



# SURVEY FUTURES

SURVEY DATA COLLECTION  
METHODS COLLABORATION

## Working Paper 15:

LLMs for survey pretesting: How well can large language models identify flaws in survey questions?

Patrick Sturgis<sup>1</sup>, Caroline Roberts<sup>2</sup>, Thomas Robinson<sup>1</sup>

<sup>1</sup>London School of Economics and Political Science;

<sup>2</sup>University of Lausanne

April 2026

[www.surveyfutures.net](http://www.surveyfutures.net)

*Survey Futures* is an Economic and Social Research Council (ESRC)-funded initiative (grant grant ES/X014150/1) aimed at bringing about a step change in survey research to ensure that high quality social survey research can continue in the UK. The initiative brings together social survey researchers, methodologists, commissioners and other stakeholders from across academia, government, private and not-for-profit sectors. Activities include an extensive programme of research, a training and capacity-building (TCB) stream, and dissemination and promotion of good practice. The research programme aims to assess the quality implications of the most important design choices relevant to future UK surveys, with a focus on inclusivity and representativeness, while the TCB stream aims to provide understanding of capacity and skills needs in the survey sector (both interviewers and research professionals), to identify promising ways to improve both, and to take steps towards making those improvements. *Survey Futures* is directed by Professor Peter Lynn, University of Essex, and is a collaboration of twelve organisations, benefitting from additional support from the Office for National Statistics and the ESRC National Centre for Research Methods. Further information can be found at [www.surveyfutures.net](http://www.surveyfutures.net).

This paper is a product of *Survey Futures* Research Strand 9, “Generative AI for questionnaire design,” led by Professor Patrick Sturgis.

Prior to citing this paper, please check whether a final version has been published in a journal. If so, please cite that version. In the meanwhile, the suggested form of citation for this working paper is:

Sturgis P, Roberts C & Robinson T (2026) ‘LLMs for survey pretesting: How well can large language models identify flaws in survey questions?’, *Survey Futures Working Paper* no. 15. Colchester, UK: University of Essex. Available at <https://surveyfutures.net/working-papers/>.

# LLMs for Survey Pretesting: How Well Can Large Language Models Identify Flaws in Survey Questions?

Patrick Sturgis<sup>1</sup>

Caroline Roberts<sup>2,1</sup>

Thomas Robinson<sup>1</sup>

<sup>1</sup>Department of Methodology, London School of Economics and Political Science

<sup>2</sup>Department of Sociology, University of Lausanne

March 2026

## Abstract

This paper investigates whether large language models (LLMs) can approximate canonical human question review procedures. We compare three prompt variants for simulated cognitive testing against a simpler expert review approach using a set of 20 questions containing deliberately embedded problems, with 7 items from the European Social Survey used as a control set. The best-performing configuration, a guided cognitive testing variant, detects 75% of known problems with a small false positive rate. As with their human counterparts, simulated cognitive testing and expert review detect different types of problem, suggesting the two approaches are complementary and could fruitfully be used in tandem. Detection and false positive rates are sensitive to prompt design and model choice: structured prompts that work well on open-weight models perform poorly on the proprietary model tested, and open-weight models achieve higher detection rates but much poorer discrimination between flawed and well-designed questions. A full evaluation of a 20-item questionnaire completes in under an hour at a cost of a few dollars, orders of magnitude faster and cheaper than human pretesting. These findings suggest that LLM-based question testing has the potential to provide a useful complement to human pretesting, particularly for early-stage screening of large item sets and in situations where resource constraints mean human-testing is not possible.

## 1 Introduction

Cognitive interviewing has become the pre-eminent tool for evaluating draft survey questions before fielding a survey. Its distinctive value lies in anchoring diagnosis in the respondent's perspective: participants think aloud while answering a question, or respond to targeted probes about their internal thought processes, and a trained analyst identifies difficulties across interviews. Drawing on Tourangeau's four-stage cognitive model of survey response (Tourangeau, Rips & Rasinski, 2000), the method is designed to reveal how respondents actually navigate each cognitive stage, surfacing problems that may not be apparent from the question text alone (Willis 2005; Beatty and Willis 2007). Over the past several decades, it has moved from a relatively specialised technique to a routine component of questionnaire development, with its use growing substantially in recent decades across academic, governmental, and applied settings (Caporaso and Presser 2024).

Yet the very success of cognitive interviewing has also drawn attention to its limitations. Although it offers valuable insight into individual- and subgroup-level variation in question validity (Collins 2015), it is also labour-intensive, slow, and expensive in practical terms. Conducting interviews, training interviewers, analysing recordings or notes, and iteratively revising items all require substantial resources, which means that only a limited number of questions, respondents, and design variants can usually be examined within real project constraints (Beatty and Willis 2007; Caporaso and Presser 2024). In practice, many studies rely on small convenience samples and a relatively small number of rounds, often because budgets and timelines do not permit more expansive testing (Beatty and Willis 2007). This creates a tension between the aspiration to identify as many important problems as possible and the reality that questionnaire development often proceeds under severe operational pressure.

That tension is sharpened by evidence on sample size. Cognitive interviewing has often been justified on the basis that a small number of interviews can reveal the most serious problems, and in practice this is what tends to happen: Caporaso and Presser (2024) report a median minimum of around five respondents

and a median maximum of around twenty per round among US academic organisations. Yet consequential problems may remain undiscovered unless sample sizes become much larger than these conventional rules of thumb imply, because the probability of observing a problem depends on both its prevalence and the interview’s sensitivity to it (Blair and Conrad 2011). For long or complex questionnaires, exhaustive human cognitive testing of every item is often infeasible.

Cognitive interviewing is also not equally effective at identifying all types of question problems. Because it relies on respondent experience, it is well suited to diagnosing difficulties that arise when someone actually attempts to answer a question, such as confusion over ambiguous wording or inability to map a response onto the available options. It is less well suited, however, to identifying problems that do not create experienced difficulty, such as subtle presuppositions or leading framing that respondents may accept without noticing (Presser and Blair 1994). An alternative question testing procedure, systematic expert review, relies on expert judgement rather than respondent perspective (Presser et al. 2004): experienced survey methodologists evaluate questions against known problem taxonomies and can catch design flaws that are visible from a methodological standpoint even when respondents do not experience difficulty. Comparative work confirms that the two approaches tend to identify different kinds of problems, and that neither is sufficient on its own (Presser and Blair 1994). Expert review is less widely used than cognitive interviewing as a formal evaluation procedure, no doubt in part because of the time pressures on the small number of professionals with specialist question design expertise. Both approaches are constrained by the availability of skilled practitioners, and both are too resource-intensive to apply exhaustively across large instruments.

These long-standing constraints on effective pre-testing have become newly salient with the introduction of large language models (LLMs), and in particular with their use in simulating human survey respondents, what is referred to as ‘silicon sampling’ (Argyle et al. 2023). While early studies showed that conditioned models could reproduce aggregate attitudinal patterns and classic experimental effects (Argyle et al. 2023; Aher, Arriaga and Kalai 2023; Dillion et al. 2023; Sarstedt et al. 2024), subsequent evaluations have identified persistent limitations: sensitivity to superficial ordering and label effects (Dominguez-Olmedo, Hardt and Mendler-Dünner 2024), systematic biases toward Western, educated, liberal subgroups (Atari et al. 2023; Santurkar et al. 2023), poor recovery of subgroup patterns in non-English settings (Qi, Lyu and Luo 2024; Heyde, Haensch and Wenz 2025), and a tendency for demographic prompting to flatten within-group diversity (Wang, Morgenstern and Dickerson 2025). These representational shortcomings currently make silicon sampling unsuitable for population inference. They are less consequential, however, in diagnostic applications where the goal is not to estimate population quantities but to identify question design flaws (Sarstedt et al. 2024). It is this diagnostic use case that the present paper addresses.

We present a systematic evaluation of LLM-based simulated cognitive testing and expert review. To simulate cognitive testing, we develop a three-stage pipeline in which synthetic respondents defined by realistic sociodemographic backstories work through questions in a think-aloud format in the first person, with an ‘analyst’ stage that codes identified problems and a synthesis stage that consolidates findings. We test three prompt variants that differ in how much structure and guidance the simulated respondent receives, addressing the open question of whether the principles that govern effective human cognitive interviewing transfer to LLM-based evaluation. As a comparison, we also evaluate a simpler approach, direct LLM-based expert review in which reviewers evaluate questions against known problem taxonomies, with three prompt variants of its own. We assess performance using a test set of 27 questions, 20 of which contain deliberately embedded problems spanning eight common problem types and seven of which are well-tested core items from the European Social Survey that serve as controls. We further test whether findings generalise across models by replicating all configurations on two open-weight alternatives alongside the proprietary GPT-4o. This design allows us to evaluate not only how well LLMs can simulate human pre-testing, but also how effectively they discriminate between flawed and well-designed items, what types of problems each approach is best suited to detecting, and how sensitive performance is to prompt design and model choice.

Our principal contribution is a controlled evaluation of LLM-based question testing in which detection rates and false positive rates can be computed against a known answer key. We find that the best-performing configuration detects three quarters of known problems with a small false-positive rate, and that simulated cognitive testing and expert review are complementary: the former is more effective at identifying response-mapping failures, while the latter is stronger on presupposition and leading framing. Performance is sensitive to both prompt design and model choice, with structured prompts that work well on one model performing poorly on another, and open-weight models achieving higher detection rates but

substantially poorer discrimination between flawed and well-designed items. These results suggest that LLM-based question evaluation has the potential to serve as a useful complement to human pretesting methods, particularly for early-stage screening of large item pools, or in situations where human cognitive testing and expert review are beyond project resource constraints.

## 2 Methodology

### 2.1 Simulated cognitive testing

To simulate cognitive testing, we develop an LLM-based pipeline for question evaluation, consisting of simulated respondents, analyst coding and synthesis. We implement this approach in a prototype system (what we refer to as 'CogBot'), but the architecture is model-agnostic and not tied to any specific interface.

The pipeline proceeds in three stages (Figure 1). First, a random sample of  $n$  synthetic respondents is drawn from a pool of backstories derived from individual respondent records in the 2018 ESS UK dataset, a probability sample of the UK general population aged 15 and over. The pool contains 2,204 respondent backstories, each built from demographic variables in the ESS record: age, gender, education, employment status, region, household composition, and income band. All analyses reported here use these short demographic-only backstories (approximately 580 characters). Each backstory is constructed deterministically by inserting the ESS record's demographic values into a natural-language template: the backstories are not themselves generated by an LLM, so there is no feedback loop from the model used for cognitive testing back into the respondent profile.

