

THE QUESTION & THE SETUP

Every month the ECB asks ~19,000 euro-area residents what they expect for inflation, jobs, income, and spending. Can a large language model stand in for a real respondent, and for which kinds of questions does the imitation hold?

We test it person by person. Each of 300 real respondents from the March 2026 wave (the first collected entirely after the February 28 war) becomes a profile card; the model answers all 68 questions as that person, scored against their real answers. The 300 are stratified by country, fixed-seed and frozen, so every model and round answers for the same people — what makes person-by-person pairing possible.

A REAL PROFILE CARD · RESPONDENT p_93726244, WAVE 75

Country: Austria · Age group: 18–34 · Income: 4th quintile · Employed · Male · Household of 4 · Children: 2 · Partner: present · Industry: health & social work · Education: tertiary · Housing: rented — “You are this person. Answer the survey as they would.”

Each person is answered twice. Arm A sees only the profile card. Arm B first reads four lines of context, all pre-war official data; nothing published after the survey closed enters any prompt.

THE FOUR CONTEXT LINES OF ARM B (GERMANY)

“Consumer price inflation over the past 12 months: 2.0% (February 2026 flash estimate). Unemployment rate: 4.0% (January 2026). ECB deposit facility rate: 2.00%. On 28 February 2026, a war began in the Middle East.” A fixed framing sentence follows, leaving it to the persona how closely it follows such news.

How we score. Two levels matter and can disagree. Individual: does each synthetic answer match that person's real one? Distribution: do the 300 answers together have the right shape (option shares, spread, share of zeros)? A model can nail the distribution while every individual is wrong, so we report both.

THE ARMS AT A GLANCE

A: profile card only, the floor.
B: card + the four context lines, the headline configuration.
C: B minus the war sentence, isolating that one line (screening).
D: the same information as B rewritten as a digit-free paragraph (30-person pilot).
The allocation block runs with no context, so the first uncertainty reading is unanchored.

THE SAME LINES REWRITTEN FOR ARM D · GERMANY (EXCERPT)

“Life in Germany these days feels steady but watchful. Prices at the supermarket and for everyday goods keep creeping up at a mild, familiar pace ... Jobs remain fairly easy to come by ... Borrowing money, whether for a car or a home, still carries a noticeable cost, though it feels manageable rather than punishing ...”

Discipline. Four models were screened at 100 respondents against verified training cutoffs (three completed); Claude Sonnet 5 was selected. Comparisons are paired within person, with respondent-level bootstrap uncertainty. Decision rules and predictions were pre-registered before each run, every number traces to an archived file, and misses sit next to hits.

Reading the charts

■ real respondents ■ model, no context ■ model, with statistics ■ model, digit-free



See the survey for yourself — questionnaire, methodology, and the anonymised microdata:
ecb.europa.eu/stats/ecb_surveys/consumer_exp_survey

1 CHOICE QUESTIONS (17 items)

QUESTION c3110

“Looking ahead, do you think your household will be financially better off or worse off in 12 months from now?” Options: better off · about the same · worse off

