SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

Working Paper 22:

Survey question evaluation at scale: applying large language models to the QAS-99

Caroline Roberts[1], Patrick Sturgis[2], Alice McGee[3], Rebecca Hamlyn[3], Mayca Minichino[1], Thomas Robinson[2],

[1] University of Lausanne; [2] London School of Economics and Political Science, [3] Verian Group

July 2026

Acknowledgements

Survey Futures is an Economic and Social Research Council (ESRC)-funded initiative (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.

The research reported here forms part of Survey Futures Research Strand 9: Generative AI for questionnaire development and testing, led by Professor Patrick Sturgis.

Citation

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:

Roberts C., Sturgis, P., McGee, A., Hamlyn, R., Minichino M, & Robinson T. (*2026*) 'Survey question evaluation at scale: applying large language models to the QAS-99', *Survey Futures Working Paper* no. 22. Colchester, UK: University of Essex. Available at https://surveyfutures.net/working-papers/.

Contents

Survey Question Evaluation at Scale: Applying Large Language Models to the QAS-99

Caroline Roberts*[1], Patrick Sturgis[2], Alice McGee[3], Rebecca Hamlyn[3], Mayca Minichino[1], and Thomas Robinson[2]

[1]Institute of Social Sciences, University of Lausanne
[2]Department of Methodology, London School of Economics and Political Science
[3]Question Testing Group, Verian Group

Working paper
July 2026

Abstract

The Question Appraisal System (QAS-99) is a widely used protocol for the systematic assessment of survey questions, requiring evaluators to make 27 binary coding decisions across eight appraisal steps. QAS coding is slow and resource-intensive, creating the possibility that large language models (LLMs) can replace or supplement human evaluators. We compared LLM evaluations of 118 draft survey questions with those of a QDET expert and novice coders. Agreement was further assessed using additional human evaluators and a second LLM on a validation subset. We find that LLMs produced actionable QAS evaluations at scale, with agreement levels broadly comparable to those observed between novice and expert human evaluators. LLMs identified fewer problems than novices, who in turn identified fewer problems than experts, suggesting that expertise currently remains important for questionnaire review. Agreement varied substantially across QAS decisions, indicating that some question characteristics are inherently more difficult to evaluate than others. Compared with the expert coder, the LLM was notably less likely to identify certain data quality risks, particularly those related to social desirability bias. Nevertheless, the findings suggest that LLMs can support QAS-based survey appraisal at scale and large-scale research on questionnaire design and survey measurement error.

*Corresponding author: caroline.roberts@unil.ch

1 Introduction

A substantial literature documents the consequences of poor question design for response quality and argues for a theoretically and empirically informed approach to questionnaire design, evaluation and testing (QDET) (Presser et al., 2004; Campanelli, 2008; Groves et al., 2009). Yet investment in reducing measurement error consistently lags behind efforts to reduce representation error (Fowler, 1995), held back by persistent disagreement about which QDET methods are most effective (Presser and Blair, 1994; Maitland and Presser, 2018; Willis, 2020), and by the time and cost of preferred approaches such as cognitive interviewing and split-ballot experiments. Providing the survey research community with faster, more cost-efficient, and reliable ways to evaluate draft questions would help lower these barriers and contribute to a general push towards improving the accuracy of survey measurement. Recent advances in large language models (LLMs) now offer the technology to enable exactly such a development (Jansen et al., 2023; Olivos and Liu, 2025; Adhikari et al., 2025). In this paper, we implement and evaluate one concrete use case: the application of LLMs to the systematic assessment of potential risks to measurement quality using the Question Appraisal System-99 (QAS-99), a widely adopted rule-based QDET framework (Willis and Lessler, 1999).

Rule-based QDET methods apply structured checklists of question characteristics associated with poor measurement quality, derived from cognitive models of the survey response process (Cannell et al., 1981; Tourangeau, 1984; Tourangeau and Rasinski, 1988) and accumulated design recommendations (Fowler, 1995; Saris and Gallhofer, 2007; Krosnick, 2011; Dillman et al., 2014). The approach treats question quality as determined by objective, identifiable features such as wording, the type of information requested, and response format that can be evaluated without collecting respondent data (Dykema et al., 2020; Willis, 2020). Multiple rule-based frameworks have been developed on this basis (Forsyth et al., 2004; Graesser et al., 2000; Felderer et al., 2024; d'Ardenne et al., 2025; Fitzgerald et al., 2011), of which the QAS-99 is among the most widely used. Developed at the Research Triangle Institute (Willis and Lessler, 1999; Forsyth and Hubbard, 1992; Lessler and Forsyth, 1996), the instrument comprises 27 coding decisions across eight evaluation steps, covering risks to measurement quality arising from problems with reading and comprehension, instruction clarity, question wording and terminology, assumptions about respondent knowledge and behaviour, cognitive demands (recall, computation), sensitivity, response category design, and other miscellaneous features. Designed as an 'easy-to-use' tool for practitioners with varying levels of expertise (Willis and Lessler, 1999, p. 1-1), the QAS-99 provides a transparent basis for identifying

and documenting question problems at relatively low cost(Dean et al., 2007; Schaad et al., 2020).

Evidence on the effectiveness of rule-based QDET methods is broadly positive. Comparative studies find them effective at identifying problems predictive of field difficulties and reduced data quality, including those associated with complex wording, demanding recall tasks, sensitive content, and poorly designed response categories (Forsyth et al., 2004; Olson, 2010; Dykema et al., 2020). Rothgeb et al. (2007) found the QAS detected more problems overall than either informal expert review or cognitive interviewing when applied to the same questionnaires. Maitland and Presser (2016; 2018; 2020) found rule-based methods particularly effective at identifying respondent task problems, though somewhat less effective than expert review for problems affecting comprehension and of less predictive value for field outcomes compared to expert review and cognitive interviews.

Reliability of QAS completion is a more persistent concern. Assessments of intercoder agreement in rule-based applications are relatively rare (Bais et al., 2019), and few studies have formally tested the reliability of the QAS-99 itself. Schaad, Jans and Scott (2020) assessed inter-rater agreement among three evaluators with varying levels of expertise using an extended version of the QAS applied to ten questions. While agreement on whether any problem was present was high, agreement on which specific problem types were identified was substantially lower and senior evaluators identified more problems than did the junior rater. Across the original QAS-99 steps, agreement was reasonable (in the 0.7–0.9 range) but fell below chance-level for Step 2 (Instructions) and Step 7 (Response Categories). Bais et al. (2019) found similar patterns in a more extensive reliability assessment of a related coding framework, with agreement particularly low for characteristics considered subjective or difficult to operationalise clearly. These findings echo a broader literature showing considerable disagreement among expert questionnaire reviewers (Schaad et al., 2020; Presser and Blair, 1994; DeMaio and Landreth, 2004; Conrad and Blair, 2004; Olson, 2010), though the general consensus is that moderate reliability does not undermine the utility of rule-based QDET methods in detecting and fixing problems.

Various explanations for this low reliability have been proposed, including the cognitive burden of applying detailed coding schemes to large question sets, inherent subjectivity in judging ambiguous characteristics, and variation in evaluators' critical thresholds, which typically tracks their level of QDET expertise but may also reflect individual idiosyncrasy (Schaad et al., 2020; Schouten et al., 2018; Bais et al., 2019; Olson, 2010). A further factor is the lack of consensus about the relative severity of the risks to measurement quality that different question characteristics present, because the 27 QAS

coding decisions do not carry equal weight. Some steps target characteristics with well-established empirical links to measurement error, such as vague wording, demanding recall or computation tasks, sensitive content, double-barreled questions, and flawed response categories (Dykema et al., 2020; Bais et al., 2019; Maitland and Presser, 2020). Others flag features whose impact is more context-dependent and harder to judge. This variation in impact severity matters for evaluating intercoder reliability and for assessing automated systems because a classifier that reliably identifies high-impact risks is of greater practical utility than one with nominally higher overall agreement that disproportionately fails on the problems that matter most. In our analysis, we accordingly stratify comparisons of LLM and human evaluations by the type of impact risk associated (theoretically) with each QAS step.

Recent advances in LLMs have prompted growing interest in their application to survey research tasks (Buskirk et al., 2025b; Rothschild et al., 2026). Their ability to process and classify text features (Kostina et al., 2025), apply domain knowledge to structured evaluation tasks (Zhang, 2025), and produce detailed outputs at scale and speed makes them natural candidates for automating rule-based QDET procedures. A recent systematic review (Buskirk et al., 2025b) identified a rapidly expanding body of work applying LLMs across the survey lifecycle, including a growing subset focused on question evaluation and pretesting (Maiorino et al., 2023; Olivos and Liu, 2025; Buskirk et al., 2025b; Adhikari et al., 2025; Tao et al., 2024; Olson and Buskirk, 2026; Sturgis et al., 2026; Fuchs et al., 2026). Most of this work does not assess LLM performance using a formal rule-based coding framework, though some conclusions are noteworthy.