Each sampled respondent independently works through the survey question in a single-prompt think-aloud format, verbalising their thought process and noting any difficulties encountered; the three prompt variants, described below, differ in how much cognitive structure and problem-type guidance the respondent receives (Figure 1). Second, an analyst LLM reads each respondent's transcript and codes any problems that are clearly evidenced in the text. The analyst prompt (full text in Appendix C) instructs the model to identify only problems that are supported by specific transcript evidence, to return an empty list where the respondent navigated the question without difficulty, and to record for each problem a plain-language description, the supporting excerpt from the transcript, and a severity rating on a 1 to 10 scale (where 1 indicates negligible distortion to the respondent's answer and 10 indicates the answer is rendered meaningless). Third, a synthesis step consolidates findings across the  $n$  respondents, merging observations that describe the same underlying problem even when they are expressed differently. For the main results reported here we set  $n = 30$ , on the grounds that preliminary testing indicated the number of identified problems tails off above this sample size. Because each run draws a different random sample of backstories from the pool, repeated runs produce somewhat different sets of identified problems. In addition to the think-aloud transcript, each simulated respondent records the answer they have selected and rates their confidence in that answer on a 1 to 5 scale, providing a potential supplementary signal of experienced difficulty.

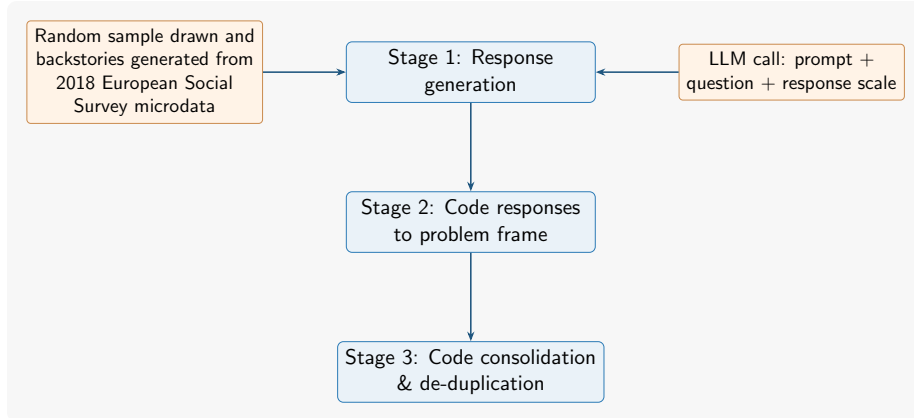


Figure 1: Simulated cognitive testing pipeline. Stage 1 draws a random sample of backstories from survey microdata and generates think-aloud transcripts. Stage 2 codes each transcript for problems. Stage 3 consolidates findings across respondents.

This architecture mirrors the logic of human cognitive interviewing, where recruited participants verbalise their thought process while answering a draft question and a trained analyst identifies patterns across interviews. An important simplification is that real cognitive interviews are dynamic: the interviewer follows up on the respondent’s answer with targeted probes, pursuing lines of difficulty as they emerge. In practice, however, many follow-up probes are planned in advance as part of the interview protocol and tailored to the question being tested, and much of what live probing does is keep a vague or inattentive respondent focused on the question, work that is at least partially handled by the think-aloud prompt itself in our pipeline. Fully mimicking dynamic probing would require a multi-turn exchange between an LLM interviewer and an LLM respondent, which is beyond the scope of this initial investigation. The pipeline therefore relies on a single prompt to the respondent, raising the question of how much structure this prompt should contain.

We test this systematically with three prompt variants, holding all other content constant. In the first (CI-Open), the respondent simply thinks aloud with no framework or problem types mentioned. In the second (CI-Structured), the respondent works through four stages based on the Tourangeau–Cannell cognitive model of survey response (Cannell, Miller and Oksenberg 1981) (comprehension, retrieval, judgement, response mapping) but no problem types are named. In the third (CI-List), the same four-stage structure includes explicit lists of common problem types at each stage (e.g., double-barrelled wording at comprehension, recall difficulty at retrieval, poor fit at response mapping). All three share the same analyst and synthesis prompts; full prompt texts are provided in Appendix C.

## 2.2 Expert review

For expert review, three independent LLM reviewers evaluate the question directly, without simulating respondent experience, and a synthesis step consolidates their findings, merging duplicates and forming compound codes from overlapping problems. The three reviewers are produced by calling the same model three times with a distinct reviewer identifier on each call, rather than by using three different models. The pipeline also elicits severity ratings from each reviewer on a 1–10 scale (where 1 indicates negligible distortion and 10 indicates the answer is rendered meaningless); the final severity for each problem is taken as the mean across the three reviewers. This architecture is analogous to expert appraisal, where experienced survey methodologists review questions against their knowledge of common design problems. The expert review pipeline uses three fixed reviewers with no backstory sampling; its results are from single runs. In the first variant (ER-Checklist), reviewers evaluate the question against a seven-type checklist (comprehension, double-barrelled, retrieval, response mapping, social desirability, presupposition, tautology), marking each as present or absent, and the synthesis merges by type label. In the second (ER-Open), reviewers identify problems in their own terms with no predefined types and are told that problems should be specific to the question rather than generic observations. In the third (ER-Restrained), the same label-free approach is used but with strong restraint language instructing

reviewers to report only problems genuinely likely to affect respondents and to return an empty list if the question is well designed. Full prompt texts are provided in Appendix C.

### 2.3 Test set and evaluation design

The test set comprises 27 survey questions. Following Blair and Conrad (2011), twenty questions are deliberately flawed items designed by the first author drawing on established taxonomies of question-design defects (Willis 2005; Presser and Blair 1994), each containing one to three commonly encountered problems in draft questions: double-barrelled questions, presupposition, vague or ambiguous wording, response-mapping failures, retrieval and recall difficulty, leading or loaded phrasing, sensitivity, and complex syntax. The embedded problems were determined by construction and constitute the answer key against which pipeline output is evaluated. The pipeline is not constrained to identifying these categories, however, and is free to find problems of any kind. The remaining seven questions are well-tested core items from the European Social Survey (ESS) that have been fielded across more than thirty countries and many survey rounds. Having been through the ESS’s rigorous multi-stage development process and repeated real-world use, these items serve as controls: a well-functioning evaluation tool should report few, or no problems on them. The full test set is provided in Appendix B.

This design allows us to assess two distinct aspects of performance. Detection rate measures the proportion of the 48 known problems that the pipeline successfully identifies. The false positive rate measures how often the pipeline reports problems that are not in the answer key: on flawed questions, these are any identified problems beyond the deliberately embedded ones; on ESS questions, they are any problems reported at all. Both sets of false positive rates are likely to be overestimates, since the flawed questions may contain genuine problems beyond those deliberately embedded and the ESS items, while developed to the highest standards, cannot be assumed to be entirely problem-free. We return to this point in the Results, where independent human adjudication is used to assess the extent of this overestimation.

The use of deliberately constructed items with known embedded problems is not intended to approximate the full complexity of real-world questionnaire development. Rather, it provides a necessary identification strategy, without a known answer key, it is not possible to distinguish true positives from false positives or to compute detection rates in a principled way. The present design should therefore be understood as a benchmark test of diagnostic capability under controlled conditions, analogous to test batteries used in psychometric validation, rather than a direct simulation of applied questionnaire development.

To evaluate each configuration, we compare its identified problems against the answer key. Each pipeline-identified problem is paired with each answer-key problem for the same question, and an LLM judge classifies each pair as an *exact match* (same problem, same framing), an *approximate match* (same underlying issue, different emphasis or scope), or *no match* (different problems). Each pipeline problem is then matched to at most one answer-key problem and vice versa, with exact matches prioritised over approximate matches. Unmatched pipeline problems are classified as false positives; unmatched answer-key problems are misses. To avoid dependence on a single judge, we ran the matching independently with two frontier models from different providers: Claude Opus 4.6 (Anthropic) and GPT-5.4 (OpenAI), both at temperature 0 with identical prompts. The two judges agreed on 94.4% of 764 problem pairs. The 43 discordant pairs were adjudicated by the first author. The reliability of the matching procedure is analysed further in the Results.

### 2.4 Models

All six configurations are first tested using GPT-4o (OpenAI; temperature 0.7), a proprietary frontier model accessed via the OpenAI API. GPT-4o was chosen as the primary model because its cost per token at the time of testing made the full benchmark, which involves many thousands of API calls across replicates, configurations, and questions, feasible within a modest compute budget. The same configurations are then replicated on two open-weight models to test whether the pipeline’s performance generalises beyond a single proprietary system.

*Llama-3.3-70B-Instruct* (Meta) is an open-weight model with 70 billion parameters, served locally via SGLang.<sup>1</sup> Llama-3.3 is instruction-tuned but has undergone less extensive alignment than GPT-4o.

---

<sup>1</sup>Llama-3.3-70B is run with AWQ 4-bit quantisation (approximately 36GB VRAM) on a single NVIDIA RTX 6000 Ada 49GB GPU.

*Qwen3-32B* (Alibaba) is an open-weight model with 32 billion parameters, served locally via SGLang on the same hardware.<sup>2</sup>

For each model, all pipeline stages (respondent simulation, analyst coding, synthesis, and expert review) use the same model throughout. This tests each model’s end-to-end capability rather than mixing models across stages. Prompts, test set, backstories, temperature, and all other parameters are held constant across models; only the model itself varies.

### 3 Results

To give a sense of what the pipeline produces, we present output from both testing approaches applied to question F5: “Overall, how satisfied are you with your terms of employment?” (scale: Completely satisfied / Very satisfied / Fairly satisfied / Not very satisfied / Not at all satisfied). This question contains three deliberately embedded problems: “terms of employment” is vague, the response scale is asymmetric (three positive options, two negative), and the question presupposes current employment.

Table 1 shows output from both approaches. For CI-Open, two of the 30 simulated respondents are shown; neither received any cognitive framework or problem-type cues. For ER-Checklist, one of the three independent reviewers is shown alongside the consolidated output.

Table 1: Illustrative pipeline output for question F5, “Overall, how satisfied are you with your terms of employment?” Top panel: CI-Open (respondent transcript and analyst coding). Bottom panel: ER-Checklist (expert reviewer assessment and consolidated problem code).

