

Synthetic personas: a thinking tool, not a respondent

What AI customers are good for, where they fail, and four rules before you trust one

Falk Uebernickel

University of St.Gallen (HSG)

Human thinking, AI-assisted writing: the ideas and the argument are the author's; the text was written together with AI.

Abstract. Synthetic personas, language models prompted to answer as customers, are being bought faster than they are being tested: in a 2025 survey of market researchers, 81 per cent use or plan to use synthetic data, while 31 per cent rate the value of AI-made data as great. This note, written from a keynote at the E-Commerce Best Practice Day 2026, asks whether synthetic personas can replace the people we ask. Our answer is yes as a thinking tool, no as a respondent, and the difference is measurable. We separate the description (persona, synthetic persona) from what answers and is counted as data (participant, respondent); synthetic panels sell the second as if it were the first. Six uses pay off today, and none counts a synthetic answer as data. We then report a preregistered experiment with 116,570 answers from five frontier models on the anchoring items of Many Labs 1 and three invented twins. Where the model knows the answer, the typical answer does not move with the anchor, while humans move about half the way; where it cannot know, on two of the three invented items one number in the question moves the answer; and in the most extreme case 1,198 different profiles give the same number 1,171 times. A richer profile barely helps. From this we derive four rules, whether you build or buy: check diversity, freeze the model, calibrate against people, label the boundary.

1 The market runs ahead of the test

Before you simulate customers, you should have met them. The talk opened with a photograph of a synthesis session after needfinding interviews in Nairobi in 2016; the decisive insight of that project was on no sticky note until we had been there. That effort is why the shortcut through a model tempts, and it is the yardstick for what follows.

AI now sits at both ends of commerce. On one side the customer asks a model about the brand: generative search, agentic commerce, chatbots. On the other side the brand asks a model in place of the customer: synthetic personas, synthetic panels, digital twins. Whoever fills both ends with the same model ends up measuring model against model. This note is about the second end.

The second end is already a market. In a survey of more than 170 market research practitioners and users reported in the *Harvard Business Review*, 90 per cent use or plan to use generative AI in their insights work, 81 per cent use or plan to use synthetic data, and more than 40 per cent are already experimenting with digital twins (Korst et al., 2025). In the same survey, only 31 per cent rate the value of AI-made data as great, one of the lowest scores in the whole study, and 77 per cent name biased results as their concern, the most frequent one (Figure 1). Adoption runs years ahead of validation, and the people least happy with the result are the ones who already use it.

The question of this note is therefore not whether synthetic personas will be used. They are. The question is: can synthetic personas replace the people we ask? Our answer in one line: yes, as a thinking tool; no, as a respondent; and the difference is measurable.

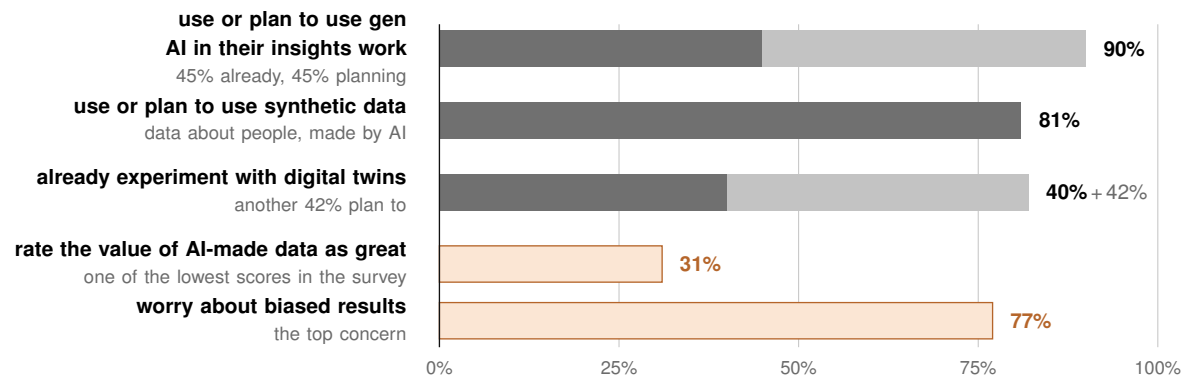


Figure 1. Everyone is buying it, a third is happy with it. Share of market research practitioners and users (n > 170) in the GBK Collective survey reported by Korst et al. (2025). Dark: already doing it; light: planning to. Orange: the two answers about the result rather than the intent.

Table 1. Real people validate, the model answers. Columns: where it comes from. Rows: description or instance. The bottom row is what synthetic panels swap.