Accuracy on the choice questions

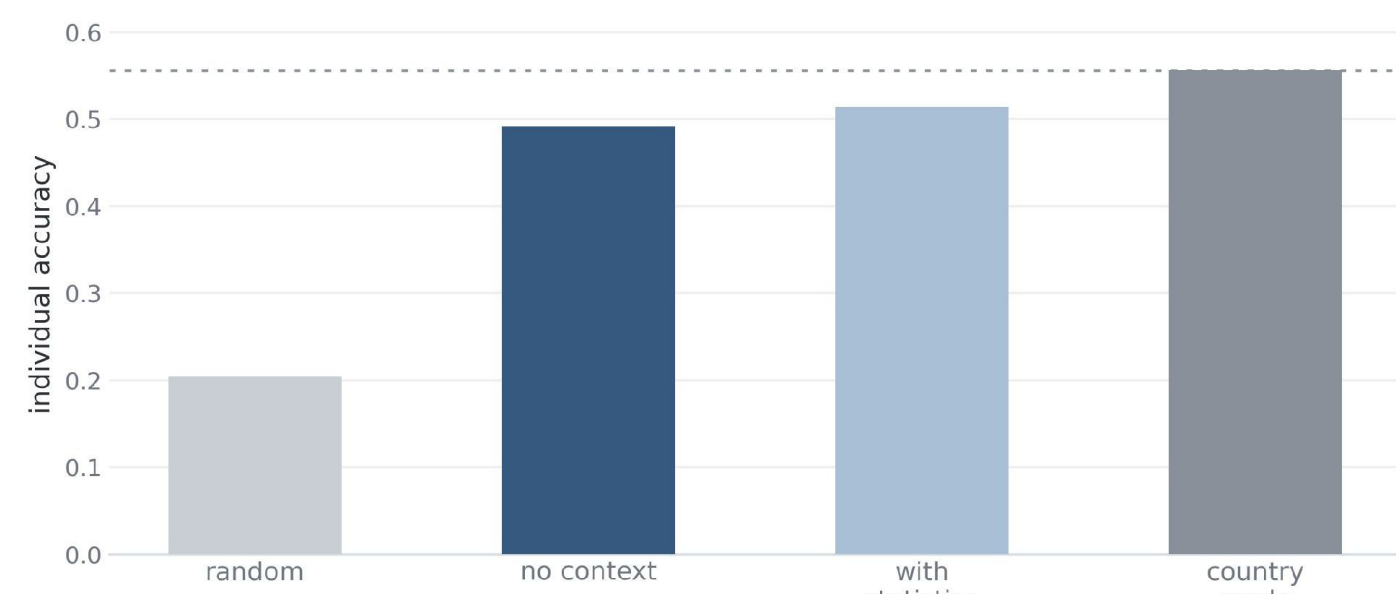


Fig 1 · Individual accuracy, 17 choice questions, 300 respondents.

Scoring is exact match. Random guessing scores 0.204; country mode (giving everyone their country's most common answer) scores 0.556 while knowing nothing about individuals. Context helps for real (+0.022, significant, settled only at the full 300), but the bar stands: both arms sit well below country mode.

Distributions are friendlier: with context, pooled option shares move toward the real ones (Jensen-Shannon divergence 0.451 → 0.400, significant). The synthetic crowd votes like the real crowd, though the individuals in it are no more recognizable.

One context sentence earned its place. An arm identical to B minus the war sentence scored measurably worse on two rulers at once: option shares (+0.027) and the zero-share match (+0.012), both significant. Reading that a war broke out makes the model answer 'no change' and 'hard to say' more often, which is what real people do in uncertain times. The sentence stayed, by a rule fixed before the run.

Diagnosis. To beat one answer per country, the model must tell compatriots apart, and that ability is broken. Gender: men and women get essentially identical answers in both arms, against a documented real gap. Age: the real pattern inverts, even though age is printed on every card. The limit is the model's beliefs about people, not missing information. Under context, between-group differences shrink on all eight demographic dimensions.