Olson and Buskirk (2026) evaluated the ability of multiple LLMs — including editions of ChatGPT and Claude Sonnet — to calculate readability scores for 60 survey questions using five established readability measures, finding considerable between-model and within-model variation over time, and systematic calculation errors with implications for score accuracy.Olivos and Liu (2025) compared feedback generated by GPT-4 on question comprehensibility with expert judgements from Lenzner (2012), focusing on seven characteristics associated with comprehension difficulties, including low-frequency words, vague or ambiguous terms, and complex syntax. GPT-4 was effective at identifying problems and suggesting improvements, though its suggestions did not always coincide with Lenzner's, highlighting the continued importance of 'researchers' judgement calls' (p. 12). Buskirk et al. (2025a) examined how prompt phrasing affects the structural and linguistic quality of survey items generated by ChatGPT 3.5, confirming that providing LLM's with more explicit instructions in prompts consistently

produced better-structured, less ambiguous, and more readable questionnaire items — with quality assessed against Dillman et al. (2014) design criteria and QUAID (Graesser et al., 2000). Tao et al. (2024) evaluated five prompt variants for detecting flaws in 66 survey questions using Gemini 1.5 Pro, comparing model performance against expert ratings. Including best practice recommendations and examples of both flawed and flawless questions improved performance, while aggregating across multiple replications of LLM judgements improved precision and reduced over-identification of problems.

The studies most directly comparable to the present work are those that apply a structured coding framework to evaluate question quality. Metheney and Yehle (2025) applied a truncated version of the QAS to 262 survey questions using GPT-3.5 and GPT-4. Comparing model output to a single human expert, overall agreement was approximately 80%. The LLM flagged problems the human did not in around 15% of cases, particularly for unfair presumptions and undefined reference periods, while the human identified problems the LLM missed in around 5% of cases. Timbrook et al. (2026) explored the accuracy of ChatGPT 5 in applying ten research-driven indicators of high-quality question construction to a sample of 100 questions, comparing evaluations generated by a prompt with problem examples (few-shot learning) and a prompt without examples (zero-shot learning) to human coding. To our knowledge, no published study has evaluated LLM performance on the full QAS-99 against the complete range of human QDET expertise, from novice to experienced evaluators.

We develop, implement and evaluate a pipeline for applying LLMs to the QAS-99 at scale, and compare LLM evaluations with those of human QDET experts and novice evaluators on a corpus of 118 draft survey questions. Going beyond aggregate agreement rates, we also examine whether LLMs identify certain types of data quality risk more reliably than others, stratifying our comparisons in terms of the relative impact the problems identified at each QAS step might have on measurement quality. We thereby contribute both to the rapidly developing body of research on GenAI applications in survey workflows, and to the longer-standing literature on the performance and reliability of rule-based QDET tools. Specifically, our analysis addresses the following questions:

- **RQ1:** Do LLMs agree with human experts and novices in their application of the QAS-99 to evaluate draft survey questions?
- **RQ2:** Does disagreement between LLMs and human experts and novices concentrate in particular QAS steps or decisions?
- **RQ3:** Can LLMs identify question problems that matter most for data quality?

2 Methods

2.1 Research design

To assess LLM performance of the QAS task, we use a range of evaluation metrics, focusing on levels of agreement across the different QAS steps between and within groups of LLM, expert and novice evaluators. As the QAS-99 was designed to support question evaluation by survey practitioners with less QDET experience to reduce reliance on expert review (Schaad et al., 2020), we also compare its application by LLMs to both human experts and novices in questionnaire design.

Our basic design prompted an LLM to undertake QAS-99 evaluations of a corpus of 118 draft survey questions (written in English), for which novice evaluations using the QAS were already available. The questions were designed and evaluated by graduate Questionnaire Design students as part of their MA coursework (a total of 76 coders, each reviewing up to two of their own questions as part of a pre-testing assignment). The questions were designed for use in a paper self-completion questionnaire, to collect a mix of factual and nonfactual information covering a broad range of different topics. A QDET expert completed evaluations of the same question set. A subset of 25% of these (n=30) were evaluated by two other experts [1], a second novice coder (a research assistant with no prior training in questionnaire design and evaluation) and by a second LLM for validation purposes. For simplicity, we refer to the original student novices (who evaluated their own questions) collectively as Novice 1 throughout.

2.2 Pipeline and prompt design

The pipeline consisted of a Python script (programmed in Jupyter Notebooks on Anaconda) including processes to relay the input database of draft questions and their response categories (an .xlsx file or .csv), together with the user prompt (stored in a text file) to the LLM via the API. For this purpose, we used OpenAI's GPT-4, accessed via the OpenAI API. The decision to use GPT-4 was mainly pragmatic. At the time, it was one of the most widely-used LLMs (accessed via ChatGPT), and had demonstrated ability to match human abilities on a number of advanced benchmarks and to outperform other LLM's on a wide range of NLP tasks (OpenAI, 2023).

Once the QAS evaluation is complete, the data is output to an .xlsx file, structured to match the human data, which were captured in an online

[1] Expert evaluations for this paper were completed by three of the authors: Caroline Roberts, Alice McGee and Rebecca Hamlyn, who have each been involved in questionnaire design and evaluation for over 20 years and have extensive experience of applying the QAS, or similar question evaluation tools, as well as cognitive testing.

questionnaire programmed in Qualtrics. The form was designed to match the original QAS-99 pencil-and-paper version (adapted from Willis and Lessler 1999), consisting of yes/no answers for each of the 27 QAS steps, and text entry fields for recording details of problems identified). Some minor modifications to wording were made to enhance usability. These also concerned the formulation of the items at Step 1 (Reading), which was originally designed for assessing questions to be used in interviewer-based surveys, so were adapted to make them relevant to challenges respondents might have reading questions themselves. Some of the original novice evaluations used a paper-and-pencil version of the QAS-99 and were entered manually in the Qualtrics version. Data from all three sets of coders were combined in a single database for analysis.

The design of the initial system and user prompts used in the pipeline were also based on the original QAS-99 form, using the same formulations used at each step. In addition, some contextual information about the QAS-99 and the cognitive steps involved in answering survey questions was included, together with instructions for coding taken directly from the Willis and Lessler (1999) manual (but without explicit examples of question problems). The prompt also specified the format of the output data so it could be merged directly with the human data. This requirement provided a clear task specification - to provide a yes/ no response for each of the 27 decisions in the QAS, and for each 'Yes' answer, to provide a short summary of the problem identified.

The prompt was iteratively revised with the aim of improving the clarity of the coding task (e.g., rewording the original QAS coding decisions in the form of direct questions specifying more explicitly what to look for in questions) and encouraging more critical evaluations. Again, no explicit examples of question problems were provided and no other parameter (e.g., model temperature) was adjusted. Evaluations of the full 118 questions presented here for GPT-4 were generated using this revised prompt (shown in Appendix) in September 2025. Because of the speed with which LLM performance and capabilities are advancing, and to provide a comparison point with the main GPT-4 evaluations, the same prompt was used to generate new evaluations of the 30 questions in the validation set in May 2026 using a second LLM, Anthropic's Claude Sonnet 4.6 (accessed directly via Claude Code and not via the API, but with the instruction to use the same prompt and input file as in the pipeline).

2.3 Analytic approach

We focus on agreement rates between coder pairs (the percentage of questions in the corpus for which the coding decision concurred (i.e., a problem was

flagged or not flagged by both coders)) for 1) all 27 decisions; 2) each of the eight steps; and 3) overall. Specifically, we compare LLM 1 (GPT-4), Expert 1 and the Novices (Novice 1) for the full corpus of 118 questions. We then compare these with agreement rates for the validation subset of 30 questions between a) LLM 1 and LLM 2 and the three experts (individually and with the average of their ratings); b) the LLMs, novice 1 and novice 2 (individually and on average); and c) GPT-4 and Claude Sonnet 4.6. In the same way, we also compare agreement between and within the experts and the novices to contextualise the comparisons between LLMs and humans. As a complement, we also calculate Gwet's AC1 (Gwet, 2002, 2008) measure of agreement (using the irrCAC package in R), which estimates and adjusts for the probability that two raters would agree purely by chance using the average proportion assigned to each category across both raters. This option is preferred over more commonly used metrics such as Cohen's kappa, which are unstable when base rates are low (i.e., at QAS steps where problems are rarely flagged) (Wongpakaran et al., 2013).

To assess the nature of disagreements between LLMs and human coders, and the extent to which they vary across QAS steps (RQ2), we first consider coder evaluations at the aggregate level, computing for all questions and coders the total number of problems flagged across all 27 QAS steps, calculating means across all questions and the proportion of questions for which problems were flagged. We then consider variation in flagging rates across steps, considering the mean number of problems flagged as a proportion of the possible number of problems (which varies by step) and the extent to which coders agreed on the relative prevalence of problem types. We also examine which coding decisions contribute most to the total number of disagreements between coder types at each of the eight steps (comparing LLMs vs Experts, LLMs vs. Novices, Experts vs. Novices and Experts vs. Experts). This makes it possible to examine the extent to which reduced reliability of the QAS-99 results from varying levels of QDET expertise among coders or whether particular coding decisions may be inherently more ambiguous (or less clearly operationalised) than others.

Finally, to address RQ3, we focus on the direction of disagreement between coders, comparing GPT-4 and Novice 1 against Expert 1 for the full question set. Based on the literature reviewed, evaluations between experts were expected to differ, in part due to the subjective nature of question evaluation and perceptions of problem severity (Willis and Lessler, 1999). It is not appropriate, therefore, to deem one set of evaluations to be ground truth. Nevertheless, comparing novices and LLMs to expert review is meaningful when considering ways to make questionnaire evaluation procedures more cost-efficient. Thus, for all 27 QAS decisions, we compare

LLM 1 and Novice 1 'hit rates' (the percentage of problems flagged by Expert 1 that they also flagged) and unconfirmed flag rates (the percentage of problems flagged by the coders that were not flagged by Expert 1 – akin to false positives/discovery - were Expert 1's ratings indeed considered to be the gold standard).