	From people	From a model
THE DESCRIPTION written down; nothing answers yet	Persona. An archetype distilled from research with participants. A description of a customer, not a customer.	Synthetic persona. A profile written for a language model to play: age, job, region, traits. Written, not observed.
WHAT ANSWERS and whose answer is counted as data	Participant. A real person in a study: interviewed, observed, surveyed. Only they can validate. In marketing: the customer.	Respondent. The description instantiated: a model runs it and the answer is counted as if a person had given it.

2 Four terms, one swap

Much of the confusion in this market comes from one word doing two jobs. A *persona*, in the tradition of Cooper (1999), is an archetype distilled from research with participants: a description of a customer, not a customer. A *synthetic persona* is also a description: a profile written for a language model to play, with age, job, region and traits. Written, not observed. Neither of the two answers anything by itself.

What answers, and whose answer is counted as data, is a different row of the table (Table 1). From people, it is the *participant*: a real person in a study, interviewed, observed, surveyed. In marketing this is the customer. From a model, it is the *respondent*: the description instantiated, a model runs it, and the answer is counted as if a person had given it. Synthetic panels sell the bottom right of the table as if it were the bottom left. Everything in this note turns on whether that swap holds.

The same description can play three roles, and the question decides which. Asked “what did our customers say?”, the synthetic persona is a *condensate*: it summarises real interviews and is valid as long as the answer is in the data. Asked “what would you object to?”, it is a *thinking tool*: it keeps the team close to the customer, and its value arises in the conversation, not in a number. Asked “what would you pay?”, it is an *instrument*: the answer gets counted, and from here on the base model answers in costume. Only the third role demands validation against real people, and that is the role synthetic panels sell (Uebernickel, 2026).

3 Value: six uses, none of them a respondent

The value is real; we use it ourselves. Synthetic personas buy three things fieldwork cannot: speed (a first set of personas before the kick-off meeting ends; no recruiting, no scheduling), dissent without politics (a persona that objects has no manager in the room and nothing to lose), and scale for the small (in an automated build, every screen, label and flow can be

COUNTERPOINT

The swap may hold more often than this note allows. Argye et al. (2023) conditioned GPT-3 on demographic backstories and found that it reproduced the attitude patterns of the corresponding human subgroups closely enough to coin the term algorithmic fidelity. If a model panel passes on attitudes, is anchoring the right test, or merely the hardest one?

Table 2. Six uses pay off today. Each buys speed, dissent or scale that fieldwork cannot give. After Newhook (2025), Rosala and Moran (2024) and Korst et al. (2025); the review use from our own practice.

Use	What it does	Buys
Draft	Shopper personas before the first workshop: a starting point the team can argue with, then correct with research.	speed
Pre-test	Questionnaires and interview guides: ambiguous questions and missing options show up before real people spend their time.	speed
Rehearse	Objections before a pitch or launch: a persona per stakeholder role raises the questions the room will ask.	dissent
Review	Every increment of an automated build: microcopy, flows, error states, challenged as they are generated, at machine pace.	scale
Interview	Let the AI ask real people, at scale: it moderates, probes and follows up; the people answer. Nothing to hallucinate.	scale
Condense	Interview transcripts into a persona: the one grounded use. It summarises what participants actually said.	fidelity

reviewed as it is created; no participant could keep up). Table 2 lists the six uses that pay off today. Each replaces an assumption or saves participants' time.

Two of the six deserve a comment. The *interview* use turns the arrangement around: the model does not answer, it asks. Korst et al. (2025) describe AI-moderated interview platforms that probe participants with follow-up questions, and relay a client's experience that participants open up more to an AI interviewer; that is a user's claim, not a measurement. The *condense* use is the only one where the synthetic persona is grounded; the moment someone asks that condensed persona about a new product or price, it is a respondent again, and the grounding has ended without anyone noticing. We return to that boundary in Section 7.

Everything so far replaces an assumption, saves time or sharpens a question. Nothing so far counts a synthetic answer as evidence. The value ends exactly where measurement begins.

4 Limits: what independent tests keep finding

Asked as a respondent, it is not the synthetic persona that answers but the model in costume. Practice has reported four failure modes, and independent tests keep finding the same flaw: too little reaction, too much approval.