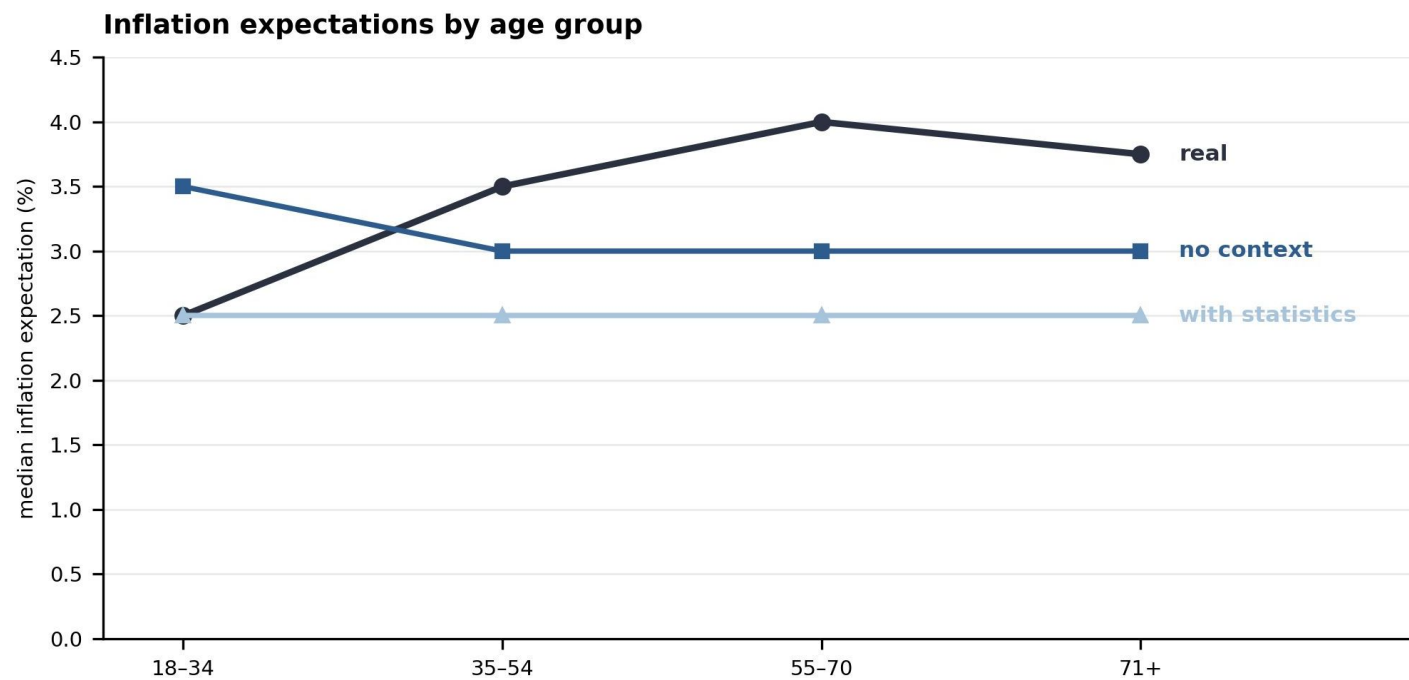


Fig 2 · Median inflation expectation by age band; statistics flatten every group onto the anchored 2.5.

The model has learned how countries answer, and it has no idea how any particular person would.

2 NUMERIC QUESTIONS (12 items)

QUESTION c1120

“How much higher (lower) do you think prices in general will be 12 months from now? Please give your best guess in percentage terms.”