<b>Simulated cognitive testing (CI-Open)</b>	
Respondent transcript	Analyst coding
<i>Respondent 2 (retired, former shopkeeper):</i> “As I read the question about my satisfaction with my terms of employment, I first think about my situation. Being retired and a former shopkeeper, I’m not currently employed in the traditional sense, so the question feels a bit out of place for me. However, I can think back to when I was actively running my shop. [...] Since I’m retired, none of these apply directly to me anymore.”	The respondent is retired and self-employed, making the question about “terms of employment” feel irrelevant. They had to reflect on past experiences and adapt the question to fit their unique situation, which could lead to inaccuracies. (Severity: 7/10)
<i>Respondent 8 (kitchen helper, female, 54):</i> “As I read the question [...] I start to think about my current job as a kitchen helper. I consider the hours I work, the pay I receive, and the conditions I work under. These are what I think about when I hear ‘terms of employment.’ [...] The phrasing ‘terms of employment’ could be a bit vague, but I interpret it as encompassing these aspects: pay, hours, and work conditions.”	The respondent found the term “terms of employment” to be vague, which might have led to some uncertainty about what aspects to consider when answering the question. (Severity: 5/10)
<b>Expert review (ER-Checklist)</b>	
Expert reviewer assessment	Consolidated problem code
<i>Comprehension</i> (severity 6): “The phrase ‘terms of employment’ may be interpreted differently by respondents, potentially including salary, benefits, work hours, or job security.”	(1) “Terms of employment” is ambiguous (mean severity 5.7; flagged by all 3 reviewers).
<i>Presupposition</i> (severity 5): “The question assumes that respondents have defined and clear terms of employment, which may not be true for all, such as temporary or freelance workers.”	(2) The question assumes defined employment terms (mean severity 5.0; flagged by 1 reviewer).

These examples are illustrative rather than representative, selected to show cases where the pipeline

<sup>2</sup>Qwen3-32B is run with FP8 quantisation. Qwen3 generates reasoning blocks by default; a wrapper suppresses these by appending a no-think instruction and stripping reasoning tags from responses.

correctly identifies embedded problems. Both approaches successfully detected known problems in this question but the contrast between them is instructive. The simulated cognitive test surfaced the presupposition through the lived experience of a retired respondent encountering a question that did not apply to him, and the vagueness through a respondent who had to decide for herself what “terms of employment” meant. The expert review identified the same issues as logical properties of the question structure. This illustrates the importance of the backstory pool for the simulated cognitive testing approach, which depends on generating respondents for whom particular problems are salient.

Next we turn to the overall results across all test questions. Table 2 presents the detection rate, false positive rate, LLM-assigned severity, and respondent confidence for all six configurations on GPT-4o, with separate rows for flawed and ESS questions within each testing method.

Table 2: Overall results across six testing methods on GPT-4o. Detection = proportion of 48 known problems matched. FP/Q = mean false positives per question. FP/Q (sev  $\geq 5$ ) = mean false positives per question after excluding problems with severity below 5. Confidence = mean respondent confidence (1–5); available for cognitive testing methods only.

Configuration	Prompt architecture	Set	Detection	FP/Q	FP/Q ( $\geq 5$ )	Confidence
CI-List	Simulated respondent works through the four cognitive stages of the Tourangeau–Cannell model with an explicit list of common problem types provided at each stage.	Flawed	75.0%	1.8	1.50	4.08
		ESS	—	1.6	1.00	4.44
ER-Checklist	Three expert reviewers evaluate the question against a seven-type problem checklist (comprehension, double-barrelled, retrieval, response mapping, social desirability, presupposition, tautology).	Flawed	72.9%	1.4	1.05	—
		ESS	—	2.3	0.57	—
ER-Restrained	Three expert reviewers identify problems in their own terms, with strong restraint language instructing them to report only problems genuinely likely to affect respondents and to return an empty list if the question is well designed.	Flawed	68.8%	1.4	1.40	—
		ESS	—	3.1	3.14	—
CI-Open	Simulated respondent thinks aloud while working through the question, with no cognitive framework or problem types mentioned in the prompt.	Flawed	60.4%	1.4	1.00	4.21
		ESS	—	1.4	1.00	4.41
ER-Open	Three expert reviewers identify problems in their own terms, with no predefined problem types and no additional restraint language.	Flawed	47.9%	0.8	0.55	—
		ESS	—	0.3	0.29	—
CI-Structured	Simulated respondent works through the four cognitive stages of the Tourangeau–Cannell model, but no problem types are named at any stage.	Flawed	31.2%	0.6	0.50	4.28
		ESS	—	0.7	0.57	4.42

CI-List achieves the highest detection rate, 75% with a moderate false positive rate (1.6 per ESS question).

ER-Checklist is close behind at 72.9% detection but generates more false positives on ESS questions (2.3 per question). ER-Restrained achieves 68.8% detection and has the highest ESS false positive rate of any configuration (3.1 per question). Whether that is a limitation or a feature depends on the use case: high recall with low specificity may be acceptable for early-stage screening where any flagged question will be reviewed by a human, but it offers little discrimination between questions that need revision and those that do not.

CI-Open achieves 60.4% detection with moderate false positive rates (1.4 per ESS question). ER-Open has far fewer false positives than any other configuration (0.3 per ESS question) but at the cost of missing more than half the known problems. CI-Structured is the weakest configuration overall at 31.2%; the four-stage framework appears to constrain GPT-4o’s simulated respondents so that they navigate even flawed questions without reporting difficulties.

Each LLM evaluator assigns a severity rating on a 1–10 scale to each identified problem (where 1 indicates negligible distortion and 10 indicates the answer is rendered meaningless). In the simulated cognitive testing pipeline, the analyst codes severity for each respondent’s transcript and the synthesis averages across respondents; in expert review, each of the three reviewers rates severity and the synthesis averages across reviewers. The final column of Table 2 shows the effect of retaining only problems rated 5 or above. Excluding lower-severity problems reduces false positive rates across most configurations while retaining the majority of true positives. The effect is strongest for ER-Checklist, which drops from 2.3 to 0.57 false positives per ESS question, a 75% reduction. CI-List and CI-Open both fall from their baselines to 1.00. ER-Restrained is the exception: almost all of its false positives carry severity ratings of 5 or above, so thresholding has virtually no effect (3.14 versus 3.1 at baseline). Above a threshold of 5, detection falls sharply with diminishing gains in precision. Pooling across all six configurations, the mean severity assigned to problems on flawed questions is 5.87 (SD = 1.22,  $n = 318$ ), compared to 5.23 (SD = 1.15,  $n = 66$ ) on ESS controls. The difference of 0.64 points is statistically significant ( $t = 4.04$ ,  $p = .0001$ ) but modest on a 10-point scale, consistent with the compressed distribution. Severity ratings thus appear to carry diagnostic signal that can serve as a filter for prioritising findings, even though the absolute calibration is compressed into a narrow range.

The false positive rates reported above treat the answer key as exhaustive and the ESS questions to be without problems entirely, neither of which is likely to hold in practice. We therefore tested how much of the reported false-positive rate reflects genuinely spurious findings as opposed to real problems missing from the answer key. All 79 unmatched problems from the two highest-performing configurations (CI-List and ER-Checklist) were submitted for independent human adjudication; these two configurations were chosen because expert adjudication is expensive (two specialists  $\times$  79 problems) and prioritising the strongest configurations is where the result carries most weight. Two questionnaire design specialists each rated every problem on a 1–5 severity scale (1 = not a problem at all, 5 = very severe problem) without seeing the answer key or matching verdicts. Ratings were collapsed into a binary classification (4–5 = genuine design problem, 1–3 = not genuine), and a problem was reclassified as genuine only where both raters independently rated it 4 or 5. Inter-rater agreement on the binary classification was 78.5% (Cohen’s  $\kappa = 0.57$ ; Spearman  $r_s = 0.71$  on the original 1–5 scale).

This additional adjudication step substantially changes how CI-List’s false-positive rate should be interpreted. For CI-List, 19 of 29 false positives on flawed questions (65.5%) were reclassified as genuine problems, reducing the false positive rate from 1.8 to 0.5 per flawed question. In other words, nearly two thirds of what the pipeline flagged beyond the answer key turned out to be real design problems that the authors of the test set had missed. For ER-Checklist, 8 of 24 (33.3%) were reclassified, reducing the rate from 1.4 to 0.8 per flawed question. On ESS questions, adjudication had minimal effect, 0 of 9 CI-List problems and 2 of 17 ER-Checklist problems were reclassified, confirming that false positives on the well-tested core items are genuinely spurious. The practical implication is that CI-List’s effective precision on flawed items is considerably higher than the headline numbers in Table 2 suggest, and that part of the apparent false-positive rate is really the pipeline surfacing defects over and above those deliberately embedded. Although adjudication was conducted only for these two configurations, the same dynamic is likely to apply at least partially to the others, meaning the false positive rates reported in Table 2 should be understood as upper bounds on genuinely spurious output rather than as direct measures of it.

The simulated cognitive testing pipeline introduces stochastic variation between runs from two sources: the random sampling of backstories (different runs draw different respondents) and the inherent non-determinism of LLM text generation (the same prompt can produce different outputs across calls at temperature > 0). To quantify this combined variation, we ran each cognitive testing configuration three

times with different random seeds, holding all other parameters constant. The expert review pipeline uses three fixed reviewers with no backstory sampling; its results are from single runs. Table 3 shows the replicate results.

Table 3: Per-replicate results for the three cognitive testing configurations on GPT-4o (short backstories,  $n = 30$ ).

Configuration	Run	Detection	FP/flawed Q	FP/ESS Q
CI-List	Rep 1	79.2%	1.4	1.4
	Rep 2	66.7%	1.1	1.8
	Rep 3	64.6%	1.4	1.6
	<i>Mean (SD)</i>	<i>70.1% (7.9)</i>	<i>1.30</i>	<i>1.58</i>
	<i>Range</i>	<i>64.6–79.2</i>	<i>1.1–1.4</i>	<i>1.4–1.8</i>
CI-Open	Rep 1	58.3%	1.2	1.0
	Rep 2	58.3%	1.2	1.2
	Rep 3	60.4%	1.1	1.1
	<i>Mean (SD)</i>	<i>59.0% (1.2)</i>	<i>1.18</i>	<i>1.12</i>
	<i>Range</i>	<i>58.3–60.4</i>	<i>1.1–1.2</i>	<i>1.0–1.2</i>
CI-Structured	Rep 1	29.2%	0.4	0.2
	Rep 2	33.3%	0.5	0.9
	Rep 3	39.6%	0.5	0.4
	<i>Mean (SD)</i>	<i>34.0% (5.2)</i>	<i>0.45</i>	<i>0.50</i>
	<i>Range</i>	<i>29.2–39.6</i>	<i>0.4–0.5</i>	<i>0.2–0.9</i>

There is notable between-run variation, but the general pattern of detection and false positive rates is broadly consistent across replicates. CI-List achieves the highest mean detection (70.1%) but is also the most variable across runs (standard deviation of 7.9 percentage points, henceforth pp), with individual replicates ranging from 64.6% to 79.2%. CI-Open is considerably more stable (SD = 1.2 pp) but at a lower detection level. CI-Structured performs poorly across all three runs, confirming that its weak detection is not a single-run anomaly. False positive rates are more stable than detection rates across all three configurations.