- **Narrow.** They know only the expected average: no context, no rare groups, no networks. The model has read more than any human, but it has still read only what people wrote down. Why a basket gets abandoned (the tram arrives, the child cries), customers who leave no trace (luxury buyers, a market of nine million in three languages), and who really decides (the partner who pays, the buyer at the supplier): none of this is in the text.
- **Flattering.** They like every concept and never drop out. The Nielsen Norman Group re-ran three of its own interview studies with synthetic users: real participants said things like "I completed three of seven courses" and explained why, synthetic users claimed to have completed everything; real participants weighed interest against barriers, synthetic users praised every concept. The verdict: too shallow to be useful, fit for desk research and hypotheses, never for decisions (Rosala and Moran, 2024).
- **Unreactive.** They move too little when a price or an anchor moves, so demand curves come out implausible. Korst et al. (2025) ran pricing experiments with digital respondents and found that the model had trouble treating the price as random, assumed other things had changed with it, and showed little change in behaviour as the price varied. The largest review to date, 28 studies, finds that what people do with their own money, ownership, sunk costs, mental budgets, the models do not reproduce (Sarstedt et al., 2024); in a replication of two corporate studies, Arora et al. (2025) find that GPT-4

COUNTERPOINT

The dissent may be flattery in disguise. Sharma et al. (2024) show across five AI assistants that preference-tuned models tend to tailor answers to the user's stated views, and that human raters often prefer such agreement. The objection you ask a persona for is still an answer it thinks you want. Can a tool trained to agree be relied on to object?

COUNTERPOINT

The glass is nine tenths full. In Bain's own backtest, digital twins built from historical respondent-level data replicated about 90 per cent of the key outcomes of a large conjoint study, with "preliminary price sensitivity curves that showed promise" (Bain & Company, 2026). A method that reproduces nine outcomes in ten is not nothing. Which tenth would you have paid a field study to find?

matches the direction of human responses but not their magnitude without further calibration.

- **Invisible boundary.** They answer invented questions as readily as known ones, and nothing in the output marks the edge of the data.

The first two were documented with real participants. We tested the mechanism behind the third and the fourth, and our data speak to the first as well. No single number tells you whether your own panel is affected, which is why you cannot correct it by hand, and why we measured it.

5 The experiment: do synthetic personas react, or recall?

Anchoring is a clean test, because the reaction can be measured. Give people a number before a question (“is the height of Mount Everest more or less than 45,500 feet? How high is it?”) and their estimate moves towards that number (Jacowitz and Kahneman, 1995; Tversky and Kahneman, 1974). For a retailer, an anchor is the crossed-out price next to the promotional one. If synthetic personas react to context the way people do, they will move the way thousands of humans moved on the same questions. If they merely recall, they will ignore the anchor wherever they know the answer and follow it wherever they do not.

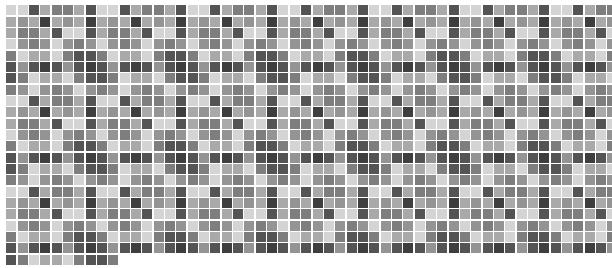
We put the anchoring items people answered in Many Labs 1 (Klein et al., 2014), a replication project with 6,344 participants in 36 samples, to five frontier models in the respondent role: Claude Sonnet 4.6, GPT-5.1, Gemini 2.5 Pro, Llama 4 Maverick and DeepSeek v3.1. Three items are known: babies born per day in the United States (anchors 100 and 50,000; true value about 11,000), the height of Mount Everest in feet (2,000 and 45,500; true value 29,032), and the population of Chicago (200,000 and 5,000,000; true value about 2.7 million). Three are invented twins with identical anchors and sentence structure but nothing to look up: the daily sales of a “Culina X2” kitchen appliance, the height of “Mount Veyra”, the population of “Halvert, Texas”. Each synthetic persona was a fresh, stateless call at temperature 1.0, its profile drawn from US census quotas (age, gender, education, region, 14 occupations, 8 consumer types), with three prompt depths: demographics only, plus psychographics, plus a behavioural instruction telling the model that real people answer quickly and intuitively, “are influenced by whatever numbers or context they just saw” and “give rough round estimates”. That instruction names the mechanism under test, so any anchoring it produces may partly be compliance. Study 1 crossed 5 models, 3 depths, 6 items, 2 anchors, 200 personas and 2 runs; Study 2 repeated the items with two instructions, recall the fact or answer intuitively. Together the two studies yielded 116,570 valid answers (70,271 of 72,000 and 46,299 of 48,000 planned calls; 2.9 per cent excluded under preregistered rules). Hypotheses, items, prompts and the analysis were frozen on 7 June 2026 at 15:46, before the first confirmatory call, and registered on OSF (ADYPE); the data were collected on 7 and 8 June 2026 (Uebersnickel, 2026).

Two points of honesty before the results. The effective sample for generalisation is six questions, not 116,570 answers; precision within a model is not generality across tasks. And the human benchmark is external: Many Labs 1 differs in population, platform and instructions from our protocol, and there is no human data at all on the invented items. A matched human arm on identical stimuli is preregistered and the decisive next step.