Inflation expectations, real vs model

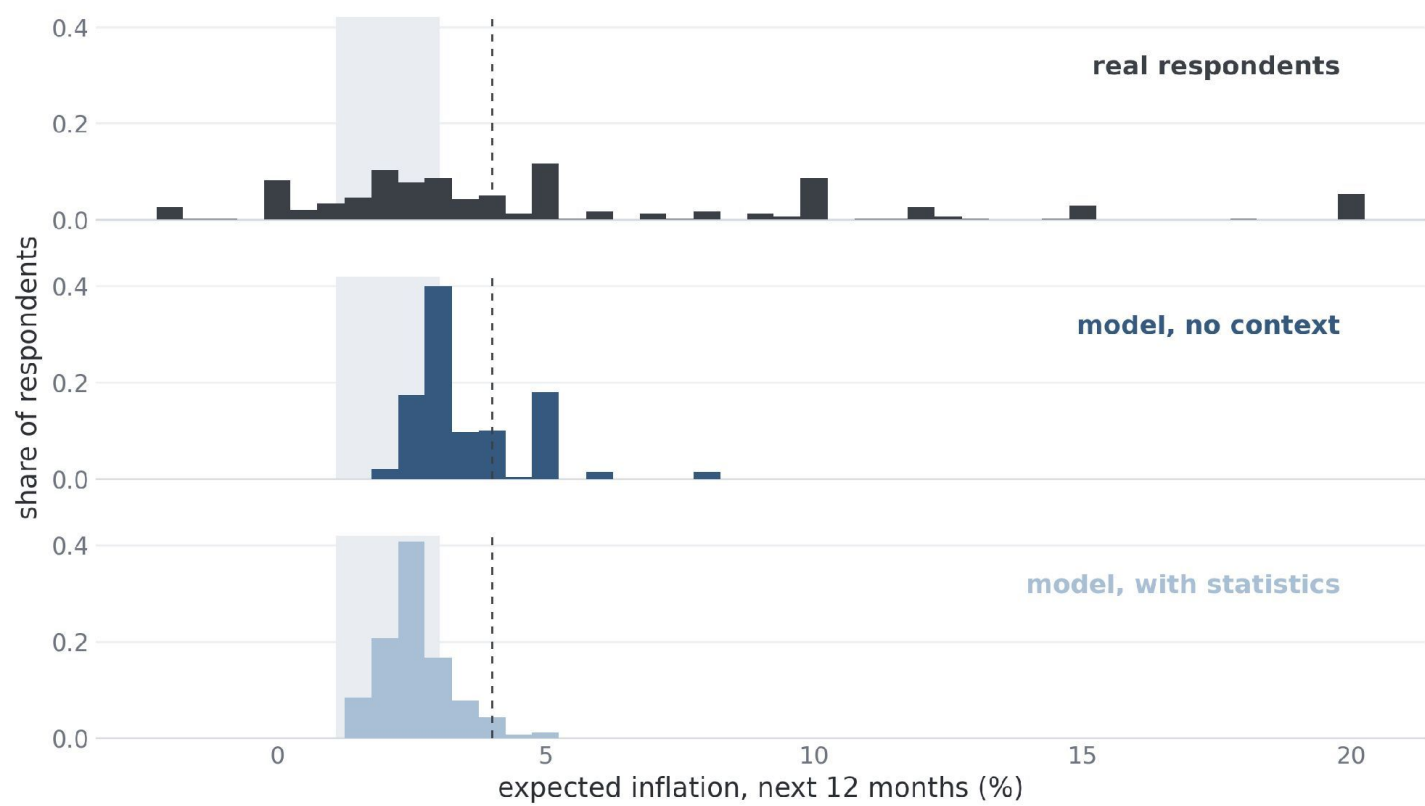


Fig 3 · Twelve-month inflation expectations, 300 respondents per panel. Grey band: supplied pre-war range, 1.1–3.0. Dashed line: full-wave real median, 4.0.

The prompt prints “inflation: 2.0%”, the survey asks what this person expects, and the model copies. The with-statistics median lands at 2.5 while reality had moved to 4.0; no-context reaches 3.0 (closer, still short), and both spreads run far narrower than the real one. A quantile comparison across all twelve numeric questions confirms the pattern holds.