## Detection rates by problem type

Table 4 shows that detection varies substantially, and different configurations have distinctive profiles. Double-barrelled problems are detected by all configurations except CI-Structured (67%), presumably because asking about two things at once is salient and well known. Presupposition is the hardest type overall and the one where configurations diverge most: ER-Checklist leads at 89%, likely because its checklist explicitly names presupposition as a category, while ER-Open manages only 11% without this cue.

The most substantively interesting contrast concerns response-mapping problems. CI-List detects 71% of these problems and CI-Open 57%, compared to 43% for ER-Checklist and 0% for ER-Open. Simulated respondents who actually attempt to map their answer onto the response scale are more likely to notice that the options do not fit than an expert reviewer analysing the question in the abstract. This mirrors the complementarity between human cognitive interviewing and expert appraisal, each approach surfaces problems that the other is more likely to miss.

Leading and loaded framing shows a large gap between configurations with and without explicit type cues. CI-Open detects only 40%, while CI-List reaches 100% and ER-Checklist 80%. Certain problem types, particularly presupposition and leading framing, are rarely identified without explicit prompting or taxonomic guidance.

There is a strong positive association between detection rate and mean severity across problem types (Spearman  $\rho = 0.69$ ), such that problem types that the pipeline detects more frequently also receive higher severity ratings, with double-barrelled problems at one extreme and response-mapping at the other. This could indicate that genuinely severe problems are more salient and therefore easier to detect. Alternatively, problems detected by more respondents accumulate stronger evidence, which may

inflate the synthesised severity rating. A different research design would be needed to disentangle these possibilities.

The confidence ratings elicited from simulated respondents in the cognitive testing configurations provide another potential diagnostic signal. Across all three CI configurations, simulated respondents report consistently lower confidence on flawed questions than on ESS controls (Table 2): CI-List 4.08 versus 4.44, CI-Open 4.21 versus 4.41, CI-Structured 4.28 versus 4.42. The direction of the difference is consistent, though the gap is modest in absolute terms. At the problem-type level (Table 4), the pattern is more differentiated. Problem types that create experienced difficulty for the respondent produce lower confidence. Leading/loaded framing (3.80) and presupposition (3.91) both fall well below the ESS baseline. At the question level, F14 (childcare satisfaction, which presupposes need for childcare) averages 2.10, and F4 (immigration attitudes, with leading preamble and complex syntax) averages 3.37. By contrast, problem types that affect data quality without the respondent noticing produce near-ceiling confidence: vague/ambiguous wording (4.37), retrieval/recall (4.32), and response-mapping (4.30) all fall close to the ESS control mean. Confidence ratings thus better identify problems that create experienced difficulty for the respondent, as would be expected.

Table 4: Detection rate (%) and mean LLM-assigned severity (1–10) by problem type across six testing methods on GPT-4o (gold-standard matching). Severity is pooled across all methods for matched true positives. Confidence is the mean respondent rating (1–5) pooled across CI configurations only.

Problem type	N	Cognitive testing			Expert review			Severity	Confidence
		Open	Struct.	List	Checklist	Open	Restr.		
Double-barrelled	6	100	67	100	100	100	100	7.18	4.07
Presupposition	9	44	33	44	89	11	56	6.18	3.91
Vague/ambiguous	9	56	11	78	67	67	78	5.57	4.37
Response-mapping	7	57	43	71	43	0	29	5.33	4.30
Retrieval/recall	5	100	0	100	100	80	80	5.64	4.32
Leading/loaded	5	40	20	100	80	40	80	6.33	3.80
Sensitivity	4	25	50	50	50	50	50	6.07	4.29
Complex syntax	3	67	33	67	33	67	100	5.97	4.04

Finally, to test whether the pipeline’s performance generalises beyond GPT-4o, all six configurations were replicated on Llama-3.3-70B-Instruct and Qwen3-32B. Table 5 presents the full results.

Table 5: Three-model comparison across all six configurations. Det. = detection rate. FP/ESS = false positives per ESS control question.

Config	GPT-4o		Llama-3.3-70B		Qwen3-32B	
	Det.	FP/ESS	Det.	FP/ESS	Det.	FP/ESS
CI-Open	60.4%	1.6	83.3%	5.4	83.3%	7.0
CI-Structured	31.2%	0.6	85.4%	4.0	87.5%	5.2
CI-List	75.0%	1.8	91.7%	5.4	83.3%	6.2
ER-Checklist	72.9%	2.6	89.6%	3.5	77.1%	2.9
ER-Open	47.9%	0.4	47.9%	0.8	6.2%	0.1
ER-Restrained	68.8%	3.1	75.0%	3.1	64.6%	2.0

On first inspection, the open-weight models appear to substantially outperform GPT-4o, with detection rates exceeding 80% on several configurations. On closer examination, however, these higher detection rates are largely a consequence of the open-weight models reporting far more problems overall. All open-weight cognitive testing configurations produce 4 to 7 problems per ESS question, compared to GPT-4o’s 0.6 to 1.8, meaning they flag as many or more problems on well-designed questions as on flawed ones. This represents an important trade-off for practical application, though not one that is necessarily disqualifying: a high-recall, low-precision configuration is well-suited to early-stage screening, where every flagged question is reviewed anyway and the cost of a missed problem is higher than the cost

of a false alarm. It is less well-suited to automated prioritisation, where the tool’s output is taken at face value, because it offers little discrimination between questions that need revision and those that do not.

Model sensitivity to the content of the prompt is further illustrated by ER-Open, which explicitly instructs reviewers to return an empty list if the question is well designed. On GPT-4o this produces a useful balance of 47.9% detection with just 0.4 false positives per ESS question (Table 5). The same prompt achieves a similar detection level on Llama (47.9%) but drops to just 6.2% on Qwen3, where the restraint language suppresses output almost entirely.

These cross-model differences should be interpreted in the context of the study design. The prompts were developed and refined on GPT-4o and then applied to the open-weight models without modification, a point to which we return in the Discussion.

## 4 Discussion

Cognitive interviewing and expert review have become indispensable components of the survey development workflow, providing the primary means by which researchers identify design flaws in draft questions before they reach the field. Yet both remain significant bottlenecks. They require scarce expertise, are time-consuming and expensive, and in practice can cover only a fraction of the items in a long or complex questionnaire. The practical consequence is that many questions go to field with less testing than would be ideal, and some go with none at all. If LLMs can simulate these diagnostic processes with reasonable fidelity, the implications for survey practice would be considerable, not as a replacement for human expertise, but as a means of extending diagnostic coverage to contexts where intensive pretesting is currently out of reach.

The results of this study suggest that the potential is real, if uneven. The best-performing simulated cognitive testing configuration, CI-List, detected three quarters of the 48 known problems embedded in the test set. This is a substantial proportion, particularly given that some of the embedded problems were deliberately subtle. After independent human adjudication of the false positives, CI-List’s false positive rate on flawed questions fell from 1.8 to 0.5 per question, indicating that many of the problems it flagged beyond the answer key were genuine design issues that the answer key had not anticipated.

The comparison between simulated cognitive testing and expert review revealed the kind of complementarity that has long been observed in the human versions of these methods. The cognitive testing configurations were more effective at identifying response-mapping failures, where the simulated respondent’s attempt to answer the question exposed the mismatch between their experience and the available options. Expert review, by contrast, was stronger on presupposition and leading framing, problems that are more readily identified through abstract analysis of question structure than through the experience of answering. Neither approach dominated the other across all problem types, suggesting that their integration is likely more powerful than either alone.

Detection also varied substantially across problem types, in ways that suggest clear avenues for improvement. Double-barrelled questions were detected near-universally, presumably because asking about two things at once is a salient and well-documented flaw. Retrieval and recall problems were similarly well detected, perhaps because they create obvious difficulty when a simulated respondent attempts to answer. At the other end of the spectrum, presupposition was the hardest problem type to detect without explicit prompting, the CI-Open configuration, which gave no guidance on what to look for, detected only 44% of presupposition problems, while ER-Checklist, whose checklist explicitly named the category, reached 89%. Sensitivity problems were also poorly detected across all configurations, with no approach exceeding 50%. These patterns suggest two complementary routes to improvement. The first is prompt refinement, adding explicit cues for problem types that the pipeline currently tends to miss, which is most likely to help with defects that can be recognised from the question text alone, such as ambiguous terms, leading framing, or double-barrelled structure. The second is adjusting the respondent sample, selecting backstories purposively to include profiles for whom a particular kind of defect is salient, which is more likely to help with defects whose visibility depends on the respondent’s circumstances, such as presupposition and applicability failures.

The open-weight models achieved higher detection rates than GPT-4o on most configurations, but this came at a substantial cost to specificity, with false positive rates on ESS control questions three to five times higher than GPT-4o’s. For any practical application, the ability to discriminate between questions

that need revision and those that do not is at least as important as raw detection, since a tool that flags everything provides no more guidance than one that flags nothing.

The cross-model comparison should be interpreted with caution, however, since it does not constitute a fair test of the models' underlying capabilities. The prompts were developed and refined on GPT-4o and applied to the open-weight models without modification. The CI-Structured reversal illustrates how consequential this is, the same four-stage cognitive framework that produces compliant, uncritical respondents on GPT-4o generates detailed problem articulation on Llama and Qwen3, suggesting that the prompt interacts with each model's alignment characteristics in ways that cannot be predicted without empirical testing. In practice, deploying the pipeline on a different model would require model-specific prompt optimisation, and it is plausible that the open-weight models could achieve a better detection-precision balance with prompts tailored to their characteristics. The key conclusion then is that prompt design is model-dependent, results obtained with prompts optimised for one model cannot be generalised to others without recalibration. This also implies that the principles governing effective prompt design for LLM-based cognitive testing cannot be derived a priori but must be calibrated empirically for each model, an important consideration as the landscape of available models continues to evolve.