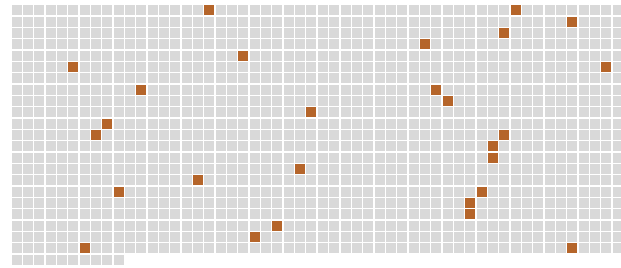
6 Finding one: a thousand profiles, one voice

Figure 2 shows the most extreme of the 60 model $\times$ item $\times$ anchor cells, and it is the one to remember. Claude Sonnet 4.6 was asked, as 1,198 different personas, how many Culina X2 devices sell per day after seeing the anchor 50,000. The profiles differ in every dimension we sampled. The answers do not: 1,171 of 1,198 say exactly 500 a day. The 65-year-old nurse from the Midwest, thirty years on the ward, price-conscious and frugal, says 500. The 29-year-old accountant from the Northeast who pre-orders every new device says 500. So

1,198 profiles · shaded by consumer type



1,198 answers · coloured by estimated daily sales



8 consumer types, 14 occupations, 4 age groups, 4 education levels, 4 regions ■ 500 a day · 1,171 personas ■ other values · 27 personas

Figure 2. 1,198 synthetic profiles give one answer. Claude Sonnet 4.6, invented item Culina X2 (daily unit sales), high anchor 50,000 a day, all three prompt depths and both runs pooled. Each tile is one persona. Counts are exact (1,171 of 1,198 answer exactly 500); the position of the 27 other tiles and the shading of the profiles are schematic. The most concentrated of the 60 model × item × anchor cells; in the same cell Gemini's most frequent answer covers 26 per cent. Source: exploratory analysis of the preregistered study (Uebernicket, 2026), OSF ADYPE.

Table 3. A model panel runs dry after a handful of answers. Number of distinct answers, most frequent first, needed to cover the given share of respondents. Humans on the real Everest, high anchor, Many Labs 1; models on Culina X2, high anchor, about 1,200 personas each; not a matched comparison. Source: exploratory analysis of the preregistered study (Uebernicket, 2026); Klein et al. (2014).

Share of respondents covered	Humans	Claude	GPT-5.1	Gemini	Llama	DeepSeek
Half	9	1	2	4	1	5
Nine in ten	121	1	6	18	6	32
95 per cent	258		not reported for the models			

do 1,169 others. Whatever the profile says, the same model answers, and it answers the same.

Claude is the extreme case; the others concentrate too. Table 3 counts how many distinct answers are needed to cover half, and nine in ten, of the respondents in a panel: humans on the real Everest need 9 and 121 and keep adding new ones; the models on the invented Culina X2 are exhausted after one to five answers for half, and after one to thirty-two for nine in ten. The comparison is not like for like: the humans answered a known item in a free format, the models an invented one in a forced number format, and the samples differ in size; it shows the shape of a model panel, not a matched test. Still, a demand curve drawn from a model panel has a handful of points, not a thousand. A persona may answer the same every time, that is what an archetype is for. But these are 1,198 different profiles, sold as a panel, and they answer like one. Whatever you build on such a spread, segments, quartiles, demand curves, rests on a spread that does not exist.

The obvious repair is a richer profile, and it barely helps (Figure 3). We measured the effective number of distinct answers, the exponential of the Shannon entropy, on the three known items pooled, about 2,400 answers per cell. Adding psychographics to demographics barely moves anything: Claude from 6.1 to 5.4, GPT-5.1 from 5.4 to 5.0, Gemini from 27.2 to 25.5, Llama from 11.3 to 10.4; only DeepSeek moves, from 10.6 to 18.4. The behavioural instruction helps two models, GPT-5.1 to 25.1 and DeepSeek to 42.7, and no one else; since that instruction tells the model to be influenced by the numbers it just saw, part of the gain may be compliance rather than a disposition. Humans on the same items come to 153. A behavioural instruction is hygiene, not a cure. On the preregistered dispersion measure, the standard deviation, this collapse did not even register, because a few outliers dominate it; entropy and effective number are exploratory measures, reported as such.

