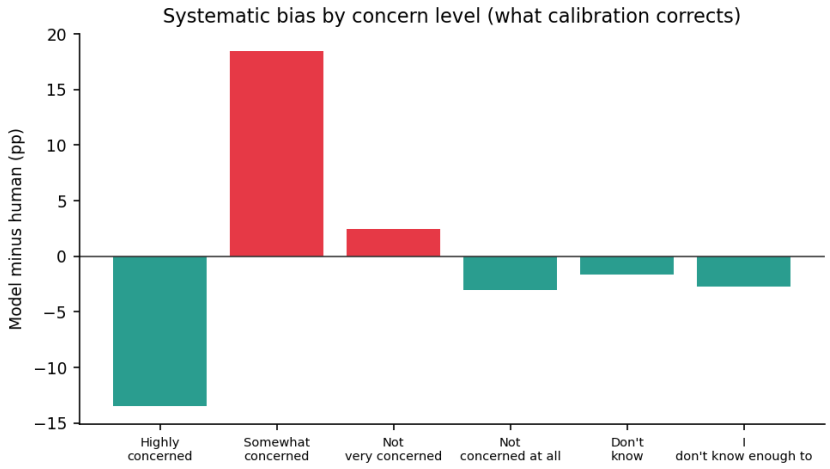

Figure 4: Model minus human, in percentage points, for each option on the concern scale. Bars above the line are options models over-select.

The bias varies by segment, and the largest errors appear there rather than in the pooled figures. Strong concern among older respondents is the worst understated, with models putting 15% of those aged 75 and over as highly concerned about ultra-processed food against a real 50%, a gap of 35 points. Among the young the error runs the other way, models placing 82% as somewhat concerned about food quality at ages 16 to 24 against a real 41%. This direction is worth noting against Santurkar et al. (2023), who find the opinions of respondents aged 65 and over among the most poorly reflected; our data reproduce that on UK ground truth.

Of the three interventions, grounding produces the smallest effect. A neutral block of real-world context improves some models, with the largest improvements among the least expensive. An editorialising block describing national sentiment degraded performance instead, amplifying the pessimism the models already exhibited and increasing collapse, which indicates that any grounding supplied to a persona should state facts without characterising them.

4.4 Ordering

Ranking the fifteen concern questions within a segment by the proportion selecting one of the top two options, the ensemble reproduces the published ordering at a Spearman correlation of 0.80 in the largest segment, with a permutation p-value of 0.0004 over 20,000 relabelings. Averaged across the eight scored segments the correlation is 0.77. Every figure reported before this point measures the level a method predicts for a single question; this measures the order it places questions in.

Table 6: Rank correlation between the ensemble's ordering of the fifteen concern questions and the published ordering, by segment.

Segment	Spearman rank correlation
Age 16–24	0.93
Age 25–34	0.93
Age 45–54	0.89
Age 35–44	0.87
Sex, female	0.80
Sex, male	0.72
Age 55–74	0.56
Age 75+	0.45

The result does not depend on ensembling. Across the 26 models individually the median correlation is 0.64, with 19 models above 0.5 and a range from 0.19 to 0.91. It also holds on the instrument where levels are least accurate: ordering the answer options within each circumstantial question reaches a mean correlation of 0.64 across 71 question-segment pairs, with 66% above 0.5, against a best-model error of 14.0 points on the same instrument. The equivalent figure on the subjective instrument is 0.83, with 94% above 0.5.

The no-model procedure of §4.3 produces no ordering. Predicting each question from the mean of the others assigns every question substantially the same level, with a standard deviation across the fifteen questions of 0.8 points, against 10.8 in the published data. The procedure that wins on level is therefore uninformative about rank, and the ensemble that loses on level recovers it.

Two properties limit how far this can be taken. The ensemble exaggerates the distance between questions, at a standard deviation of 22.0 points against the published 10.8, so the ordering is informative while the margins between adjacent questions are not. Accuracy also declines with the age of the segment, from 0.93 among 16 to 24 year olds to 0.45 among the over-75s, which is the same population in which the level errors are largest.

5. Discussion

An ensemble with calibration reaches 6.7 percentage points of mean absolute error against real UK attitudinal distributions. Placed against a sampling-noise floor of 2.9 points, that figure represents about 3.8 points of reducible error. Placed against a procedure that uses the same human inputs and no model output, it represents 2.7 points of additional error. The cross-question structure of a shared answer scale predicts a held-out question more accurately than 26 language models do.

That comparison concerns the accuracy of a predicted level. The models perform differently on the accuracy of a predicted ordering.

Ranking the fifteen concerns by the proportion expressing concern, the ensemble matches the real ordering at a Spearman correlation of 0.80 for the largest segment, with a permutation p-value of 0.0004, and at 0.77 averaged across the eight demographic segments. The result is not an artefact of ensembling: the median across individual models is 0.64, and 19 of 26 exceed 0.5. It also survives on the instrument where levels fail worst, where within-question option ordering on the circumstantial battery reaches 0.64 despite a best-model error of 14 points. The no-model procedure cannot produce a ranking at all, since it assigns every question substantially the same level, with a standard deviation across questions of 0.8 points against a real spread of 11.