Subjective perceptions of problem severity, as well as objective knowledge of likely data quality impacts, play a role in coding decisions, however. A coder may observe a potential problem but dismiss it on account of perceived low importance. For this reason, our analysis takes account of potential impact severity by classifying the QAS decisions into different types of data quality risk and examining problem flagging rates by coder type for each category. This classification (summarised in Table 1) was informed by existing literature (e.g., Krosnick, 1991; Tourangeau et al., 2020; Presser et al., 2004; Willis, 2005; Olson, 2010) as well as theoretical assumptions about the mechanisms by which effects on data quality might result. It distinguishes particular manifestations of survey error (missing data, measurement bias and variance, and non-validity), which have different implications for the accuracy of statistics and the analytic utility of the data. Note, however, that the extent of the risk presented by different problem types may vary by mode of questionnaire administration (d'Ardenne et al., 2025); we do not consider this explicitly here.

Table 1: QAS-99 problem types classified by type of data quality risk.

No.	Step	Problem type	Mechanism
Risk of item-nonresponse			
1	2a	Conflicting instructions	Increased burden; respondents may be unsure which instructions to follow, misunderstand the questions, or be unsure how to answer.
2	2b	Complicated instructions	Increased burden; respondents may ignore the instructions, not read them carefully, or apply them incorrectly.
3	3a	Complex wording/syntax	Increased burden; respondents may not read the question carefully or may misinterpret the question.
4	4a	Inappropriate assumptions	Respondents may not know how to answer, or feel the question does not apply to them.
5	7a	Open-ended question	Increased burden; respondent must reflect on and articulate response in own words.
Risk of measurement bias			
6	5c	Recall failure	Increased burden on memory may result in recall failures, forward or backward telescoping.
7	5d	Computation problem	Increased burden; complex calculations may result in 'guess-timates', rounding and heaping.
8	6a	Sensitive content	Threat of disclosure; respondent may be unwilling to share personal information.
9	6b	Sensitive wording	Even for non-sensitive topics, respondents may feel pressure to present themselves in a more favourable light.
10	6c	Socially acceptable response implied	Salient norms around what is socially desirable may influence willingness to report answers truthfully.
Risk of increased measurement variance (low reliability and validity)			
11	1a	What to read	Respondents may overlook relevant information or read only part of the item, leading to inconsistent interpretations.
12	1b	Missing information	Respondents may rely on assumptions, resulting in inconsistent interpretations.
13	1c	How to read	Questions may be misread, resulting in inconsistent interpretations.
14	3b	Undefined technical terms	Respondents may be unfamiliar with the term, or may interpret concepts differently.
15	3c	Vague wording	Respondents may interpret concepts differently, resulting in inconsistent interpretations.
16	3d	Poor/missing reference periods	Inconsistent temporal interpretations across respondents may result in variation in how the response is constructed.
17	4b	Assumes stable behaviour	Respondents may find it difficult to provide a single answer that accurately reflects experience, resulting in inconsistent interpretations.
18	7c	Technical terms in response categories	Respondents may struggle to map their answer onto the available categories, resulting in inconsistent responding.
19	7d	Vague response categories	Respondents may interpret response category labels differently or in multiple ways, resulting in inconsistent or random response selection.
20	7e	Overlapping response categories	Respondents may struggle to distinguish available options or to know which to select, resulting in inconsistent responding.

3 Results

3.1 Agreement rates between LLMs, human experts and novices (RQ1)

We first assess the overall extent of agreement between LLM 1 (GPT-4), Novice 1 and Expert 1 for the complete question set (n=118), across all 27 coding decisions combined, and aggregating across the eight QAS steps (first three rows of Table 2 and top panel of Figure 1). Agreement rates for all individual coding decisions are shown in Appendix Table 1. The average percentage agreement was highest between the LLM and Novice 1 (74% agreement) and lowest between the LLM and Expert 1 (63%) (see left side of top panel of Figure 1). The rate for Expert 1 and Novice 1 was 64 percent. Across the eight steps, the average percentage agreement between LLM 1 and Expert 1 were lowest at 52% for Step 3 (question clarity) and highest at 75% for Step 2 (instructions). Agreement between LLM1 and Novice 1 was higher overall than with Expert 1, ranging from 60% for Step 3 to 82% for Step 2. These compare with agreement rates between Expert 1 and Novice 1 of 59% at Step 3 and 69% at Step 2.

Agreement rates in the validation question set (n=30) broadly mirrored those in the main set. Averaged across coders within groups, LLMs agreed with Experts at 69% and with Novices at 71%, compared with 65% between Experts and Novices. Step 3 (clarity) again showed the lowest agreement across coder pairs, while inter-LLM agreement was highest overall (83%). Within coder types, there was greater individual variation than in the main set, particularly among the novices, though group-level patterns across steps closely mirrored the main set findings (lower panel of Figure 1).

The average Gwet's AC1 values across all coding decisions for all coder pairs is shown in column 3 of Table 2 (see Appendix Table 2 for AC1 at each coding step for all coder pairs). Gwet's AC1 can be interpreted similarly to Cohen's kappa. The overall values are relatively low in the main question set indicating only a 'fair' level of agreement between LLM 1 and Expert 1 (AC1 = 0.361) and Novice 1 and Expert 1 (AC1 = 0.376). By contrast, the level of agreement between LLM 1 and Novice 1 was more substantial (AC1 = 0.620). In the validation set, average agreement between the LLMs and the Experts was low to moderate (AC1 = 0.469 for LLM 1 and 0.461 for LLM 2) and between the LLMs and the Novices was substantial (AC1 = 0.538). By comparison, the lowest level of agreement between coder types was found between the Novices and the Experts (AC1 = 0.394). Within coder types, LLM 1 and LLM 2 had the highest level of agreement (AC1 = 0.733), while average agreement rates between the Experts and the Novices were only moderate (AC1 = 0.509 and 0.470 respectively). Thus, while

Table 2: Percentage agreement and Gwet's AC1 between coder pairs across QAS steps.

Coder pair	All QAS steps		% agreement by step							
	% agree	AC1	Step 1 Reading (n=3)	Step 2 Instruc. (n=2)	Step 3 Clarity (n=4)	Step 4 Assumpt. (n=3)	Step 5 Knowl. (n=4)	Step 6 Sensitiv. (n=3)	Step 7 Resp. cats. (n=7)	Step 8 Other (n=1)
All questions (N=118)										
LLM 1 – Expert 1	62.9	0.361	62.7	75.0	51.9	55.9	70.1	59.6	66.9	55.9
LLM 1 – Novice 1	73.5	0.620	72.6	81.8	60.2	70.3	72.5	73.4	80.1	80.5
Expert 1 – Novice 1	64.0	0.376	57.1	68.6	58.7	59.9	71.0	67.5	67.2	48.3
Validation set (N=30): LLM 1 vs Experts										
LLM 1 – Expert 1	66.0	0.351	64.4	80.0	54.2	63.3	72.5	68.9	69.5	40.0
LLM 1 – Expert 2	73.3	0.545	70.0	83.3	60.8	67.8	81.7	77.8	71.0	100.0
LLM 1 – Expert 3	66.9	0.380	77.8	86.7	54.2	60.0	70.8	60.0	64.3	90.0
LLM 1 – Mean of experts	**68.8**	**0.425**	**70.7**	**83.3**	**56.4**	**63.7**	**75.0**	**68.9**	**68.3**	**76.7**
Validation set (N=30): LLM 2 vs Experts										
LLM 2 – Expert 1	65.8	0.404	62.2	70.0	58.3	54.4	69.2	73.3	73.8	40.0
LLM 2 – Expert 2	73.6	0.553	76.7	86.7	63.3	67.8	75.0	71.1	73.3	100.0
LLM 2 – Expert 3	67.2	0.425	77.8	80.0	55.0	57.8	69.2	64.4	66.7	90.0
LLM 2 – Mean of experts	**68.8**	**0.461**	**72.2**	**78.9**	**58.9**	**60.0**	**71.1**	**69.6**	**71.3**	**76.7**
Validation set (N=30): LLMs vs Novices										
LLM 1 – Novice 1	72.1	0.607	71.1	73.3	60.0	68.9	70.0	74.4	80.0	76.7
LLM 1 – Novice 2	71.1	0.581	74.4	85.0	49.2	72.2	78.3	66.7	74.8	76.7
LLM 2 – Novice 1	70.5	0.521	78.5	70.0	55.8	64.4	68.3	74.4	77.0	76.7
LLM 2 – Novice 2	69.6	0.512	74.4	78.3	56.7	72.2	73.3	62.2	71.4	76.7
Mean LLMs – Mean novices	**70.8**	**0.555**	**74.6**	**76.7**	**55.4**	**69.4**	**72.5**	**69.4**	**75.8**	**76.7**
Validation set (N=30): Novices vs Experts										
Novice 1 – Expert 1	64.8	0.379	53.3	70.0	59.2	63.3	75.8	67.8	68.6	36.7
Novice 1 – Expert 2	69.8	0.476	73.3	76.7	55.8	67.8	75.0	65.6	72.9	76.7
Novice 1 – Expert 3	66.7	0.396	70.0	80.0	60.8	62.2	72.5	58.9	65.7	73.3
Novice 2 – Expert 1	58.6	0.257	56.7	71.7	53.3	55.6	65.8	53.3	60.5	43.3
Novice 2 – Expert 2	64.2	0.453	82.2	78.3	40.0	62.2	70.0	57.8	64.8	76.7
Novice 2 – Expert 3	63.7	0.379	83.3	85.0	56.7	50.0	60.8	55.6	62.9	73.3
Mean novices – Mean experts	**64.6**	**0.390**	**69.8**	**76.9**	**54.3**	**60.2**	**70.0**	**59.8**	**65.9**	**63.3**
Validation set (N=30): Experts vs Experts										
Expert 1 – Expert 2	73.0	0.408	67.8	70.0	68.3	66.7	80.8	80.0	78.6	40.0
Expert 1 – Expert 3	71.0	0.384	62.2	73.3	73.3	74.4	80.0	71.1	70.0	43.3
Expert 2 – Expert 3	71.4	0.529	76.7	86.7	56.7	63.3	74.2	71.1	72.4	90.0
Mean within experts	**71.8**	**0.441**	**68.9**	**76.7**	**66.1**	**68.1**	**78.3**	**74.1**	**73.7**	**57.8**
Validation set (N=30): Within-group comparisons										
LLM 1 – LLM 2	82.5	0.733	84.4	83.3	74.2	73.3	83.3	77.8	89.0	100.0
Novice 1 – Novice 2	67.2	0.496	77.8	75.0	52.5	70.0	63.3	58.9	73.3	66.7