7 Finding two: where the model knows, it does not react

The anchoring index of Jacowitz and Kahneman (1995) measures how far the typical answer moves from the low anchor to the high one: 0 means the median does not move, 1 means it

COUNTERPOINT

The one voice may be an artefact of how it was asked. Meister et al. (2025) find that models describe a group's opinion distribution better than they reproduce it when sampled one persona at a time. A panel elicited as a distribution might show the variety a thousand single calls hide. Is the one voice the model's, or the sampling method's?

COUNTERPOINT

Perhaps the profile was never going to spread the answers. Hu and Collier (2024) show that the demographic and attitudinal variables used in persona prompts explain less than 10 per cent of the variance in human annotations, and that persona prompting helps only where they predict something. If age, job and region explain so little in people, why should they spread a model?

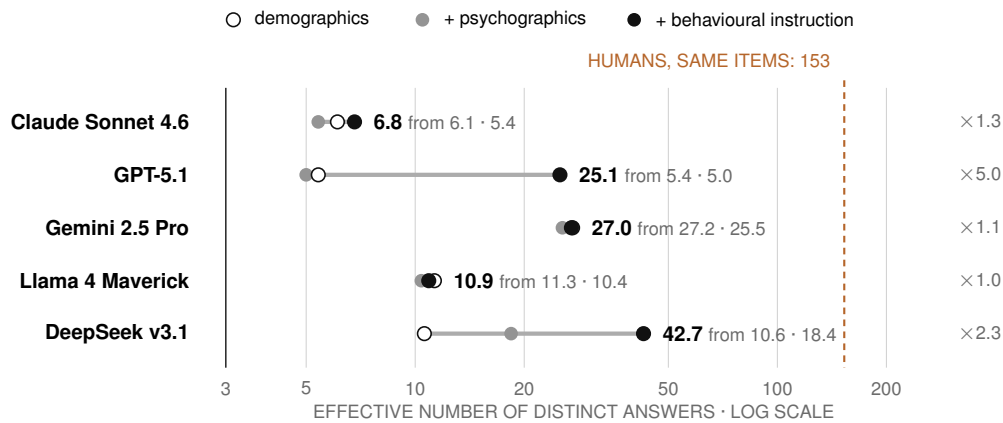


Figure 3. A richer persona profile creates little diversity. Effective number of distinct answers (exponential of the Shannon entropy), three known items pooled, about 2,400 answers per cell. Open circle: demographics only; grey: plus psychographics; black: plus behavioural instruction; right: ratio of instruction to demographics. Humans: Many Labs 1, same items, 16,520 answers. The bare-question baseline (exploratory, high format failure) is left out; the human and model cells are not matched in size or format. Source: exploratory analysis of the preregistered study (Uebersnickel, 2026), OSF ADYPE; Klein et al. (2014).

travels the full distance between the anchors. On the three known items, humans in Many Labs 1 travel about half the way: 0.61 for babies, 0.53 for Everest, 0.50 for Chicago. The five models' typical answer does not move: an index between 0.00 and 0.02, every model, every known item (Figure 4, left). Their medians sit on the true value, 11,000 births, 29,032 feet, 2.7 million residents. They are right, but they do not judge; they look it up. The median is not the whole story: a minority of answers jumps to the anchor, which is why the standardised effect sizes on known items run from 0.02 to 0.38 rather than zero; every one of them still lies far below the human 1.79 to 2.42 (Uebersnickel, 2026).

On the invented twins the picture inverts (Figure 4, right). Mount Veyra does not exist; the anchors are those of Everest. Personas that saw 2,000 feet and personas that saw 45,500 feet form two separate groups in every model, and the distance between them is the anchor effect: Claude's median moves from 15,000 to 28,000, GPT-5.1's from 7,320 to 29,000, Gemini's from 13,500 to 28,500, Llama's from 12,000 to 28,000, DeepSeek's from 11,000 to 23,000. Humans on the real Everest moved from 7,000 to 36,825. Two details matter. The high-anchor answers land on Everest's height, 28,000 to 29,000 feet: with nothing to look up, the model reaches for the nearest fact it knows. And the movement is a single mechanical shift of one voice, not a population spreading between the anchors: on Mount Veyra the models' anchoring index runs from 0.20 to 0.44, against 0.53 for humans on the real Everest, a cross-item comparison, since no human data exist for Veyra. The third invented item, the town of Halvert, Texas, adds a caveat: there the anchor did little. The models give a typical small-town population whatever number they saw, and three of the five moved slightly the wrong way (Claude's median from 5,000 to 800). So two of the three invented twins show the effect.