Questions containing presuppositions that made them clearly inapplicable, or leading framing that created discomfort, produced markedly lower respondent confidence. Questions whose problems were subtler, involving vague terminology or recall difficulty, did not. This pattern maps onto a well-established distinction in the cognitive interviewing literature between problems that create experienced difficulty for the respondent and problems that silently degrade measurement quality without the respondent noticing (Presser and Blair 1994). Confidence ratings may therefore provide a useful screening signal, flagging questions where simulated respondents sense something is wrong, but they cannot substitute for explicit problem identification.

The computational costs of LLM-based question evaluation are modest. Each simulated cognitive testing evaluation of a single question requires 61 API calls (30 respondent simulations, 30 analyst codings, and one synthesis call) and takes approximately 90 to 135 seconds of wall-clock time on GPT-4o, at an estimated cost of approximately \$0.50 per question. Each expert review evaluation requires 4 API calls (3 independent reviewers plus synthesis) and completes in approximately 10 to 20 seconds at a cost of approximately \$0.02 per question. A full simulated cognitive testing evaluation of a 20-item questionnaire therefore costs under \$10 and completes in under an hour; expert review of the same instrument costs under \$1 and completes in minutes. For open-weight models served on local hardware, the marginal cost per question is close to zero. Both approaches are dramatically faster and cheaper than human cognitive interviewing, which typically requires days of interviewer time, participant recruitment and compensation, transcription, and analysis for a comparable number of questions.

Several limitations of the present study warrant acknowledgement. The test set comprised deliberately constructed items with known embedded problems, which provides the identification strategy needed to compute detection rates but limits external validity. The embedded problems are by design more clearly defined than many issues encountered in real questionnaire development, where defects may be subtle, overlapping, or context-dependent. Reported detection rates are therefore likely to represent an upper bound on what would be achieved with naturally occurring draft questions. Evaluating LLM-based approaches on organically generated questionnaires is an important direction for future work, but would require alternative evaluation frameworks that do not rely on a fixed answer key.

The current pipeline is also a simplification of real cognitive interviewing, which is fundamentally interactive in nature. A human interviewer follows up on the respondent's initial answer with targeted probes, pursuing lines of difficulty as they emerge. Our pipeline relies on a single prompt to the respondent, omitting this dynamic element. Future work could explore multi-turn interview simulations in which an LLM interviewer probes the LLM respondent dynamically. More broadly, several parameters that were held constant in this study warrant systematic investigation, including the content and length of respondent backstories, the temperature setting (fixed at 0.7 throughout), and the number of simulated respondents per question. Each of these is likely to interact with model choice and prompt design in ways that cannot be predicted without empirical testing.

This is early work and the landscape and capabilities of available LLMs is changing rapidly. The results reported here should, therefore, be understood as an initial benchmark rather than a ceiling on what LLM-based question testing can achieve. Language models will continue to improve in their ability to follow complex instructions, simulate diverse perspectives, and exercise critical judgement, and the performance

of LLM-based question testing will surely improve with them. LLM-based question evaluation is unlikely to replace skilled cognitive interviewing conducted by experienced practitioners. But it has the potential to become a useful complement, particularly in contexts where human pretesting is constrained by budget, timeline, or the sheer volume of items that need to be screened. The ability to run a comprehensive diagnostic evaluation of an entire questionnaire in under an hour, at minimal cost, and to do so repeatedly as questions are revised, represents a genuine practical opportunity for the field.

## References

- Aher, Gati V., Rosa I. Arriaga and Adam Tauman Kalai (3rd July 2023). ‘Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies’. In: *Proceedings of the 40th International Conference on Machine Learning*. International Conference on Machine Learning, PMLR, pp. 337–371. URL: <https://proceedings.mlr.press/v202/aher23a.html> (visited on 14/10/2025).
- Argyle, Lisa P. et al. (July 2023). ‘Out of One, Many: Using Language Models to Simulate Human Samples’. In: *Political Analysis* 31.3, pp. 337–351. ISSN: 1047-1987, 1476-4989. DOI: 10.1017/pan.2023.2. URL: <https://www.cambridge.org/core/journals/political-analysis/article/out-of-one-many-using-language-models-to-simulate-human-samples/035D7C8A55B237942FB6DBAD7CAA4E49> (visited on 09/10/2023).
- Atari, Mohammad et al. (22nd Sept. 2023). *Which Humans?* DOI: 10.31234/osf.io/5b26t. URL: [https://osf.io/5b26t\\_v1](https://osf.io/5b26t_v1) (visited on 08/10/2025).
- Beatty, Paul C. and Gordon B. Willis (1st Jan. 2007). ‘Research Synthesis: The Practice of Cognitive Interviewing’. In: *Public Opinion Quarterly* 71.2, pp. 287–311. ISSN: 0033-362X. DOI: 10.1093/poq/nfm006. URL: <https://doi.org/10.1093/poq/nfm006> (visited on 19/12/2025).
- Blair, Johnny and Frederick G. Conrad (2011). ‘Sample Size for Cognitive Interview Pretesting’. In: *The Public Opinion Quarterly* 75.4, pp. 636–658. ISSN: 0033-362X. URL: <https://www.jstor.org/stable/41288411> (visited on 10/12/2025).
- Cannell, Charles F., Peter V. Miller and Lois Oksenberg (1981). ‘Research on Interviewing Techniques.’ In: *Sociological Methodology* 12.389.
- Caporaso, Andrew and Stanley Presser (1st Nov. 2024). ‘The Prevalence and Nature of Cognitive Interviewing as a Survey Questionnaire Evaluation Method in the United States’. In: *Journal of Survey Statistics and Methodology* 12.5, pp. 1278–1294. ISSN: 2325-0992. DOI: 10.1093/jssam/smad047. URL: <https://doi.org/10.1093/jssam/smad047> (visited on 19/12/2025).
- Collins, Debbie (2015). *Cognitive Interviewing Practice*. SAGE Publications Ltd. ISBN: 978-1-4739-1010-2. DOI: 10.4135/9781473910102. URL: <https://methods.sagepub.com/book/edvol/cognitive-interviewing-practice/toc> (visited on 14/04/2026).
- Dillion, Danica et al. (1st July 2023). ‘Can AI language models replace human participants?’ In: *Trends in Cognitive Sciences* 27.7, pp. 597–600. ISSN: 1364-6613. DOI: 10.1016/j.tics.2023.04.008. URL: <https://www.sciencedirect.com/science/article/pii/S1364661323000980> (visited on 26/09/2025).
- Dominguez-Olmedo, Ricardo, Moritz Hardt and Celestine Mendler-Dünner (9th Dec. 2024). *Questioning the Survey Responses of Large Language Models*. DOI: 10.48550/arXiv.2306.07951. arXiv: 2306.07951[cs]. URL: <http://arxiv.org/abs/2306.07951> (visited on 16/03/2026).
- Heyde, Leah von der, Anna-Carolina Haensch and Alexander Wenz (26th Apr. 2025). ‘Vox Populi, Vox AI? Using Large Language Models to Estimate German Vote Choice’. In: *Social Science Computer Review*, p. 08944393251337014. ISSN: 0894-4393. DOI: 10.1177/08944393251337014. URL: <https://doi.org/10.1177/08944393251337014> (visited on 26/09/2025).
- Presser, Stanley and Johnny Blair (1994). ‘Survey Pretesting: Do Different Methods Produce Different Results?’ In: *Sociological Methodology* 24, pp. 73–104. ISSN: 0081-1750. DOI: 10.2307/270979. URL: <https://www.jstor.org/stable/270979> (visited on 09/01/2026).
- Presser, Stanley et al., eds. (2004). *Methods for Testing and Evaluating Survey Questions*. Section: 1. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471654728.ch1>. John Wiley & Sons, Ltd. ISBN: 978-0-471-65472-8. DOI: 10.1002/0471654728.ch1. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/0471654728.ch1> (visited on 14/04/2026).
- Qi, Weihong, Hanjia Lyu and Jiebo Luo (16th July 2024). *Representation Bias in Political Sample Simulations with Large Language Models*. DOI: 10.48550/arXiv.2407.11409. arXiv: 2407.11409[cs]. URL: <http://arxiv.org/abs/2407.11409> (visited on 26/09/2025).
- Santurkar, Shibani et al. (3rd July 2023). ‘Whose Opinions Do Language Models Reflect?’ In: *Proceedings of the 40th International Conference on Machine Learning*. International Conference on Machine

- Learning. PMLR, pp. 29971–30004. URL: <https://proceedings.mlr.press/v202/santurkar23a.html> (visited on 08/10/2025).
- Sarstedt, Marko et al. (2024). ‘Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines’. In: *Psychology & Marketing* 41.6. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.21982>, pp. 1254–1270. ISSN: 1520-6793. DOI: 10.1002/mar.21982. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/mar.21982> (visited on 26/09/2025).
- Wang, Angelina, Jamie Morgenstern and John P. Dickerson (Mar. 2025). ‘Large language models that replace human participants can harmfully misportray and flatten identity groups’. In: *Nature Machine Intelligence* 7.3, pp. 400–411. ISSN: 2522-5839. DOI: 10.1038/s42256-025-00986-z. URL: <https://www.nature.com/articles/s42256-025-00986-z> (visited on 26/09/2025).
- Willis, Gordon (2005). *Cognitive Interviewing*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc. ISBN: 978-0-7619-2804-1 978-1-4129-8365-5. DOI: 10.4135/9781412983655. URL: <https://methods.sagepub.com/book/cognitive-interviewing> (visited on 18/03/2026).

## Data availability statement

Data and code to replicate the tables reported in this paper are available at <https://github.com/PatrickSturgis/cogbot-replication>. The pipeline backend can be installed and run from <https://github.com/PatrickSturgis/cogbot>. A prototype web-based dashboard is available at <https://huggingface.co/spaces/Patricksturg/silicon-sampling-dashboard>, allowing users to run the pipeline with their own API keys.

## Appendix A AAPOR-Required Disclosure Elements

The data analysed in this paper are primarily LLM-generated outputs produced by the simulated cognitive testing and expert review pipelines described in the Methodology. These data do not fit the canonical frame of survey research on human respondents, and several AAPOR disclosure elements therefore do not apply. We list all elements below, with an explanation where a given element is not applicable. The paper also makes secondary use of microdata from the 2018 European Social Survey (UK), which is documented as a second data source.

### Data source: LLM-generated question-evaluation outputs

**Type and source of data:** New data collected by the authors. Outputs were generated by prompting large language models via API and recording their responses. No human respondents were involved in producing these data.