Note: % agree = mean percentage agreement across QAS coding decisions within each step. AC1 = Gwet's AC1 averaged across all 27 coding decisions. Bold rows are unweighted group means. LLM 1 = GPT-4; LLM 2 = Claude Sonnet 4.6.

average percentage agreement at QAS steps may appear reasonable, low AC1 values suggest that within QAS steps, specific coding decisions are discordant between coders and coder types. In the following, we examine in more detail patterns of flagging by coder type across QAS steps, and the nature of disagreements between coders.

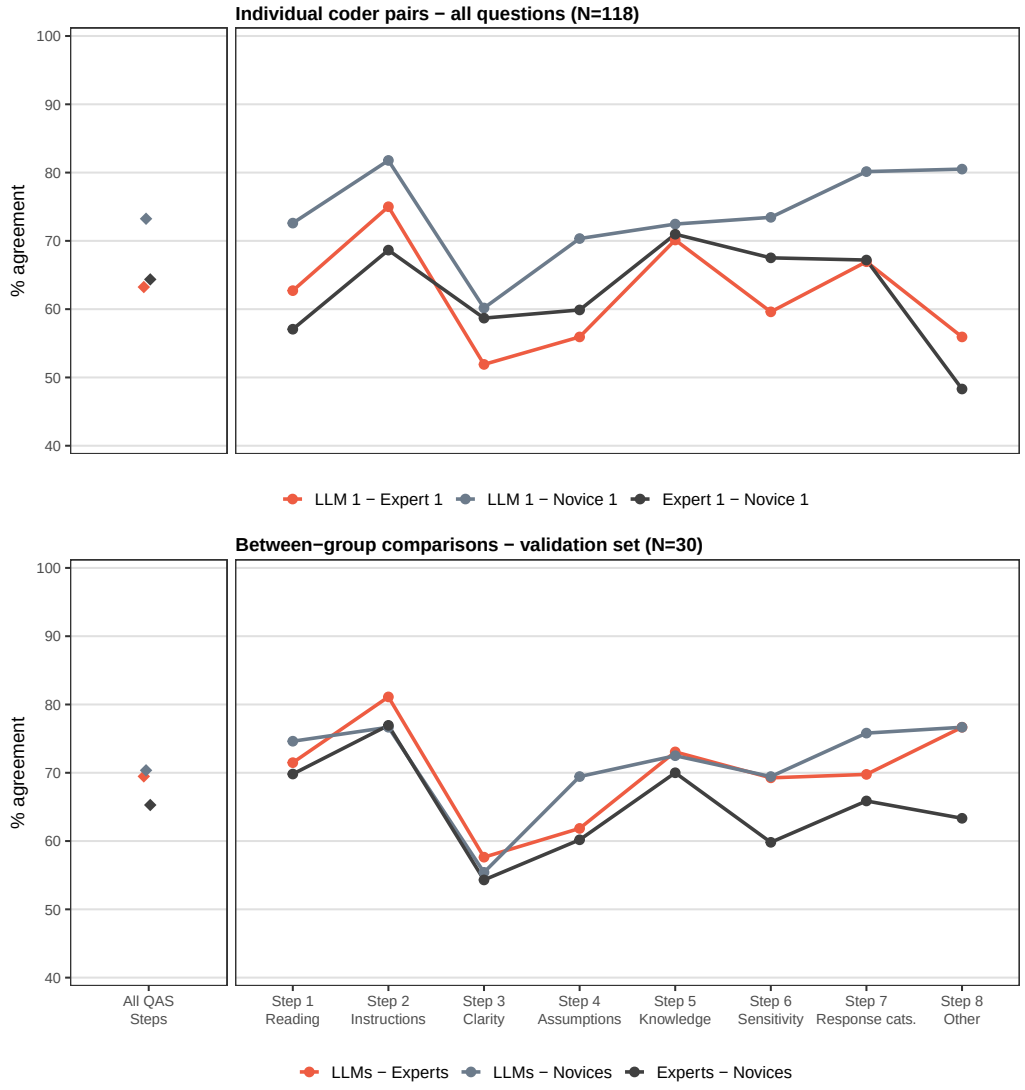


Figure 1: Agreement between individual coder pairs (full corpus, n=118) N=118) and within- and between-group comparisons (validation set, N=30).

3.2 Nature of disagreements across QAS coding decisions (RQ2)

Overall agreement rates only indicate whether or not decisions concurred (flag or no flag), but say little about the extent to which coder pairs agree on the presence of problems for a given question and their nature. In the

main question set, LLM 1 (GPT-4) flagged a mean of 3.5 problems across all 27 QAS decisions, which was a little over half the number flagged by the novices on average (6.3 problems). By contrast, Expert 1 flagged twice as many problems (12.1) on average than the novices (see Figure 2). A similar pattern was observed in the validation question set: the mean number of problems flagged was lowest for the LLMs (M=5.2) and highest for the experts (M=9.7), with novices (M=6.5) falling between the two. Within the LLM group, Claude flagged 1.9 more problems on average (M=6.1) than GPT-4 (M=4.3). Problem counts by question for the validation set are shown in Appendix Table 3.

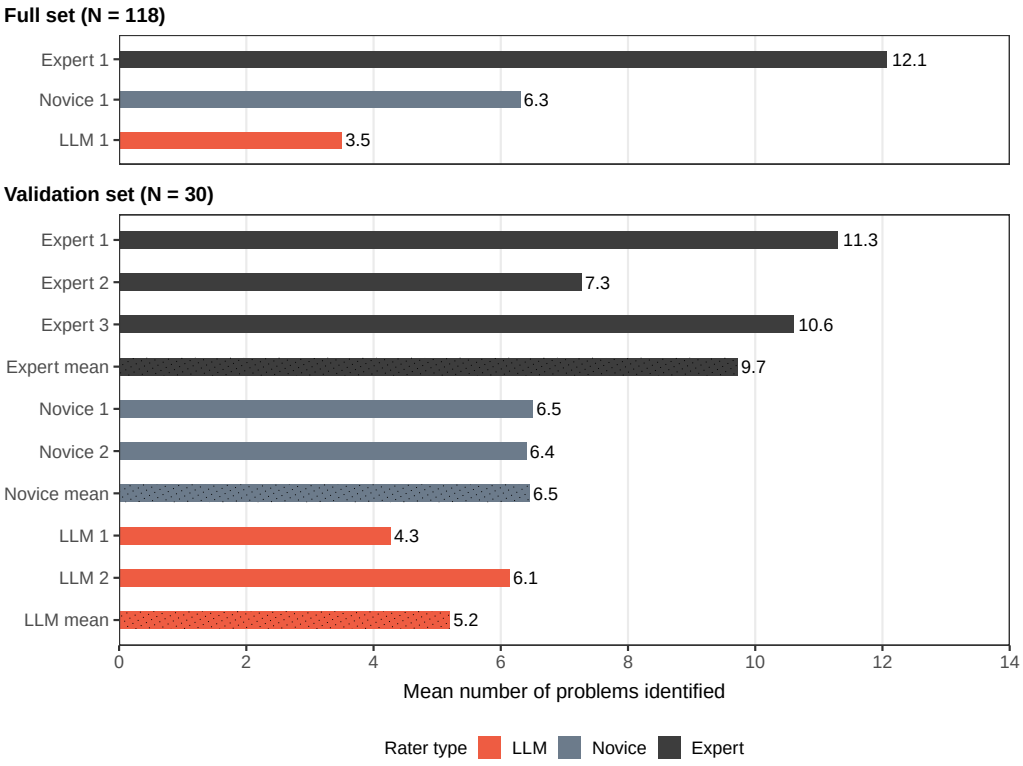


Figure 2: Mean number of problems flagged by coder and coder type.