Study 2 tested the mechanism directly (Figure 5). The same Everest question was asked with two instructions: recall the fact, or answer intuitively and forget what you know. Told to recall, every model sits at the true height almost without exception (96 to 100 per cent of answers in the 28,000 to 30,000 band). Told to forget, GPT-5.1 drops to 70 per cent, Gemini to 67, Llama to 83, DeepSeek to 89; Claude stays at 100 per cent. Claude cannot forget. And exactly where recall gives way, the anchor starts to work: on the known items the pooled effect rises from 0.03 to 0.14 (d , log scale), carried by Llama and DeepSeek, while on the invented items the same instruction changes nothing; the interaction is +0.17 with a 95 per cent interval of [0.09, 0.24]. Recall, not judgement, is what blocks the anchor. Even released, the effect stays far below the human band, so the override is not something a prompt reverses.

Put the two findings together and the respondent role fails from both sides. On a known

COUNTERPOINT

Not reacting may be what an informed person does. Smith et al. (2013) found in four studies that people who know more about a quantity anchor less, and that giving participants knowledge reduced the effect. A model that knows Everest's height and ignores the anchor behaves like a well-informed respondent, and the Many Labs participants like uninformed ones. Which of the two do your customers resemble on the questions you are paying to ask?

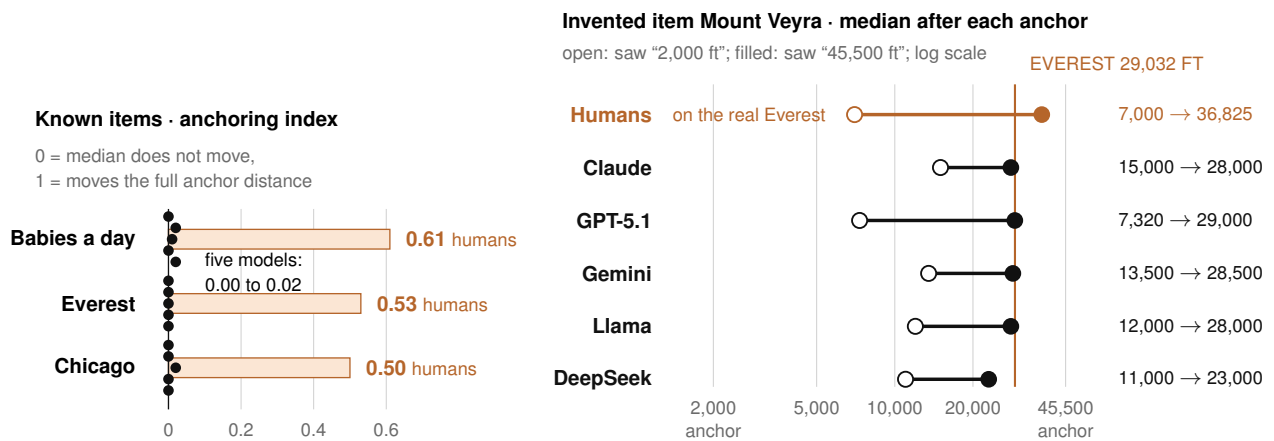


Figure 4. Where the model knows, it does not react; where it cannot know, one number in the question moves the answer. Left: log-scale anchoring index after Jacowitz and Kahneman (1995) on the three known items; orange bars: humans in Many Labs 1 (index computed from the open data); black dots: the five models (0.00 to 0.02, printed stacked). Right: median estimate after the low and after the high anchor for the invented Mount Veyra, each model; humans shown on the real Everest for scale (medians 7,000 and 36,825). No human data exist on invented items. Source: exploratory analysis of the preregistered study (Uebersnickel, 2026), OSF ADYPE; Klein et al. (2014).

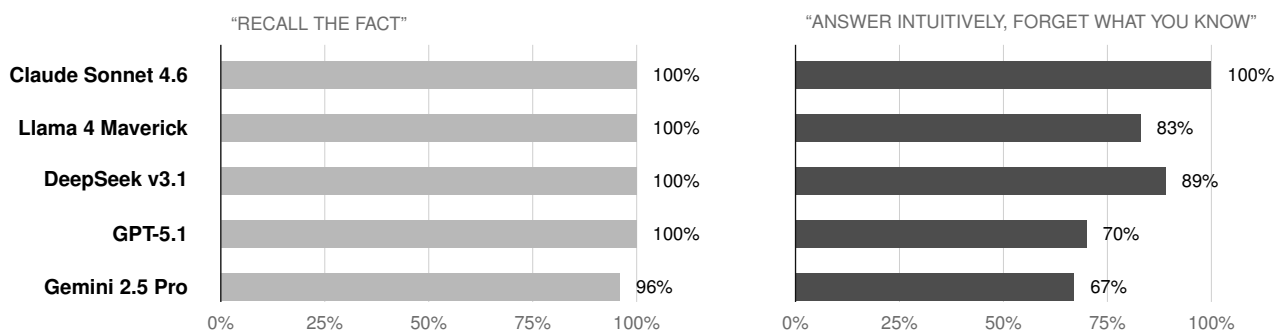


Figure 5. Told to forget the fact, the model anchors again. Study 2, Everest only, 7,902 valid answers, both anchors pooled: share of answers at the true height (28,000 to 30,000 ft) under the two instructions. Study 2 in total: 46,299 answers on six items; interaction +0.17 log- d [0.09, 0.24]. Source: Uebersnickel (2026), preregistered amendment, OSF ADYPE.