Unemployment-inflation link

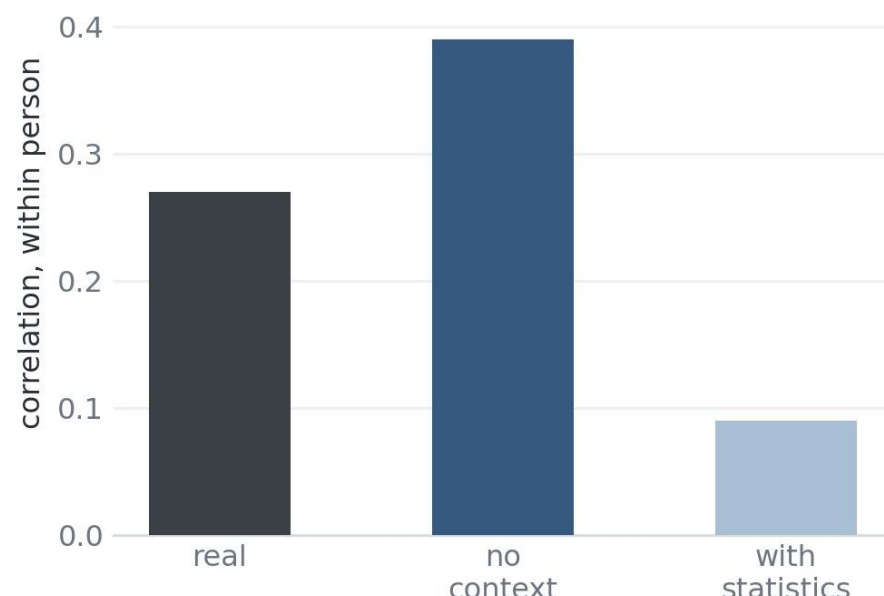


Fig 4 · The severed link: correlation between unemployment and inflation expectations, within person.

The copied number stops being a belief. Real people expecting high inflation also worry more about unemployment (+0.27); without context the model over-couples (+0.39); hand it the unemployment figure and the link collapses to +0.09. The pessimistic tail vanishes: the real tenth percentile of growth expectations is −8.5, the synthetic −0.5. Horizons merge: 49% of personas repeat one number for one- and three-year inflation (real: 26%), and expectations float free of the persona's own perceptions (0.31 vs a real 0.63).

The exception that proves the rule. The supplied January unemployment rates were still roughly true in March, and there the country medians land almost exactly on the real ones (Spain 10.0 vs 10.0, Portugal 5.8 vs 5.8), while inflation medians sit near 2.5 everywhere. What matters is whether the number was still true. Zero-heavy questions add a habit gap: about a quarter of real respondents answer exactly 0; the model, too diligent for habits, almost never does.

Give the model a number and it copies the number instead of answering the question; take the number away and the same information stops doing harm.

3 REMOVING THE DIGITS (ARM D)

If copyable digits drive the anchoring, the same information without digits should undo part of it. Each country's four lines became one no-numbers paragraph; 30 stratified respondents answered it, five predictions archived in advance.

Digit-free context: pilot results

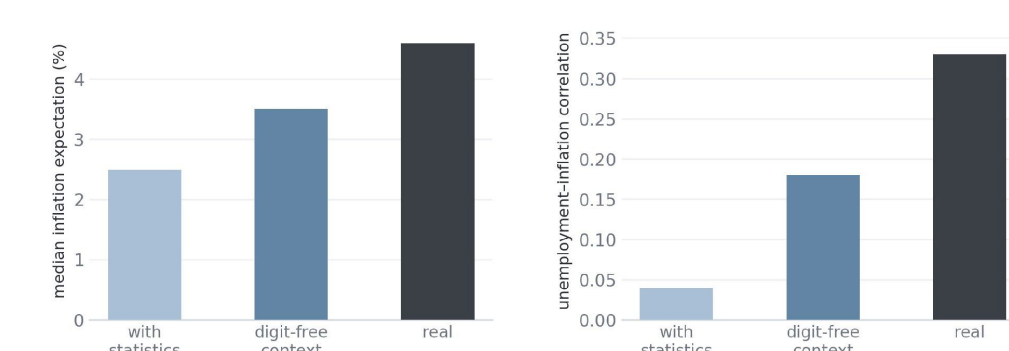


Fig 5 · The pilot, 30 respondents. Bars darkened as they approach the real value.