**Data Collection Strategy:** Automated prompting of LLMs through a Python-based pipeline (referred to in the paper as CogBot). For simulated cognitive testing, each test question was evaluated in three stages: (i)  $n = 30$  simulated respondents each produced a think-aloud transcript in response to the question, conditioned on a demographic backstory; (ii) an analyst LLM coded each transcript for evidence of question-design problems; (iii) a synthesis LLM consolidated findings across the 30 respondents. For expert review, three LLM reviewers evaluated each question directly, followed by a synthesis step. Six prompt configurations (three cognitive testing, three expert review) were run across three models.

**Research Sponsor and Conductor:** The research was conducted by the authors at their respective institutions and funded by the Economic and Social Research Council (ESRC).

**Measurement Tools/Instruments:** All prompt texts used in the pipeline are provided verbatim in Appendix C. The full test set of 27 questions (20 deliberately flawed items and 7 ESS control items) is provided in Appendix B.

**Population Under Study:** Not applicable in the conventional sense. The unit of analysis is the survey question ( $n = 27$ ), not a population of respondents. LLM outputs conditioned on 2,204 synthetic respondent backstories (see Data Source 2) serve as the source of variation in simulated cognitive testing.

**Methods Used to Generate and Recruit the Sample:** Not a probability sample in the conventional sense. For each simulated cognitive testing run, 30 backstories were drawn by simple random sampling without replacement from the pool of 2,204 backstories derived from the 2018 ESS UK (see Data Source 2). Expert review uses three fixed reviewers per question with no backstory sampling. No quotas were used. No human participants were recruited; no incentives were provided.

**Method(s) and Mode(s) of Data Collection:** API calls to three large language models, all prompts and responses in English:

- GPT-4o (OpenAI), proprietary, accessed via the OpenAI API at temperature 0.7.
- Llama-3.3-70B-Instruct (Meta), open-weight, served locally via SGLang with AWQ 4-bit quantisation on a single NVIDIA RTX 6000 Ada 49GB GPU.
- Qwen3-32B (Alibaba), open-weight, served locally via SGLang with FP8 quantisation on the same hardware.

Matching of pipeline-identified problems to the answer key was performed by Claude Opus 4.6 (Anthropic) and GPT-5.4 (OpenAI) at temperature 0.

**Dates of Data Collection:** Not applicable in the conventional sense. The study does not rely on a fixed fieldwork period; pipeline runs were executed across the course of the project and can be re-run at any time given the prompts, test set, and model identifiers documented here.

**Sample Sizes and Precision:** The unit of analysis is the question ( $n = 27$ : 20 flawed, 7 ESS controls). For each cognitive testing configuration and model, 30 simulated respondents were generated per question. For each expert review configuration and model, three reviewer outputs were generated per question. Six configurations  $\times$  three models  $\times$  27 questions yields the full cross. No sampling-error estimates are reported: the study is not a sample survey of a population, and reported detection and false positive rates are descriptive summaries of pipeline behaviour on a fixed question set. Inter-judge agreement for

answer-key matching was 94.4% across 764 problem pairs evaluated by the two LLM judges (Claude Opus 4.6 and GPT-5.4); the 43 discordant pairs were adjudicated by the first author.

**Whether and How the Data Were Weighted:** Not applicable. Data were not weighted.

**How the Data Were Processed and Procedures to Ensure Data Quality:** Temperature, prompts, backstories, and all other parameters were held constant across models. Identical prompt texts were used for every question within a configuration. Coding of pipeline outputs into problem reports was performed by the models themselves under the analyst and synthesis prompts provided in Appendix C. Matching of pipeline-identified problems to the answer key was performed by two independent LLM judges with disagreements resolved by the first author. Pipeline outputs were stored as structured records (JSON) with full transcripts retained for audit.

**Panel Description:** Not applicable. No human or persistent-agent panel was used.

**Interviewer Details:** Not applicable. No human interviewers were involved. The simulated interviewer role is performed by the LLM under the think-aloud prompts documented in Appendix C. The LLM judges used for answer-key matching are described above.

**Screening Criteria and Process:** Not applicable. No human respondents were screened.

**Study Stimuli:** The 27 test questions (20 deliberately flawed items and 7 ESS control items) constitute the study stimuli and are reproduced in full in Appendix B.

**Dispositions or Response or Participation Rates:** Not applicable. LLM API calls do not have meaningful dispositions or non-response. All scheduled calls in the benchmark returned usable output; isolated API errors were retried.

**Measurement and Model Specification:** The pipeline architecture is fully described in the Methodology section and the prompts are reproduced in Appendix C. Replication code, the full test set and answer key, and archived pipeline outputs are available at the repository listed in the Data Availability Statement.

**Limitations:** Reported detection rates are based on a test set of deliberately constructed items with known embedded problems and therefore represent an upper bound on what would be achieved with naturally occurring draft questions. The pipeline relies on a single-prompt think-aloud rather than multi-turn dynamic probing. Temperature, number of simulated respondents, and backstory content were held constant; other settings are likely to affect results. These limitations are discussed in the Discussion section.

## Appendix B Test Set

### Appendix B.1 Flawed questions (n = 20, 48 known problems)

ID / Topic	Question	Scale	Embedded problems
F1. Local council services	“How satisfied or dissatisfied are you with the recycling services and street cleaning provided by your local council?”	Very satisfied to Very dissatisfied	(P1) double-barrelled: recycling and street cleaning are distinct services; (P2) presupposition: assumes council provides both services; (P3) presupposition: assumes respondent has experience of services.
F2. Financial well-being	“How would you describe your current financial situation?”	Living comfortably to Finding it very difficult	(P1) vague/ambiguous: “current financial situation” could refer to income, savings, debt, or wealth; (P2) response-mapping: categories use inconsistent linguistic constructions.
F3. Physical activity	“In a typical week, how many hours do you spend doing vigorous physical activity and moderate physical activity combined?”	None to More than 10 hours	(P1) double-barrelled: vigorous and moderate activity are distinct behaviours; (P2) retrieval/recall: summing hours across multiple activities requires difficult mental calculation.
F4. Immigration attitudes	“Given the growing pressure on public services, don’t you think that the government should do more to reduce the number of immigrants coming to this country, particularly those who may not contribute to the economy?”	Strongly agree to Strongly disagree	(P1) leading/loaded: premise stated as fact, negative interrogative implies expected agreement; (P2) complex-syntax: conditional premise, negative interrogative, main proposition, and subordinate clause.
F5. Job satisfaction	“Overall, how satisfied are you with your terms of employment?”	Completely satisfied to Not at all satisfied	(P1) vague/ambiguous: “terms of employment” is ambiguous; (P2) response-mapping: asymmetric scale (3 positive, 2 negative); (P3) presupposition: assumes respondent currently has defined terms of employment.
F6. Household income	“What is your household’s total monthly income from all sources, before tax and deductions?”	Less than £1,000 to £5,000 or more	(P1) sensitivity: precise income is a highly sensitive question; (P2) presupposition: assumes identifiable household income.
F7. Transport and commuting	“How easy or difficult is it for you to get to work and to access essential services such as healthcare from where you live?”	Very easy to Very difficult	(P1) double-barrelled: work access and service access are distinct; (P2) response-mapping: no way to distinguish components, no N/A option; (P3) presupposition: assumes respondent has a job.
F8. Internet use	“Roughly how many hours per week do you spend on social media platforms?”	None to More than 20 hours	(P1) vague/ambiguous: “social media platforms” is undefined; (P2) retrieval/recall: estimating weekly usage across fragmented sessions is difficult.
F9. Criminal justice attitudes	“People who commit serious crimes deserve to be punished severely, even if this means longer prison sentences. To what extent do you agree or disagree?”	Strongly agree to Strongly disagree	(P1) leading/loaded: frames punishment as deserved; (P2) sensitivity: social desirability around criminal justice attitudes.

ID / Topic	Question	Scale	Embedded problems
F10. Neighbourhood safety	“How safe do you feel in your neighbourhood during the day and after dark?”	Very safe to Very unsafe	(P1) double-barrelled: day and night safety may differ; (P2) vague/ambiguous: “neighbourhood” has no agreed boundary.
F11. Healthcare access	“In the past 12 months, how easy has it been to get an appointment with a healthcare professional when you needed one?”	Very easy to Very difficult	(P1) vague/ambiguous: “healthcare professional” could mean GP, consultant, dentist, etc.; (P2) retrieval/recall: aggregating across 12 months of varied experiences; (P3) presupposition: assumes respondent has sought an appointment.
F12. Gender roles	“Research consistently shows that children benefit from having a parent at home in the early years. How important do you think it is that mothers stay at home to look after their children before they start school?”	Very important to Not at all important	(P1) leading/loaded: appeal to authority frames proposition as evidence-based; (P2) sensitivity: shift from “parent” to “mothers” introduces gender role discomfort; (P3) response-mapping: no neutral midpoint.
F13. Life satisfaction	“Taking everything into account, how satisfied are you with the way things are going in your life at this moment?”	0 (Extremely dissatisfied) to 10 (Extremely satisfied)	(P1) vague/ambiguous: “the way things are going” conflates trajectory with current state; (P2) response-mapping: only endpoints labelled on 0 to 10 scale.
F14. Childcare arrangements	“How satisfied are you with the availability and affordability of childcare services in your area?”	Very satisfied to Very dissatisfied	(P1) double-barrelled: availability and affordability may diverge; (P2) presupposition: assumes need for childcare.
F15. Environmental behaviour	“Considering the urgent need to address climate change and the fact that individual actions collectively make a significant difference, to what extent have you made changes to your daily habits specifically to reduce your personal carbon footprint over the past year?”	A great deal to Not at all	(P1) leading/loaded: preamble presents contested claims as fact; (P2) complex-syntax: long sentence with conditional premise, domain specification, purpose clause, and time reference.
F16. Volunteering	“How many times in the past year have you done voluntary or unpaid work for a community organisation?”	None to More than 10 times	(P1) vague/ambiguous: “community organisation” is vague; (P2) retrieval/recall: counting instances over 12 months is difficult; (P3) vague/ambiguous: “past year” could mean last 12 months or last calendar year.
F17. Housing situation	“How adequate is the amount of space in the home where you currently live for the needs of your household?”	Far too much space to Far too little space	(P1) presupposition: assumes respondent has a fixed home; (P2) response-mapping: options address only space quantity but question asks about adequacy for needs.
F18. Relationship quality	“How often do you and your partner argue about money and about how to raise your children?”	Very often to Never	(P1) double-barrelled: money and child-rearing may have different frequencies; (P2) sensitivity: reporting arguments is sensitive; (P3) presupposition: assumes partner and children.