question the typical answer does not move, because it is looked up. On an invented one it tends to follow the number in the question and lands on the nearest fact. Neither is judgement. The more the model knows, the less it judges; this is an effect within a model across questions, not a ranking of models, and on our descriptive human-likeness index the model nearest to human behaviour is GPT-5.1.

The boundary is invisible. The practical hazard is that none of this is visible in the output. Picture one synthetic persona, condensed from twenty customer interviews, and two questions in the same chat window. "What did our customers say about their coffee routine?" The answer is precise, cites recurring themes, reads like a clean summary. It is one: the data cover it. "What would this customer pay for the new concept?" Same voice, same confidence: \$4.99. No hint that the data ended one question ago; the number goes into the deck. The vignette is an illustration, not a measured case. A synthetic persona does not say "I don't know": the protocol asked for a number in a fixed format, and the models supplied one in about 97 per cent of calls, on invented items as readily as on known ones, with nothing in the output to mark the difference. The role switch happens in the same window, and nobody validates at the boundary, because in the workflow as lived there is no boundary to see.

Where the data boundary is invisible and the stakes are high, a person decides, not an average: in health, where a wrong assumption about patient behaviour means a missed diagnosis; in finance, where saving and debt are learned in conversations about fear; and for

groups underrepresented in the training data, exactly the voices a synthetic persona erases and a project meant to amplify. Rosala and Moran (2024) draw the same line for niche populations.

8 Rules: four, whether you build or buy

Trust is a measurement, not an attitude. The seven rules in the paper (Uebernickel, 2026) condense, for practice, into four, and each comes with a question to put to a vendor in writing.

1. **Check diversity.** Ask for the frequency table of one key question. If one value covers more than 40 per cent of the answers, the panel is a single voice: no segments, no sizing on it. A share threshold alone is not enough: our most concentrated model gave only 28 distinct values on the known items with a modal share of 29 per cent, so also compare the number of distinct answers with a human benchmark on the same question. The question to the vendor: how many distinct answers, and show me the table.
2. **Freeze the model.** The models are not interchangeable: on our human-likeness index they rank GPT-5.1 > DeepSeek > Llama > Gemini > Claude, and the models that spread out under a behavioural instruction (GPT-5.1, DeepSeek) only partly overlap with those that anchor again when told to forget (Llama, DeepSeek). Pin the version, write the result next to the model name, re-check on every change. To the vendor: which model, which version, which date, and what happens to my numbers when the version changes next month.
3. **Calibrate against people.** A behavioural instruction is hygiene, not a cure. Only the comparison with real answers to the same questions tells you how far to trust the persona, and “internally validated” is not an answer. To the vendor: how was accuracy measured, against which human study, on which questions, by whom.
4. **Label the boundary.** Mark every number as from participants or from the model, and record which questions each answered. Model numbers are hypotheses, not evidence. To the vendor: which questions did participants answer, and which did the model answer alone; and what does the panel not cover: where do you say “we do not know”?

If you want to test it yourself. The recipe is in the paper and takes about an hour: one known anchoring item with a published human effect, one invented twin, two anchors, about a hundred personas per anchor and item, then compare the movement with the published human value. A model that collapses on the known item and under-reacts on the invented one behaves as ours did. For the invented twin you compare model against model; the human comparison exists only for the known item. Most teams will get further with the four questions above.

Where they fit. Scope and phase tell where synthetic personas belong (Table 4). Early phase and small increments: synthetic. Scope decisions and new ground: participants. The workflow that follows is human at both ends: interview and observe real people (human); let the AI read the transcripts and analytics, find patterns and write the persona (AI); let that persona object in the team as a dissenting voice (AI); and take every hypothesis that comes out of it back to the field before it goes into a deck (human).

Monday. Three uses and one rule are enough to start. Allowed from tomorrow: proto-personas before the workshop, a questionnaire pre-test, objections rehearsed before the board meeting; in a team account, without customer data in the prompt. Not allowed: a number from a persona in a presentation, no price, no conversion, no segment size; the price test stays with real customers. The one line: every persona output carries the note “model, not participants”, plus model name and version. For the discussion in your own team: which number in your last deck came from a model, and who would have noticed?