Because the number of coding decisions varies at each step (e.g., from seven at Step 7 to just one at Step 8), raw step means are not directly comparable. Expressing each step mean as a proportion of the number of possible problems that could be flagged (shown in Table 3) confirms (in the main set) that LLM 1 (GPT-4) is consistently the most conservative overall, under-flagging problems relative to Expert 1 and Novice 1 across almost every step, with the exception of Step 1 (Reading), where its flagging rate (19%) slightly exceeds that of Novice 1 (12%), compared with 46% for Expert 1. LLM 1 evaluations come closest to Novice 1 at Step 3 (Question Clarity), with flagging rates of 31% and 42% respectively (compared with

69% for Expert 1). The model diverges most from human coders at Step 2 (Instructions), where it flagged almost no problems (a rate of 5%, compared with 14% for Novice 1 and 29% for Expert 1); at Step 6 (Sensitivity), where its flagging rate was 8% (compared with 28% for Novice 1 and 46% for Expert 1); and at Step 7 (Response Categories), where its rate was 7% (compared with 16% for Novice 1 and 37% for Expert 1).

In the validation question set, the relative pattern across steps was broadly consistent with the main set findings. The main divergence occurred at Step 5 (knowledge/recall), where LLMs flagged far fewer problems (LLM 1: 8%; LLM 2: 18%) compared with novices (18–30%) and experts (20–31%), suggesting that recall and knowledge demands may be harder for an LLM to assess from text alone.

Table 4 presents the percentage of questions evaluated in each set where at least one problem was flagged at each QAS step, across coder types. Again, the proportion of questions with flagged problems varies by coder type and across steps. In the main set, LLM 1 identified problems for the fewest number of questions, except at Step 1 where it flagged reading-related problems for 36% of the questions, compared with the 32% of questions flagged by the novices and 75% questions by Expert 1. LLM 1 was also least likely to identify questions with problems at Steps 2 (instructions) and step 6 (sensitivity), where just 7% and 22% of questions were flagged. By contrast, Expert 1 consistently flagged more questions as having problems at each step than both the LLM and the Novices. The average differences in the proportion of questions with problems flagged across all steps between the LLM and Expert 1 was 43%, and between the LLM and the novices was 20%, compared with 24% between Expert 1 and the Novices (not shown in table).

The relative pattern of flagging across steps was broadly consistent across coder types: Question Clarity (Step 3) was the most frequently flagged step for all three coders, and Instructions (Step 2) and Other (Step 8) were consistently the least flagged. Rank-order correlations between coders confirm this broad agreement (LLM 1 vs Expert 1: $r = 0.81$; LLM 1 vs Novice 1: $r = 0.71$; Novice 1 vs Expert 1: $r = 0.93$; see Appendix Table 4 for full correlation matrices for all coders in both question sets). The main exception is Step 7 (Response Categories), which ranked second for both Expert 1 (94% questions flagged) and Novice 1 (68%), but fifth for LLM 1 (28%), suggesting LLM 1 is less attentive to response category problems than human coders. In the validation set, LLM 2 showed substantially higher rank-order correlations with human coders (0.84–0.98) than LLM 1 (0.41–0.72), suggesting greater consistency between Claude and human evaluators in their ordering of which QAS steps presented the most problems.

Table 3: Mean flagging rate per QAS step by coder and coder group.

Group	Coder	Flagging rate (% of possible problems flagged)								
		All steps (n=27)	Step 1 Reading (n=3)	Step 2 Instruc. (n=2)	Step 3 Clarity (n=4)	Step 4 Assumpt. (n=3)	Step 5 Knowl. (n=4)	Step 6 Sensitiv. (n=3)	Step 7 Resp. cats. (n=7)	Step 8 Other (n=1)
All questions (N=118)	LLM 1	13.0	19.2	4.7	30.5	13.8	11.4	7.9	7.1	0.0
	Novice 1	23.7	12.7	14.7	42.5	23.7	28.2	28.2	16.7	16.7
	Expert 1	44.7	45.8	28.8	69.3	51.1	35.8	45.5	36.8	44.1
Validation set (N=30)	LLM 1	14.7	25.0	9.2	33.3	13.3	9.2	10.0	9.0	0.0
	LLM 2	22.7	20.0	20.0	35.8	20.0	18.3	31.1	20.5	0.0
	LLM mean	**18.7**	**22.5**	**14.6**	**34.5**	**16.6**	**13.8**	**20.6**	**14.8**	**0.0**
	Novice 1	23.7	12.5	15.0	43.1	24.4	27.1	25.0	16.7	26.7
	Novice 2	21.7	3.9	7.5	51.7	20.6	16.7	26.7	17.4	21.7
	Novice mean	**22.7**	**8.2**	**11.2**	**47.4**	**22.5**	**21.9**	**25.9**	**17.0**	**24.2**
	Expert 1	45.2	51.7	33.3	68.8	51.1	31.7	42.8	38.3	46.7
	Expert 2	26.9	14.4	20.0	42.5	30.0	20.0	26.7	31.9	0.0
	Expert 3	39.1	17.5	16.9	63.6	44.1	34.3	46.3	42.1	11.8
	Expert mean	**37.1**	**27.9**	**23.4**	**58.3**	**41.7**	**28.7**	**38.6**	**37.4**	**19.5**

Note: Flagging rate = mean percentage of of possible problems flagged at each step. Bold rows are unweighted group means. Step n = number of QAS coding decisions. LLM 1 = GPT-4o; LLM 2 = Claude 3.5 Sonnet.

Table 4: Percentage of questions with at least one problem flagged per QAS step, by coder.

Group	Coder	All steps (n=27)	Step 1 Reading (n=3)	Step 2 Instruc. (n=2)	Step 3 Clarity (n=4)	Step 4 Assumpt. (n=3)	Step 5 Knowl. (n=4)	Step 6 Sensitiv. (n=3)	Step 7 Resp. cats. (n=7)	Step 8 Other (n=1)
All questions (N=118)	LLM 1	94.9	35.6	6.8	81.4	39.8	31.4	22.0	28.0	0.0
	Novice 1	100.0	32.2	26.3	93.2	53.4	61.9	50.8	67.8	16.7
	Expert 1	100.0	74.6	42.4	98.3	85.6	80.5	72.9	94.1	44.1
Validation set (N=30)	LLM 1	96.7	40.0	23.3	86.7	36.7	26.7	23.3	36.7	0.0
	LLM 2	93.3	40.0	36.7	86.7	40.0	53.3	50.0	56.7	0.0
	LLM mean	**95.0**	**40.0**	**30.0**	**86.7**	**38.3**	**40.0**	**36.7**	**46.7**	**0.0**
	Novice 1	100.0	26.7	26.7	93.3	63.3	63.3	46.7	73.3	23.3
	Novice 2	100.0	16.7	13.3	96.7	50.0	56.7	63.3	70.0	23.3
	Novice mean	**100.0**	**21.7**	**20.0**	**95.0**	**56.7**	**60.0**	**55.0**	**71.7**	**23.3**
	Expert 1	100.0	70.0	46.7	100.0	73.3	76.7	70.0	93.3	60.0
	Expert 2	100.0	36.7	40.0	83.3	66.7	53.3	46.7	96.7	0.0
	Expert 3	100.0	36.7	30.0	100.0	80.0	83.3	83.3	100.0	10.0
	Expert mean	**100.0**	**47.8**	**38.9**	**94.4**	**73.3**	**71.1**	**66.7**	**96.7**	**23.3**

Note: A question is counted as flagged at a step if at least one of the step's coding decisions is coded as a problem. Bold rows are unweighted group means. LLM 1 = GPT-4; LLM 2 = Claude Sonnet 4.6.

A clear pattern emerges when considering which of the eight QAS steps contribute most to the total number of disagreements between coders (across both question sets) (see top panel of Figure 3). Across all comparisons and question sets, the two QAS steps that contribute most disproportionately to disagreements (crossing the dotted reference line in the figures) are Step 3 (clarity), where both flagging and disagreement rates were high, and Step 7 (response categories), where flagging rates were lowest for the LLMs, but disagreement rates were also generally lower.

Step 3 (which includes four coding decisions) accounted for 19% of disagreements between LLM 1 and Expert 1 in the full question set (validation set: 20% of disagreements between LLMs and Experts); and 22% of disagreements between LLM 1 and Novice 1 (validation set: 23% of disagreements between LLMs and Novices). For the human comparisons, Step 3 was responsible for 17% of the disagreements between Expert 1 and Novice 1 (validation set: 19% of disagreements between experts and novices and 18% of disagreements between experts). The middle panel of Figure 3 shows that for LLM-Novice disagreements, Step 3c (Vague) contributes most at 31%, followed by Step 3b (Technical Terms) at 27%, while Step 3a (Wording) falls clearly below the expected rate of 25% at 17%. For LLM-Expert disagreements the picture is more even: Steps 3b (27%), 3c (26%), and 3d (Reference Periods; 26

Step 7 (which includes seven separate coding decisions) accounts for 23% of disagreements between LLM 1 and Expert 1 (validation set: 25% aggregated across both LLMs and all experts), 19% of disagreements between LLM 1 and Novice 1 (validation set: 22% across both LLMs and novices). Three sub-steps of Step 7 contribute disproportionately to discordant flagging between LLM 1 and Expert 1: Step 7b (mismatches; 17%), Step 7d (vague response categories; 28%), and Step 7f (missing response categories; 28%). For LLM 1 and Novice 1, Step 7c (technical terms in response categories; 20%) also exceeds the expected rate of 14%, alongside Steps 7b (19%), 7d (26%), and 7f (17%). For the human comparisons, Step 7 accounts for 24% of disagreements between Expert 1 and Novice 1 (validation set: 25% across all experts and novices), with Steps 7b (16%), 7d (21%), and 7f (27%) again contributing above the expected rate (see lower panel of Figure 3). In the validation question set, this pattern was broadly consistent: Steps 7b, 7d, and 7f again accounted for the largest share of disagreements across LLM comparisons (LLM-Expert: 19%, 21%, 28%; LLM-Novice: 26%, 21%, 24%). The one divergence was for Expert-Novice comparisons, where Step 7e (overlapping response categories; 17%) also exceeded the expected rate in the validation set, a pattern not observed in the main question set.