ID / Topic	Question	Scale	Embedded problems
F19. Political trust	“Given the recent scandals in government, how much trust do you have in politicians to act in the public interest?”	A great deal of trust to No trust at all	(P1) leading/loaded: negative preamble primes low-trust responses; (P2) response-mapping: asymmetric scale interacts with leading preamble.
F20. Time use	“On a typical weekday, approximately how many hours do you spend on leisure activities, not including time spent on domestic tasks, caring responsibilities, or personal grooming, but including any time spent on hobbies, socialising, watching television, reading, or using the internet for personal enjoyment?”	None to More than 6 hours	(P1) complex-syntax: extremely long with inclusion/exclusion lists; (P2) vague/ambiguous: boundaries between included and excluded activities are unclear; (P3) retrieval/recall: estimating fragmented leisure time is unreliable.

## Appendix B.2 ESS-tested control questions (n = 7, ESS core items)

ID / Topic	Question	Scale
C1. General happiness	“Taking all things together, how happy would you say you are?”	0 (Extremely unhappy) to 10 (Extremely happy)
C3. Political interest	“How interested would you say you are in politics: very interested, quite interested, hardly interested, or not at all interested?”	Very interested to Not at all interested
C4. Self-rated health	“How is your health in general? Would you say it is very good, good, fair, bad, or very bad?”	Very good to Very bad
C5. Life satisfaction	“All things considered, how satisfied are you with your life as a whole nowadays?”	0 (Extremely dissatisfied) to 10 (Extremely satisfied)
C6. Climate change worry	“How worried are you about climate change?”	Not at all worried to Extremely worried
C7. Fruit consumption	“How often do you eat fruit, excluding drinking juice?”	Three times or more a day to Never
C8. Control over life	“How much control do you feel you have over your life in general nowadays?”	0 (No control at all) to 10 (Complete control)

## Appendix C Prompt texts

The following are the full prompt templates for each configuration. Placeholders in braces (`{backstory}`, `{question.text}`, `{response.options}`) are populated at runtime.

### CI-Open (respondent prompt)

WHO YOU ARE:  
`{backstory}`

You are this person. Respond in first person throughout (I, me, my).  
Do not break character or refer to yourself in the third person.

SURVEY QUESTION:  
{question\_text}

RESPONSE OPTIONS: {response\_options}

Think aloud as you work through answering this question. Talk through your thought process step by step | what comes to mind as you read the question, how you go about forming an answer, and any difficulties you run into along the way. Note anything confusing, unclear, or awkward as you encounter it, referring to specific words, phrases, or features of the question and the response options.

CHOSEN ANSWER: State your final answer.

CONFIDENCE: Rate your confidence 1-5, where 1 means you felt very unsure about your answer and 5 means you felt completely certain.

## CI-Structured (respondent prompt)

WHO YOU ARE:  
{backstory}

You are this person. Respond in first person throughout (I, me, my).  
Do not break character or refer to yourself in the third person.

SURVEY QUESTION:  
{question\_text}

RESPONSE OPTIONS: {response\_options}

Think aloud as you work through answering this question, following the stages below:

1. **COMPREHENSION:** Put the question in your own words. What is it actually asking you? This is about whether the meaning of the question is clear | whether you understand what information it wants from you and whether any words or phrases are ambiguous or confusing.
2. **RETRIEVAL:** What from your life or experience are you drawing on to answer this? This is about the mental work of finding the relevant information | whether you can recall what you need, whether the question asks you to remember things that are hard to remember, or to make estimates that feel uncertain.
3. **JUDGEMENT:** Are you having to weigh up, combine, or summarise different things to arrive at a single answer? This is about whether the question forces you to compress multiple views, experiences, or considerations into one response, and whether that feels straightforward or uncomfortable.
4. **RESPONSE MAPPING:** Look at the response options. Which one fits best? How well does it capture what you actually want to say? This is about whether your answer maps cleanly onto the available options, or whether something is missing, the categories feel uneven, or you have to force your answer into an option that doesn't quite fit.
5. **CHOSEN ANSWER:** State your final answer.

6. CONFIDENCE: Rate your confidence 1-5, where 1 means you felt very unsure about your answer and 5 means you felt completely certain.

## CI-List (respondent prompt)

WHO YOU ARE:  
{backstory}

You are this person. Respond in first person throughout (I, me, my).  
Do not break character or refer to yourself in the third person.

SURVEY QUESTION:  
{question\_text}

RESPONSE OPTIONS: {response\_options}

Think aloud as you work through answering this question, following the stages below. At each stage, consider whether you experience any of the specific difficulties described. Not all of the difficulties listed below will apply | only mention those you actually experience as you work through the question.

1. COMPREHENSION: Put the question in your own words. What is it actually asking you?

As you do this, consider:

- DOUBLE-BARRELLED: Is the question asking about two or more distinct things at once, where you might want to give different answers to each part?
- VAGUE OR AMBIGUOUS TERMS: Are any words or phrases unclear, open to multiple interpretations, or likely to mean different things to different people?
- PRESUPPOSITION: Does the question assume something about you or your situation that may not be true?
- LEADING OR LOADED: Does the phrasing push you toward a particular answer, for example through emotive language, built-in justifications, or a question structure that implies one answer is more natural or correct?

2. RETRIEVAL: What from your life or experience are you drawing on to answer this?

As you do this, consider:

- RECALL DIFFICULTY: Is it hard to remember the information the question asks about?
- REFERENCE PERIOD PROBLEMS: Does the question specify a time frame that is difficult to recall over, or is the time frame vague or missing entirely?

3. JUDGEMENT: Are you having to weigh up, combine, or summarise different things to arrive at a single answer?

As you do this, consider:

- COMPLEX AGGREGATION: Does the question require you to combine multiple different experiences into a single summary response in a way that feels forced or oversimplified?
- KNOWLEDGE DEFICIT: Is the question asking you to make a judgement about something you don't feel you know enough about?

4. RESPONSE MAPPING: Look at the response options. Which one fits best? How well does it capture what you actually want to say?

As you do this, consider:

- POOR FIT: Does your answer not map clearly onto any of the available options?
- MISSING OPTIONS: Is there a response you want to give that isn't available?
- UNEQUAL INTERVALS: Do the response categories feel unevenly spaced?

5. CHOSEN ANSWER: State your final answer.

6. CONFIDENCE: Rate your confidence 1-5, where 1 means you felt very unsure about your answer and 5 means you felt completely certain.

## Cognitive interview analyst prompt

System: You are an expert in cognitive interviewing and survey methodology. You are reviewing a think-aloud transcript from a cognitive interview. Your job is to identify problems with the question based on evidence in the transcript.

User:

SURVEY QUESTION:

{question\_text}

RESPONSE OPTIONS: {response\_options}

RESPONDENT THINK-ALoud TRANSCRIPT:

{transcript}

Identify ONLY problems that are clearly evidenced in this transcript. Do NOT go looking for problems that are not there. Most transcripts will have only 1-2 genuine problems, and some will have none. If the respondent navigated the question without difficulty, return an empty problems list.

A problem must be supported by specific evidence in the transcript - something the respondent actually said or demonstrably struggled with. Do not infer problems that the respondent did not experience.

When a genuine problem is found, classify it using whichever of these categories fits best:

- COMPREHENSION PROBLEMS: Respondent misunderstood or misinterpreted terms
- DOUBLE-BARRELED: Question asks about two distinct things requiring different answers
- RETRIEVAL DIFFICULTIES: Respondent struggled to recall or estimate
- RESPONSE MAPPING FAILURES: Answer did not fit the available options
- SOCIAL DESIRABILITY / SENSITIVITY: Respondent hedged or showed discomfort
- PRESUPPOSITION FAILURES: Question assumed something untrue for this respondent
- TAUTOLOGY / LOGICAL PROBLEMS: Circularity between question wording and response options

If the problem does not fit any category above, use a short descriptive label.

For each problem:

- Cite specific evidence from the transcript (what the respondent actually said)
- Rate SEVERITY 1-10: the degree to which this problem would distort the accuracy of the answer (1 = negligible distortion, 10 = answer rendered meaningless)

Return ONLY valid JSON, no markdown fencing:

```
{"respondent_id": "{respondent_id}", "problems": [{"type": "...", "description": "...", "evidence": "...", "severity": 7}]}
```

## Cognitive interview synthesis prompt

System: You are a senior survey methodologist writing a cognitive testing report. You are synthesising the results of individual transcript analyses from a cognitive interview study.

User:

SURVEY QUESTION:

{question\_text}

RESPONSE OPTIONS: {response\_options}

Below are the coded problems identified by an analyst for each respondent's think-aloud transcript. Each problem has a severity score (1-10) indicating the degree to which it would distort accuracy of the answer. Your job is to aggregate these into a summary of the distinct problems found across respondents. Use the SAME problem type labels as the individual analyses (e.g. if analysts coded "DOUBLE-BARRELLED", use that label, not a paraphrase).

INDIVIDUAL ANALYSES:

{analyses\_block}

For each distinct problem type that was identified:

- Use the same type label as in the individual analyses
- List which respondents showed evidence of it
- Calculate the mean severity score across respondents who had this problem
- Summarise the evidence across respondents

Return ONLY valid JSON, no markdown fencing:

```
{"question_id": "synthesis", "problems_detected": [{"type": "...", "description": "...", "respondents_affected": [...], "mean_severity": 6.5, "evidence_summary": "..."}]}
```

## ER-Restrained (expert review prompt)

SURVEY QUESTION:

{question\_text}

RESPONSE OPTIONS: {response\_options}

You are expert reviewer {expert\_id}. Evaluate this survey question for problems that would cause respondents difficulty when trying to answer it, or that would reduce the accuracy of the data collected.

A problem must be specific to THIS question | not a generic observation that could apply to any survey question using a rating scale or asking about an abstract concept.

For each genuine problem:

- Describe the problem in plain language
- Cite the specific words, phrases, or structural features that cause the problem
- Rate SEVERITY 1-10: the degree to which this problem would distort the accuracy of responses (1 = negligible, 10 = meaningless)

Return ONLY valid JSON, no markdown fencing:

```
{"expert_id": "{expert_id}", "problems": [{"description": "...",
"evidence": "...", "severity": 7}]}
```

## ER-Open (expert review prompt)

SURVEY QUESTION:  
{question\_text}

RESPONSE OPTIONS: {response\_options}

You are expert reviewer {expert\_id}. Evaluate this survey question for problems that would cause respondents difficulty when trying to answer it, or that would reduce the accuracy of the data collected.