COUNTERPOINT

Calibration may be a repair, not a verdict. Brand et al. (2026) report that baseline GPT estimates of willingness to pay were often wrong-signed or about threefold inflated, and that fine-tuning on human data corrected them, even for held-out attributes, though not across product categories. The twin that fails the test may pass after one calibration study. Is the rule “never”, or “not yet calibrated”?

COUNTERPOINT

Grounding may rescue the respondent role. Park et al. (2024) built agents from two-hour interviews with 1,052 real people; the agents reproduced the participants’ General Social Survey answers at 85 per cent of the participants’ own test-retest consistency. That is a respondent grounded in a person, not a costume. Is the problem the respondent role, or the thin profile behind it?

Table 4. Scope and phase tell where synthetic personas fit. Shading: share of the work synthetic personas can carry, from participants (orange) to synthetic (grey). A sensitive scope (health, money, rarely seen groups) moves every cell one step towards participants.

	Before the project	During: small increments	During: scope decisions	After the project
Known scope documented, standard	Draft and prepare. Three shopper personas for a checkout redesign.	Small increments. Each microcopy change in the checkout.	Scope decisions. A price test, with real customers.	Condense and monitor. A persona from the twenty interviews; a monthly self-test.
New scope undocumented, rare, new	Prepare fieldwork. Screening questions for patients with MS.	Small increments. Each screen reviewed, then checked with patients.	Scope decisions. Interviews with patients about daily routines.	Ground, then label. A patient persona from the study; all else labelled as model output.

9 Limitations

The experiment measures one bias family, anchoring on numerical estimates with a right answer; whether it holds for preference tasks, where there is nothing for knowledge to override, is the next study. Generalisation is across items, and there are six; per-model interaction tests are underpowered. There is no matched human arm: Many Labs 1 is an external benchmark and the human anchoring index was computed from its open data; the results are a snapshot of June 2026 model builds. The concentration measures (mosaic, effective number) are exploratory; the preregistered dispersion hypothesis, on the standard deviation, was not supported. The 40 per cent threshold in Rule 1 is a heuristic (the paper uses roughly half); neither is benchmarked against a human panel. And the vignette of the invisible boundary is an illustration; whether people notice the role switch in practice has not been measured. None of this changes the practical line: use synthetic personas to think, test them before you count them, and keep the people you ask.

References

- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.
- Arora, N., Chakraborty, I., and Nishimura, Y. (2025). AI–human hybrids for marketing research: Leveraging large language models (LLMs) as collaborators. *Journal of Marketing*, 89(2):43–70.
- Bain & Company (2026). Synthetic customers earn their stripes. Bain.com, 4 May 2026.
- Brand, J., Israeli, A., and Ngwe, D. (2026). Using GPT for market research. Marketing Unit Working Paper 23-062, Harvard Business School. Revised version; first version 2023.
- Cooper, A. (1999). *The Inmates Are Running the Asylum*. Sams Publishing, Indianapolis, IN.
- Hu, T. and Collier, N. (2024). Quantifying the persona effect in LLM simulations. In *ACL 2024 (Long Papers)*, pages 10289–10307.
- Jacowitz, K. E. and Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, 21(11):1161–1166.
- Klein, R. A., Ratliff, K. A., Vianello, M., et al. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45(3):142–152.
- Korst, J., Puntoni, S., and Toubia, O. (2025). How gen AI is transforming market research. *Harvard Business Review*, 103(3). May–June issue.
- Meister, N., Guestrin, C., and Hashimoto, T. (2025). Benchmarking distributional alignment of large language models. In *NAACL 2025 (Long Papers)*, pages 24–49.
- Newhook, J. (2025). Are AI-generated synthetic users replacing personas? What UX designers need to know. Interaction Design Foundation, <https://www.interaction-design.org/literature/article/ai-vs-researched-personas>.
- Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. arXiv preprint 2411.10109.
- Rosala, M. and Moran, K. (2024). Synthetic users: If, when, and how to use AI-generated “research”. Nielsen Norman Group, 21 June 2024, <https://www.nngroup.com/articles/synthetic-users/>.
- Sarstedt, M., Adler, S. J., Rau, L., and Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, 41(6):1254–1270.
- Sharma, M., Tong, M., Korbak, T., et al. (2024). Towards understanding sycophancy in language models. In *ICLR 2024*.
- Smith, A. R., Windschitl, P. D., and Bruchmann, K. (2013). Knowledge matters: Anchoring effects are moderated by knowledge level. *European Journal of Social Psychology*, 43(1):97–108.
- Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.
- Uebersnickel, F. (2026). Oracles or anchor machines? Knowledge override and dampened anchoring in LLM-simulated consumers. Working paper, preregistered at OSF (ADYPE), June 2026.