The anchored median moves half the gap to truth; the severed link recovers half; identical horizon answers fall from 50% to 27%, near the real 23%. Three of four predictions held as written; the fourth missed, though its mechanism moved as predicted. Limits: the digit-free arm never beats no-context, and same-country personas still answer identically.

4 ALLOCATION QUESTIONS (3 items)

QUESTION c1152

“Please distribute 100 points across the following ranges according to how likely you think each one is.” Ten ranges, from an increase of 12% or more down to a decrease of 12% or more.

How sure is each person of their own expectation? Never before posed to a model: same 30 respondents, no context, four predictions pre-registered; 30 of 30 rows valid.

Largest stake per respondent

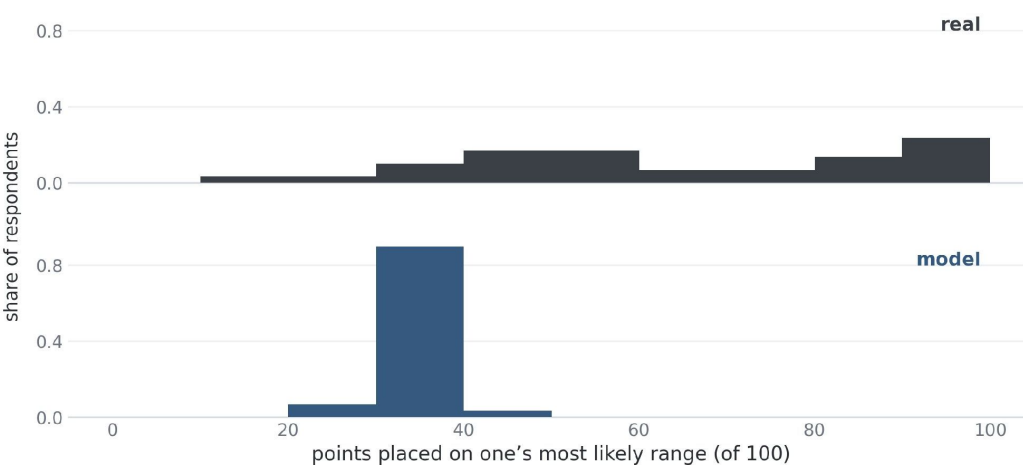


Fig 6 · Largest stake per respondent (inflation). Real medians run 58–62; personas place 22–30, and nine in ten stake the same 30–40.

Average allocation per outcome range

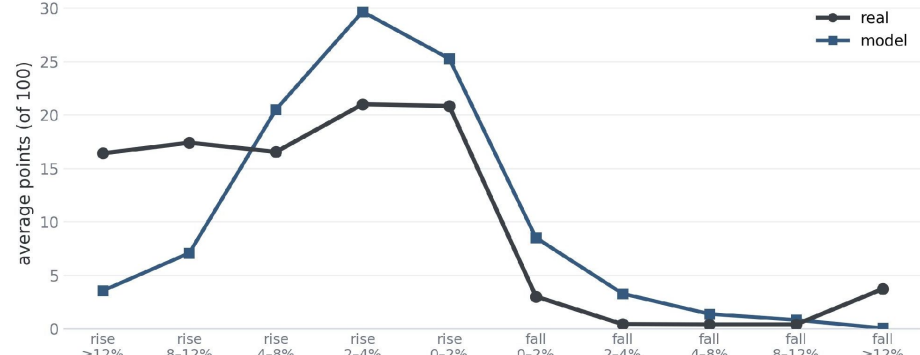


Fig 7 · Average allocation per range. Implied spread: real 1.8 / 1.2 / 1.0; synthetic 3.9 / 5.2 / 5.0. The extreme range gets 3.7 real points and 0 synthetic.

Diagnosis. The real average is broad because each person bets hard on a different range; the model's is a dome because every persona spreads the same one. Allocation centres correlate with the persona's own point forecasts at only 0.49 vs a real 0.83. Both concentration predictions were falsified, reported as written.

The model hedges where humans commit.