Human data supplies the levels and the models supply the ordering, and the two are separable in practice as well as in measurement. A procedure with access to human answers on a shared scale already knows approximately what proportion of a segment will express concern, but not which of fifteen issues that proportion attaches to, and it is the second of those questions that the models answer.

Estimating a proportion from synthetic respondents carries error that we can measure and cannot remove, so a client requiring a number should obtain it from human respondents. Establishing which of a set of issues, propositions or price points ranks highest for a given demographic segment is a different task, and one the models perform at a correlation of 0.80 in minutes and at negligible cost. The blended design follows from the partition rather than from a preference for either method.

Two properties of the models constrain how far the ordering result can be pushed. The spread between questions is exaggerated, at a standard deviation of 21 points against a real 11, so the ordering is informative while the margins between adjacent items are not. Accuracy also degrades with age, from a correlation of 0.93 in the 16 to 24 segment to 0.45 among the over-75s, which is the same population in which the level errors are largest.

On model selection, tier means span under two points and are not monotone in size, so frontier capability purchases no measurable fidelity on this benchmark. The caveat from §4.2 applies: the ranking of models correlates with their output diversity at 0.66, and that diversity is partly a function of the sampling temperature each provider recommends.

6. Limitations

6.1 Scope of the claim

The result concerns demographic aggregates rather than individuals. We score distributions for an age band or a sex, and accuracy of that kind is compatible with poor accuracy for any single respondent. It covers one attitudinal domain measured on a single six-point scale, and one further instrument on a different scale. The correction we report is restricted to calibrations tested on

questions held out of the fit, since a correction fitted and evaluated on the same questions would not be applicable to a client's own instrument.

6.2 Measurement

The published bases from both sources are unweighted, so intervals derived from them ignore the surveys' design effects and are narrower than the true intervals. This makes the CI hit rate a conservative measure of agreement and accounts in part for the low absolute rates we report. The sources also publish age and sex as separate marginals rather than as a crossed table, with the consequence that region and socioeconomic class shape the realism of the personas without themselves being scored.

6.3 Procedure

Each model runs at its provider's recommended temperature, between 0.3 and 1.0, which confounds fidelity with sampling configuration. A fixed-temperature replication is the single change that would most improve the comparison. Answers shift with option order and phrasing, which the fixed schema mitigates without eliminating. Two models suffered substantial whole-respondent dropout, raising their effective noise floor unevenly across segments.

6.4 Reproducibility

Each persona answers each instrument once, with no fixed sampling seed and no repeated runs, so we cannot report run-to-run variance. Differences of a few tenths of a point between models are therefore not interpretable. Repeating a single model several times and reporting the spread would establish the resolution of the benchmark and is the first extension we would make.

6.5 Contamination

The FSA tracker data are from June 2026 and the models carry training cutoffs in 2025 and 2026, so we cannot exclude the possibility that some reference distributions appear in training corpora. The direction and magnitude of the errors are inconsistent with straightforward memorisation, since a model with access to the answers would not understate strong concern among the over-75s by 35 points, but we have not tested for contamination directly.

6.6 Population frame

The sampling frame is the England and Wales census, the ONS instrument covers Great Britain, and the FSA instrument covers England, Wales and Northern Ireland, as set out in §3.1. The panel is additionally uniform in age and sex rather than population-weighted, as set out in §3.3. Re-weighting the panel to real age shares moves the ensemble error by 0.01 points; the effect of the frame mismatches we have not quantified.

Finally, the results are confined to the United Kingdom and to English. The biases these systems carry, comprising variance compression, a pull toward socially acceptable answers and the flattening of small subgroups, set a limit on what any post-hoc correction can achieve.

7. Conclusion and Future Work

Persona-conditioned language models estimate UK attitudinal distributions to within 6.7 percentage points once ensembled and calibrated. Against a sampling-noise floor of 2.9 points, and against a procedure that uses the same human inputs with no model output and scores 4.0, that figure represents no improvement on the cheapest statistical use of existing survey data. The same models rank fifteen concerns against their true ordering at a correlation of 0.80, where the no-model procedure produces no ranking at all. The two results describe different properties, and separating them is the main contribution of this paper: synthetic respondents recover the ordering of a set of issues but not the level of any one of them.

Neither property holds for questions concerning a respondent's own circumstances. The best model remains 14 points from the published distribution, with every model in the pool biased in the same direction toward overstating hardship.

The applied conclusion is a division of labour between the two kinds of respondent. Synthetic respondents establish which of a set of issues, propositions or price points ranks highest for a demographic segment, at a cost and speed that a human panel cannot match. Human respondents establish the levels, and in doing so supply the cross-question structure on which the correction of synthetic estimates depends.