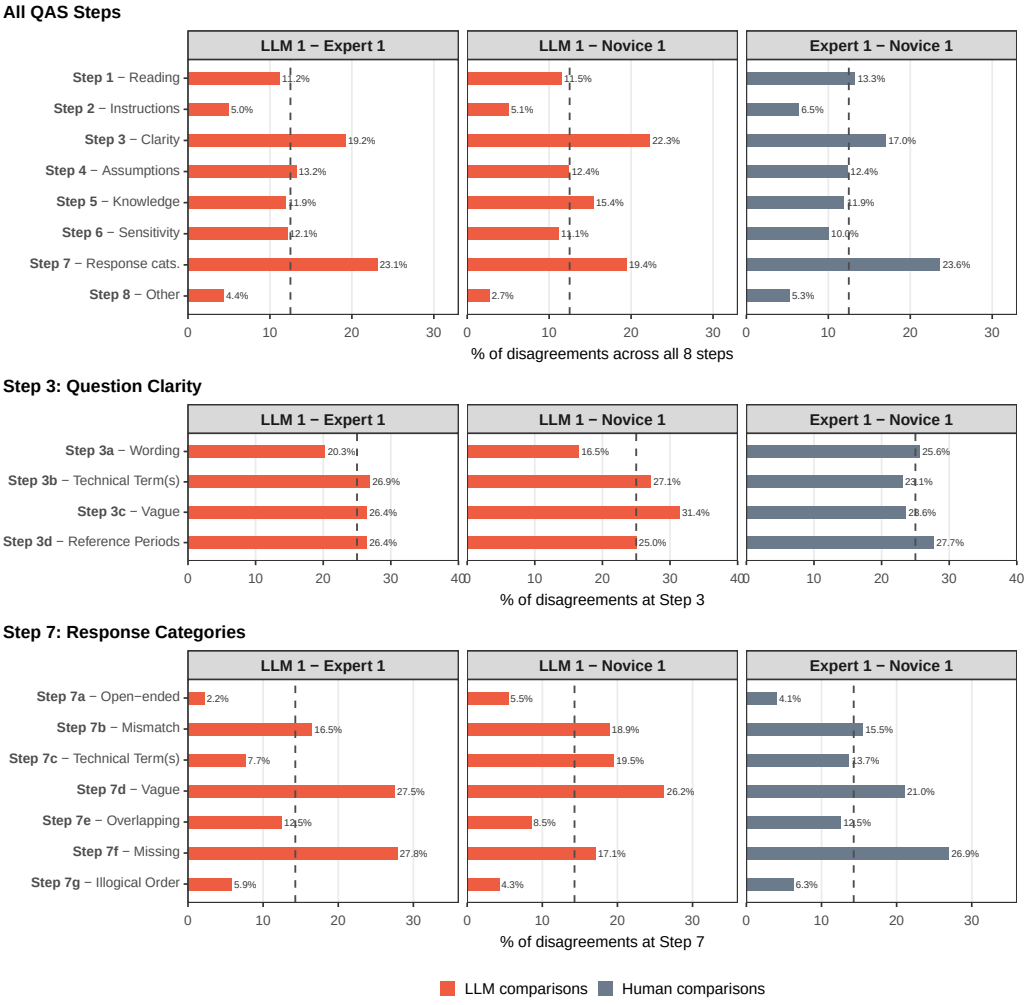


Figure 3: Contribution to total disagreements between coders by each QAS step

3.3 LLM hit rates for different types of data quality risk (RQ3)

The above findings confirm that while LLMs are consistently more conservative in their appraisal of draft survey questions compared to human evaluators, flagging propensity and patterns also vary idiosyncratically, and (apparently) as a function of ambiguity around certain coding decisions. To control for some of this variability, the final part of the analysis uses evaluations by Expert 1 (who had the highest flagging rate) as a comparison point against which to assess the appraisals made by GPT-4 and the novices.

Using the classification of 26 QAS decisions (excluding Step 8) presented in Table 1, Figure 4 shows the 'hit rates' for LLM 1 and Novice 1 – i.e., the percentage of problems flagged by Expert 1 that each also flagged. The figure also shows the 'unconfirmed flag rate' – the number of problems flagged by the LLM and the novices that were not flagged by Expert 1 (Appendix Table 5 shows average hit and unconfirmed flag rates for all coders, for the main set and the validation set). Low hit rates indicate 'under-flagging' (more conservative assessments compared to Expert 1) while unconfirmed flags indicate 'over-flagging' ('false discovery'), or simply flagging other problems Expert 1 missed (liberal assessments generally, or for specific QAS codes). The combination of a low hit rate and high unconfirmed flag rate indicates systematic divergence from Expert 1 – i.e., a coder who is interpreting specific QAS coding decisions differently altogether.

As previously noted, across all four types of data quality risk, LLM 1's (red circles) hit rate is lower on average than that of Novice 1 (blue circles) at 22% compared with 37%. This is markedly so for question problems presenting a risk of measurement bias, where the model's average hit rate was 24% compared with 51% for the novice). GPT-4 was especially likely to underflag questions with sensitive wording and socially acceptable answers, while the novices had the highest hit rates (over 50%) for questions presenting recall and computation challenges. Hit rates for both coders were lowest for question problems presenting a risk of low validity (though not necessarily low reliability), though Novice 1's rate was still just over twice that of the LLM (34% versus 15%). GPT-4 was far less likely to detect double-barreled questions, questions where knowledge might be lacking and questions with missing response categories. The hit rates for the two coders were most similar for problems presenting a risk of item-nonresponse (26% for the LLM versus 29% for Novice 1), though the LLM was better at detecting problems with instructions and complex wording. Problems relating to response categories presenting a risk of increased measurement variance (e.g., vague, overlapping, or illogical ordering) were also far less likely to be detected by the LLM compared to the novices, though the hit rate was

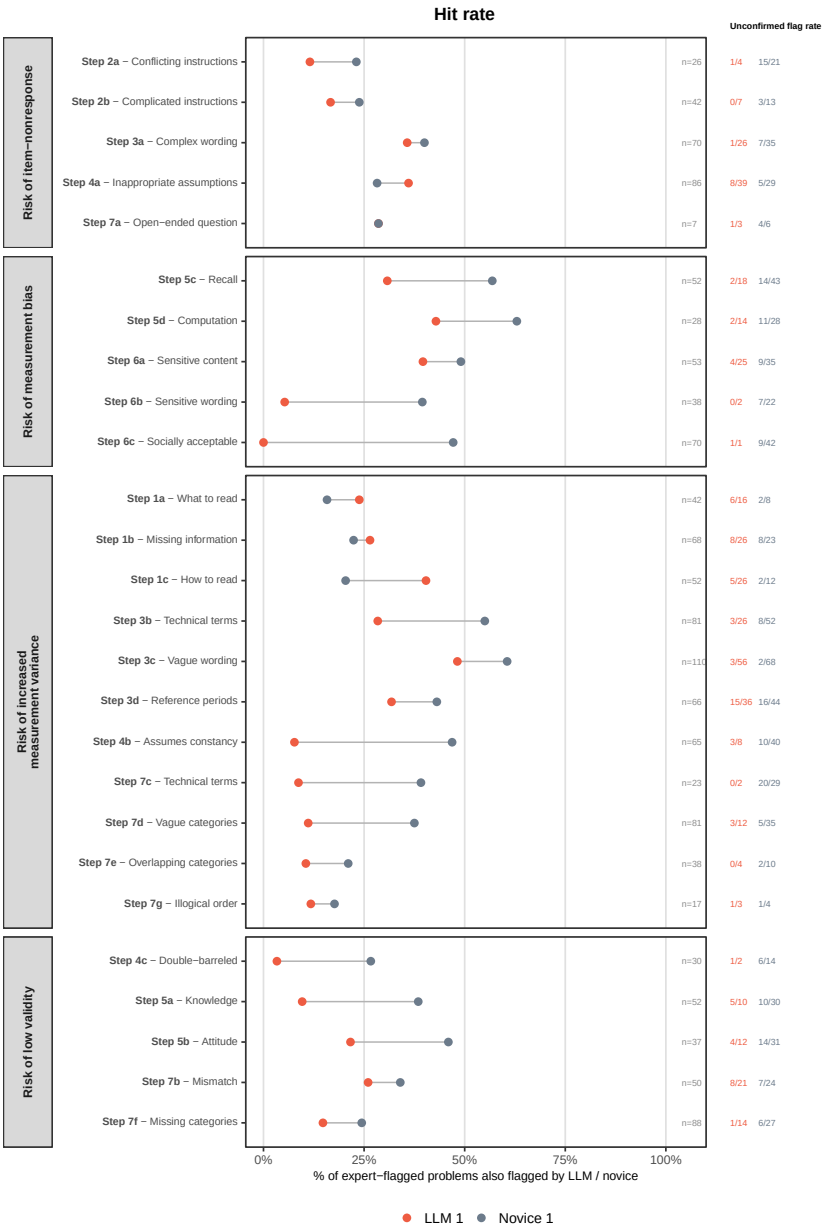


Figure 4: Hit rates and unconfirmed flag rates for LLM 1 and Novice 1 compared to Expert 1 by data quality risk type

still below 50% for the latter, and only exceeded 50% in this category for questions with vague wording or with undefined technical terms.