Identify ONLY problems that are genuinely likely to affect respondents answering this question. Do NOT go looking for problems that are not there. Many well-designed survey questions will have no significant problems. If the question is clear, well-structured, and the response options are appropriate, return an empty problems list.

A problem must be specific to THIS question | not a generic observation that could apply to any survey question using a rating scale or asking about an abstract concept.

For each genuine problem:

- Describe the problem in plain language
- Cite the specific words, phrases, or structural features that cause the problem
- Rate SEVERITY 1-10: the degree to which this problem would distort the accuracy of responses (1 = negligible, 10 = meaningless)

Return ONLY valid JSON, no markdown fencing:

```
{"expert_id": "{expert_id}", "problems": [{"description": "...",
"evidence": "...", "severity": 7}]}
```

## Expert review synthesis prompt

SURVEY QUESTION:  
{question\_text}

RESPONSE OPTIONS: {response\_options}

Below are the independent evaluations from 3 expert reviewers. Each identified problem has a severity score (1-10). Your job is to aggregate these into a summary of the DISTINCT problems found across reviewers.

Focus on the UNDERLYING ISSUE. If multiple experts describe the same fundamental problem | even in different words or from different angles | merge them into a single problem. Two descriptions count as the same problem if fixing one would also fix the other.

Do NOT retain problems that are generic observations applicable to any survey question. Only retain problems specific to this question.

EXPERT REVIEWS:  
{reviews\_block}

For each distinct underlying problem:

- Describe the problem clearly
- List which experts flagged it
- Calculate the mean severity score across experts who flagged it
- Summarise the evidence across experts

Return ONLY valid JSON, no markdown fencing:

```
{"question_id": "synthesis", "problems_detected": [{"description": "...", "experts_affected": ["e1", "e2"], "mean_severity": 6.5, "evidence_summary": "..."}]}
```

*Note: The ER-Checklist prompt is embedded in the pipeline source code (`dashboard_backend.py`) and is reproduced in the replication package.*

## Appendix D Illustrative CI-Structured transcripts across models

The following transcripts illustrate the contrast between GPT-4o and Qwen3-32B on the CI-Structured configuration, applied to question F1 (“How satisfied or dissatisfied are you with the recycling services and street cleaning provided by your local council?”). Both respondents share the same backstory profile and follow the same four-stage cognitive framework. The GPT-4o respondent navigates each stage compliantly, acknowledging the double-barrelled structure but proceeding without flagging it as a problem. The Qwen3 respondent treats each stage as an opportunity to articulate difficulties, explicitly questioning whether the two services should be evaluated separately and scrutinising the response scale.

### GPT-4o (Respondent r3)

1. COMPREHENSION: The question is asking me how happy or unhappy I am with the recycling services and street cleaning that my local council provides. It wants to know my level of satisfaction or dissatisfaction with these services. The question seems clear to me, and I don’t find any words or phrases confusing.
2. RETRIEVAL: To answer this, I’m thinking about my experiences with recycling and street cleaning in my area. I need to recall how often I see recycling collected, whether it’s done efficiently, and how clean the streets are. I’m also considering if there have been any recent issues or improvements that stood out to me.
3. JUDGEMENT: I need to consider both recycling and street cleaning together, which means I have to weigh my satisfaction with both services to come up with one overall answer. This feels a bit tricky because I might feel differently about each service. However, generally, if one is noticeably better or worse, that might tip the scale for my overall judgment.
4. RESPONSE MAPPING: Looking at the options, I think "Fairly dissatisfied" might fit best. I recall some issues with recycling collections being inconsistent, and the streets don’t always seem as clean as they could be. This option seems to capture my general feeling that there’s room for improvement, but I’m not completely unhappy with the services.
5. CHOSEN ANSWER: Fairly dissatisfied.
6. CONFIDENCE: I’d rate my confidence at 4. I’m fairly sure about my answer based on my recent experiences, but there’s a bit of uncertainty because I might not be fully considering all aspects of the services.

*Analyst coding: One problem identified — difficulty combining satisfaction with two distinct services into a single response.*

### Qwen3-32B (Respondent r3)

1. COMPREHENSION: Okay, let me read the question again. It says, "How satisfied or dissatisfied are you with the recycling services and

street cleaning provided by your local council?" So, basically, they're asking me to rate how happy or unhappy I am with the work my local council does when it comes to recycling and keeping the streets clean. I think I get it. It's asking about two things: one is the recycling service -- like how often they collect, how easy it is, whether they sort it properly -- and the other is street cleaning -- like how clean the streets are, how often they sweep or clean up litter, that kind of thing. I don't think there's any confusion in the wording, but it does combine two separate services into one question. I wonder if I need to consider them as one or if I can treat them separately in my mind.

2. RETRIEVAL: Hmm, let me think. I live in the North West, so I guess that's somewhere like Greater Manchester, maybe Salford or Wigan -- I don't know the exact area, but I know the local council is responsible for these services. I don't do much recycling myself, to be honest. I've tried a bit, but it can be confusing sometimes. I think the bins are collected once a week, but I've seen people in my area put out the wrong stuff in the wrong bins, and then the council just ignores it or throws it all in the same lorry. So that feels like a bit of a waste. As for street cleaning, I don't think they do it very often. I've noticed a lot of litter in the streets, especially near the bus stops and outside supermarkets.

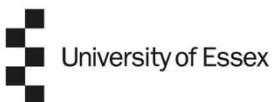
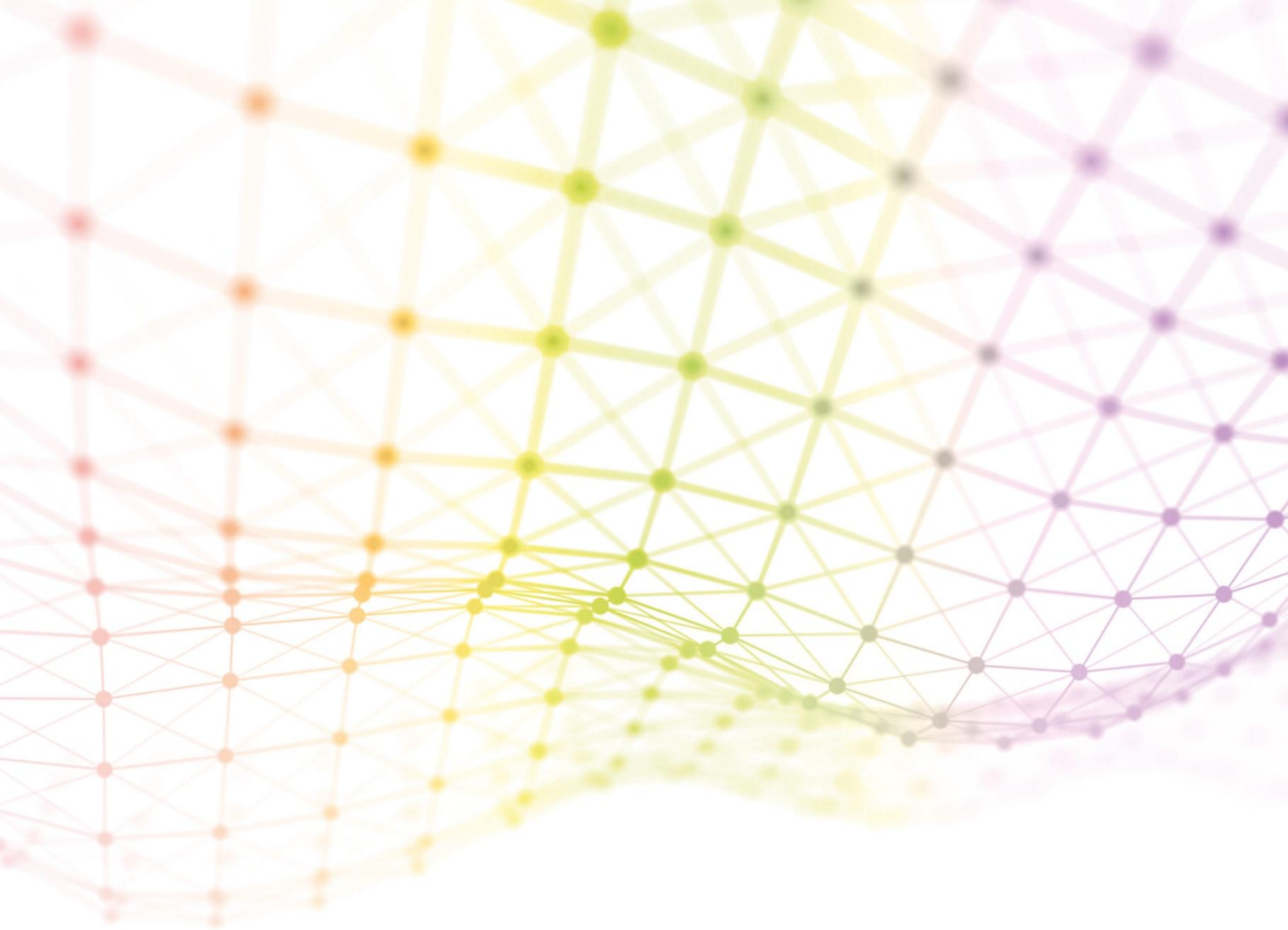
3. JUDGEMENT: So, I need to rate my overall satisfaction with both services combined. I don't hate it, but I'm not happy with it either. The recycling is a bit hit and miss, and the street cleaning is just not frequent enough. I would say that I'm leaning towards being dissatisfied, but not very dissatisfied. I wonder if the question wants me to consider how the service compares to other areas or just to my own expectations. I think it's about how I personally feel about the service I'm getting.

4. RESPONSE MAPPING: Looking at the options: Very satisfied, Fairly satisfied, Neither satisfied nor dissatisfied, Fairly dissatisfied, Very dissatisfied. I don't feel I can say I'm satisfied at all. I'm not neutral either -- there's definitely a sense of disappointment. So I think Fairly dissatisfied is the closest. I don't think the options are missing anything, but I do wonder if the categories are evenly spaced in terms of how people might feel. I don't think I'm being forced into a category that doesn't fit, but it's a bit of a compromise.

5. CHOSEN ANSWER: Fairly dissatisfied.

6. CONFIDENCE: I'd say I'm fairly confident in that answer -- maybe a 3 out of 5.

*Analyst coding: Two problems identified — (1) uncertainty about whether to evaluate the two services separately or together; (2) limited personal perspective constraining the evaluation.*



[www.surveyfutures.net](http://www.surveyfutures.net)