Widening the benchmark beyond food-related concern to the British Social Attitudes survey and to consumer batteries would establish whether the ordering result generalises across domains and answer scales, which is the claim most in need of replication. Running every model at a common temperature would separate differences in fidelity from differences in sampling configuration. Grounding a persona in retrieved responses from similar people, and fine-tuning on real UK survey data, are the two methods with the strongest evidence of further gains, the latter shown by Suh et al. (2025) to narrow the model-to-human gap by up to 46% and to generalise to unseen surveys. A standing human panel is positioned to supply the reference data that all four require.

Appendix A. Full model results

All 26 models scored on the subjective battery, across all 16 questions, ordered by error. Columns are error, steerability, mode collapse and coverage, all in percent except error in percentage points; "Temp" is the provider's recommended sampling temperature and "Output" the rung of the structured-output ladder the model settled on.

Model	Tier	Error	Steer	Coll	Cov	Temp	Output
glm-4.7-flash	tiny/small	12.5	49.7	1.6	100.0	0.6	native
ministral-3-8b	tiny/small	12.7	61.4	0.0	84.4	0.7	native
minimax-m2.5	large	12.8	54.5	0.0	100.0	1.0	native
deepseek-r1	large	13.5	64.1	0.8	100.0	0.3	json
kimi-k2-thinking	large	13.9	63.4	0.0	100.0	1.0	native
nova-pro	mid	14.1	46.9	0.0	100.0	0.7	tool
glm-5	large	14.6	55.2	0.0	99.7	0.6	native
claude-sonnet-4.5	frontier	14.6	50.3	2.3	100.0	1.0	native
kimi-k2.5	large	14.7	52.4	0.8	98.6	0.6	native
claude-opus-5	frontier	14.9	48.3	1.6	100.0	1.0	tool

Model	Tier	Error	Steer	Coll	Cov	Temp	Output
llama3-8b	tiny/small	15.0	55.9	4.7	100.0	0.6	json
nemotron-super-3-120b	large	15.0	51.0	0.0	99.7	0.6	native
claude-haiku-4.5	tiny/small	15.1	57.2	3.9	100.0	1.0	native
deepseek-v3.2	large	15.2	56.6	0.0	100.0	0.3	native
qwen3-next-80b-a3b	large	15.3	47.6	1.6	100.0	0.7	native
claude-opus-4-8	frontier	15.5	49.0	10.2	100.0	1.0	tool
mixtral-8x7b-v0:1	mid	16.0	59.3	1.6	100.0	0.7	json
mistral-large-3-675b	large	16.0	47.6	0.0	100.0	0.7	native
claude-sonnet-5	frontier	16.2	60.0	10.9	100.0	1.0	tool
nova-lite	tiny/small	17.0	40.7	4.7	83.3	0.7	tool
qwen3-32b	mid	17.2	54.5	24.2	100.0	0.7	native
llama3-70b	mid	17.3	52.4	3.1	100.0	0.6	json
nova-micro	tiny/small	17.4	39.3	0.0	100.0	0.7	tool
gemma-3-27b-it	mid	17.8	49.7	25.8	100.0	1.0	native
gemma-3-4b-it	tiny/small	18.8	52.4	30.5	100.0	1.0	json
nemotron-nano-9b-v2	tiny/small	19.2	49.7	10.9	98.3	0.6	native

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., and Wingate, D. (2023). *Out of One, Many: Using Language Models to Simulate Human Samples.* Political Analysis, 31(3), 337–351. arXiv:2209.06899.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., and Larson, J. M. (2024). *Synthetic Replacements for Human Survey Data? The Perils of Large Language Models.* Political Analysis, 32(4), 401–416.

Choi, E. C., Kim, Y., Pugalenthi, P., Chen, H.-E., and Huang, B.-R. (2026). *Beyond the Mean: Three-Axis Fidelity for Aligning LLM-Based Survey Simulators from Small Pilot Data.* arXiv:2606.28963.

Durmus, E. et al. (2023). *Towards Measuring the Representation of Subjective Global Opinions in Language Models (GlobalOpinionQA).* arXiv:2306.16388.

Jia, M., Chen, Y., Sharma, D., and Diaz-Rodriguez, J. (2026). *When Can Digital Personas Reliably Approximate Human Survey Findings?* arXiv:2605.10659.

Park, J. S. et al. (2024). *Generative Agent Simulations of 1,000 People.* arXiv:2411.10109.

PyMC Labs (2025). *Synthetic Consumers and AI Market Research* (Colgate-Palmolive collaboration).

Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., and Eichstaedt, J. C. (2024). *Large Language Models Show Human-like Social Desirability Biases in Survey Responses.* PNAS Nexus. arXiv:2405.06058.

Santurkar, S. et al. (2023). *Whose Opinions Do Language Models Reflect? (OpinionQA).* arXiv:2303.17548.

Suh, J., Jahanparast, E., Moon, S., Kang, M., and Chang, S. (2025). *Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions (SubPOP).* ACL

2025. arXiv:2502.16761.

Taday Morocho, E. E., Cima, L., Fagni, T., Avvenuti, M., and Cresci, S. (2026). *Assessing the Reliability of Persona-Conditioned LLMs as Synthetic Survey Respondents.* arXiv:2602.18462.