In the validation question set, findings were broadly consistent with the main set. The most notable divergence was for measurement bias risks: Claude performed markedly better than GPT-4 (average hit rate 62% versus 21%), bringing the LLM mean above that of the novices (42%) and approaching the expert mean (67%) for this category. This confirms measurement bias as the most skill-dependent data quality risk category in the QAS.

The unconfirmed flag rates presented in the right-hand panel of Figure 6 show the number of problems flagged by each coder that were not flagged by Expert 1 (in the numerator) of the total number of problems each coder flagged (in the denominator). Summarising across the problem types, in the main question set, GPT-4's average 'overflagging' rate was 25%, which was lower than that of Novice 1 (31%). Mean overflagging rates varied across the data quality risk categories. Similar 'overflagging' rates between coders were observed for problems presenting threats to validity and problems presenting risks of measurement bias. The lowest unconfirmed flag rates were for problems presenting threats to reliability (23% for the LLM versus 26% for Novice 1). Overflagging was markedly lower for the LLM for problems presenting risks of item-nonresponse (17% versus 40%). These summary statistics obscure the actual number of incidents concerned, however. In the validation question set, unconfirmed flag rates followed a similar pattern: rates were higher for novices than LLMs overall (46% versus 28%), and expert rates were comparable to those of the LLMs (32%).

4 Discussion

Understanding how best to capitalise on the strengths of LLMs and to establish their current limitations is a key priority for survey methodology and practice (Eloundou et al., 2023; Brand et al., 2023; Jansen et al., 2023; United Nations Economic Commission for Europe (UNECE), 2023; Learner, 2024; Wuttke et al., 2025). This study addressed two related aims: to demonstrate a scalable procedure for automating systematic question evaluations using a rule-based coding framework, and to validate LLM performance in applying RTI's Question Appraisal System-99 (Willis and Lessler, 1999) - one of the most widely used QDET frameworks — against human coders of varying expertise. The pipeline we developed returns structured data suitable for quantitative analysis or for generating actionable reports for users (see Appendix). To our knowledge, the inter-rater reliability assessment reported here is the largest formal evaluation of the QAS-99 undertaken to date

(Schaad et al., 2020), and the first to compare LLM and human ratings across all 27 QAS codes.

Across all three research questions, a consistent picture emerged. Overall agreement rates between LLMs and human coders were comparable to those between Experts and Novices, but Gwet's AC1, which corrects for chance agreement and is robust to low base rates (Gwet, 2008), revealed only slight to fair agreement between Expert 1 and all other coders. LLMs thus appear to detect question problems more like novices than experts, while both are comparably reliable relative to expert coding.

Focusing on aggregate agreement rates obscures significant variation in flagging thresholds and patterns across individual coders and coder types - variation attributable both to differences in expertise and to inherent subjectivity in certain QAS coding decisions (Willis and Lessler, 1999). Two steps contributed disproportionately to disagreements across all coder pairs: Step 3 (Clarity), where both flagging and disagreement rates were high, and Step 7 (Response Categories), where LLMs flagged notably fewer problems than human coders. Technical terms, vague wording, and mismatched or missing response categories were the specific sources of disagreement — coding decisions that likely require more nuanced linguistic and structural judgements than others.

Using Expert 1's evaluations as a benchmark, LLM 1 (GPT-4) under-flagged problems at a higher rate than the student novices across most data quality risk types, and was particularly unlikely to detect double-barreled questions, questions for which knowledge may be lacking, and questions with potentially sensitive or socially desirable answers. Among response-category problems, LLM 1 also under-flagged at a higher rate than Novice 1. Claude Sonnet 4.6, by contrast, detected markedly more measurement bias risks than GPT-4 in the validation set (62% vs 21% hit rate), though at the cost of a higher unconfirmed flag rate — a pattern consistent with a more liberal detection strategy overall.

Our findings align broadly with previous studies demonstrating disagreement between coders in their evaluation of questionnaires (e.g., Olson, 2010; Bais et al., 2019; Schaad et al., 2020). Schaad et al. (2020) similarly found that senior raters flagged more problems and that Step 7 (Response Categories) was among the steps with lowest agreement. Their finding that Step 2 (Instructions) also showed low agreement contrasts with our results, however, underscoring that observed levels of inter-rater agreement reflect not only coder expertise but also the characteristics of the specific questions evaluated and idiosyncratic code use by individual raters. Research at Statistics Netherlands likewise found considerable individual-level variation, especially for question characteristics deemed 'harder' to code due to weaker

definitions (Bais et al., 2019; Schouten et al., 2018). Notably, 'sensitive information' was one such characteristic, which is consistent with our finding that potentially questions with socially acceptable answers were among the problems least likely to be flagged by LLMs. This implies that the cultural sensitivity influencing human perceptions of such features may similarly result in their underdetection by AI, though our study suggests this varies by model.

Importantly, we also observed significant variation in agreement rates between individual QDET experts, consistent with a broader literature finding considerable inconsistencies in expert evaluations of questionnaires - especially for sensitivity-related characteristics (e.g., Presser and Blair, 1994; DeMaio and Landreth, 2004; Conrad and Blair, 2004; Olson, 2010). This context highlights the need for caution when evaluating LLM performance through comparison with human coders, without adequately accounting for variability in individual applications of coding schema and possible ambiguities in the framework itself. We do not consider Expert 1's coding to be ground truth; her high flagging rate may include substantial unsubstantiated over-flagging. As Rothgeb et al. (2007) noted of the QAS specifically, the method may encourage 'a low threshold for problem identification, producing a large number of false positive results' (p. 3). Finding ways to disentangle expertise effects from framework ambiguity is key to making valid judgements about the utility of rule-based QDET tools and the skill of those applying them.

Our results also emphasise that LLMs are capable of producing human-like evaluations of survey questions at scale and at speed. The expert QAS evaluations undertaken here took around 15 minutes on average per question, compared to less than a minute for the LLM. This confirms the cost efficiencies to be gained for questionnaire pre-testing, as well as for post-hoc studies of the link between question characteristics and measurement error (Maitland and Presser, 2020; Dykema et al., 2020). Performance nonetheless varies across models and will continue to improve. Results from a single prompt and single run should be interpreted cautiously, given known variability in LLM output across successive runs (e.g., Haase et al., 2026), though in our pretesting we found within-model variability to be lower than cross-model differences. When developing our prompt, we were able to significantly improve GPT-4's performance for some QAS codes by reformulating the original wording of the QAS steps to help improve task clarity and through minor adaptations to the system prompt designed to encourage more critical appraisals. Staying as close as possible to the instrument used by the human coders was a priority, however, so we did not experiment extensively with the prompt to try to optimise model performance.

Our conclusions about the relative QAS flagging rates of LLMs compared to humans should, therefore, be interpreted taking this limitation into account.

Further improvements to model performance are achievable through technical adjustments not yet explored here, such as adding few-shot examples, prompt chaining, or tuning model temperature. Practitioners using the prompt developed here should investigate further how prompt configuration and model choice affect task performance. The pattern of under-flagging for GPT-4 appeared to be most pronounced for problems requiring social or cultural judgements (social norms, sensitivity, adequacy of response format), while surface linguistic features were detected more reliably. Claude Sonnet 4.6 showed a different profile, performing better on conceptual and social codes but less well on some linguistic ones. This suggests that the interpretive strategies of different models may vary substantively, making them more or less adapted to certain tasks. These observations highlight the importance of model choice and configuration when using LLMs for QDET tasks, as well as the potential value of fine-tuning.

Perceptions of problem severity may arguably play an influential role in different raters flagging decisions, with some characteristics not being flagged when present, if they are not deemed especially problematic. While this seems likely to vary as a function of expertise, it is not helped by contradictory recommendations and empirical findings in the literature as to which question formats should be preferred and about the severity of the impact of different question problems on actual measurement quality. Paradoxically, unreliable application of rule-based QDET frameworks also contributes to this lack of coherence in the empirical literature on the relation between problematic question characteristics and survey measurement error (Bais et al., 2019; Schouten et al., 2018). A contradictory knowledge base also serves as a less than ideal basis for training LLMs to detect serious problems.

We attempted to assess whether detection rates varied as a function of the type of data quality risk presented by each problem, on the assumption that different risk types carry different consequences for measurement quality and call for different remedies. This classification was not unproblematic: certain codes could plausibly belong to more than one risk category, and the ultimate impact on data quality is determined by the combined role of multiple item characteristics, many of which are not addressed in the QAS-99 (but see Dean et al., 2007; Schaad et al., 2020, for important extensions). Alternative classifications may yield different conclusions about the relative utility of the framework or the skill of those applying it.

5 Conclusion

This study demonstrates that LLMs can apply a systematic QDET framework at scale with reliability broadly comparable to novice human coders, while exhibiting identifiable and model-specific patterns of under-detection. Ultimately, however, the goal should not only be an evaluation tool that is reliably applied, but one that can also lead to significant improvements to question validity and reliability, and by extension, data quality. This was the primary motivation for the QAS — to catch the significant threats to measurement quality so they could be addressed in subsequent phases of pretesting, such as cognitive interviews (Willis and Lessler, 1999). Integrating LLMs into QDET processes presents new opportunities to streamline such procedures, particularly because of their apparent strengths at proposing effective revisions to flawed questions based on problems identified. Generative AI also provides novel ways to strengthen the empirical base for knowledge about links between questionnaire design and measurement error, through large-scale coding of item characteristics in existing surveys (particularly those with experimental designs) and study of their associations with observed measurement quality. Along with greater conceptual clarification of the mechanisms by which different question characteristics can impact measurement inference, advances in GenAI can support meaningful improvements in the design and application of rule-based QDET systems, and in what is understood about effective social measurement methodology.

Acknowledgements

This research was made possible through the use of data extracted from assignments completed by students enrolled in the MA course 'Questionnaire Design' at the University of Lausanne. We are grateful to these 'novice' coders for their contribution to this study.

Funding

This research was undertaken as part of Research Strand 9 of the UKRI-ESRC funded research programme 'Survey Futures - Survey Data Collection Methods Collaboration'.

Data availability

Replication data and code will be made available via GitHub.

References

D. M. Adhikari, A. Hartland, I. Weber, and V. K. Cannanure. Exploring LLMs for automated generation and adaptation of questionnaires, 2025.

F. Bais, B. Schouten, P. Lugtig, V. Toepoel, J. Arends-Tòth, S. Douhou, N. Kieruj, M. Morren, and C. Vis. Can survey item characteristics relevant to measurement error be coded reliably? A case study on 11 Dutch general population surveys. *Sociological Methods & Research*, 48(2):263–295, 2019. doi: 10.1177/0049124117729692.

J. Brand, A. Israeli, and D. Ngwe. Using GPT for market research, 2023. SSRN working paper.

T. D. Buskirk, A. Eck, and J. Timbrook. The task is to improve the ask: An experiment for developing prompts to generate high quality survey items from large language models, 2025a. SSRN working paper.

T. D. Buskirk, F. Keusch, L. von der Heyde, and A. Eck. More parameters than populations: A systematic literature review of large language models within survey research, 2025b. URL https://arxiv.org/abs/2509.03391.

P. Campanelli. Testing survey questions. In E. D. de Leeuw, J. J. Hox, and D. A. Dillman, editors, *International Handbook of Survey Methodology*. Lawrence Erlbaum, New York, NY, 2008.

C. F. Cannell, P. V. Miller, and L. Oksenberg. Research on interviewing techniques. *Sociological Methodology*, 12:389, 1981. doi: 10.2307/270748.

F. G. Conrad and J. Blair. Data quality in cognitive interviews: The case of verbal reports. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, and J. Martin, editors, *Methods for Testing and Evaluating Survey Questionnaires*, pages 67–87. John Wiley & Sons, 2004.

J. d'Ardenne, R. Bull, A. Das, Z. Perera, and O. Sexton. Survey practice guide 2: How to mitigate against measurement effects when surveys move online. A Measurement Effect Risk Framework (MERF) and web questionnaire guidance. Technical report, Survey Futures, 2025. URL https://surveyfutures.net/wp-content/uploads/2025/06/survey-practice-guide-2-mitigate-measurement-effects-surveys-move-online.pdf.

E. Dean, R. Caspar, G. McAvinchey, L. Reed, and R. Quiroz. Developing a low-cost technique for parallel cross-cultural instrument development: The Question Appraisal System (QAS-04). *International*

Journal of Social Research Methodology, 10(3):227–241, 2007. doi: 10.1080/13645570701401032.

T. DeMaio and A. Landreth. Examining expert reviews as a pretest method. In P. Prüfer, M. Rexroth, and F. J. Fowler, editors, *QUEST 2003: Proceedings of the 4th Conference on Questionnaire Evaluation Standards*, pages 60–73. ZUMA, 2004.

D. A. Dillman, J. D. Smyth, and L. M. Christian. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Wiley, 4th edition, 2014.

J. Dykema, N. C. Schaeffer, D. Garbarski, and M. Hout. The role of question characteristics in designing and evaluating survey questions. In P. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. Willis, and A. Wilmot, editors, *Advances in Questionnaire Design, Development, Evaluation and Testing*, pages 119–152. Wiley, 2020.

T. Eloundou, S. Manning, P. Mishkin, and D. Rock. GPTs are GPTs: An early look at the labor market impact potential of large language models, 2023. arXiv preprint.

B. Felderer, L. Repke, W. Weber, J. Schweisthal, and L. Bothmann. Predicting the validity and reliability of survey questions, 2024. OSF Preprints.

R. Fitzgerald, S. Widdop, M. Gray, and D. Collins. Identifying sources of error in cross-national questionnaires: Application of an error source typology to cognitive interview data. *Journal of Official Statistics*, 27(4): 569–599, 2011.

B. H. Forsyth and M. Hubbard. A method for identifying cognitive properties of survey items. In *Proceedings of the Section on Survey Research Methods*, pages 470–475, Washington, DC, 1992. American Statistical Association.

B. H. Forsyth, J. M. Rothgeb, and G. B. Willis. Does pretesting make a difference? An experimental test. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer, editors, *Methods for Testing and Evaluating Survey Questionnaires*, pages 525–546. Wiley, 2004.

F. J. Fowler. *Improving Survey Questions: Design and Evaluation*. SAGE Publications, Thousand Oaks, CA, 1995.

A. Fuchs, A.-C. Haensch, and W. Weber. AI for survey design: Generating and evaluating survey questions with large language models, 2026. SocArXiv preprint.

A. C. Graesser, K. Wiemer-Hastings, R. Kreuz, P. Wiemer-Hastings, and K. Marquis. QUAID: A questionnaire evaluation aid for survey methodologists. *Behavior Research Methods, Instruments & Computers*, 32(2): 254–262, 2000. doi: 10.3758/bf03207792.

R. M. Groves, F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey Methodology*. Wiley, 2nd edition, 2009.

K. L. Gwet. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Statistical Methods for Inter-Rater Reliability Assessment, no. 1. AgreeStat, 2002. URL https://agreestat.com/papers/kappa_statistic_is_not_satisfactory.pdf.

K. L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008. doi: 10.1348/000711006X126600.

J. Haase, J. Gonnermann-Müller, P. H. P. Hanel, N. Leins, T. Kosch, J. Mendling, and S. Pokutta. Within-model vs between-prompt variability in large language models for creative tasks, 2026. URL https://arxiv.org/abs/2601.21339. arXiv preprint.

B. J. Jansen, S. Jung, and J. Salminen. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020, 2023. doi: 10.1016/j.nlp.2023.100020.

A. Kostina, M. D. Dikaiakos, D. Stefanidis, and G. Pallis. Large language models for text classification: Case study and comprehensive review, 2025. URL https://arxiv.org/abs/2501.08457. arXiv preprint.

J. A. Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236, 1991. doi: 10.1002/acp.2350050305.

J. A. Krosnick. Experiments for evaluating survey questions. In J. H. Madans, K. Miller, A. Maitland, and G. B. Willis, editors, *Question Evaluation Methods: Contributing to the Science of Data Quality*, pages 215–238. Wiley, 2011.

J. Learner. The promise and pitfalls of AI-augmented survey research. NORC at the University of Chicago, 2024. URL https://www.norc.org/research/library/promise-pitfalls-ai-augmented-survey-research.html.

T. Lenzner. Effects of survey question comprehensibility on response quality. *Field Methods*, 24(4):409–428, 2012. doi: 10.1177/1525822x12448166.

J. T. Lessler and B. H. Forsyth. A coding system for appraising questionnaires. In N. Schwarz and S. Sudman, editors, *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, pages 259–292. Jossey-Bass, 1996.

A. Maiorino, Z. Padgett, C. Wang, M. Yakubovskiy, and P. Jiang. Application and evaluation of large language models for the generation of survey questions. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5244–5245, 2023. doi: 10.1145/3583780.3615506.

A. Maitland and S. Presser. How accurately do different evaluation methods predict the reliability of survey questions? *Journal of Survey Statistics and Methodology*, 4(3):362–381, 2016.

A. Maitland and S. Presser. How do question evaluation methods compare in predicting problems observed in typical survey conditions? *Journal of Survey Statistics and Methodology*, 6(4):465–490, 2018. doi: 10.1093/jssam/smx036.

A. Maitland and S. Presser. A comparison of five question evaluation methods in predicting the validity of respondent answers to factual items. In P. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. Willis, and A. Wilmot, editors, *Advances in Questionnaire Design, Development, Evaluation and Testing*, pages 75–90. Wiley, 2020.

E. A. Metheney and L. Yehle. Generative AI as a safety net for survey question refinement, 2025. URL https://arxiv.org/abs/2509.08702. arXiv preprint.

F. Olivos and M. Liu. ChatGPTest: Opportunities and cautionary tales of utilizing AI for questionnaire pretesting. *Field Methods*, 37(1), 2025. doi: 10.1177/1525822X241280574.

K. Olson. An examination of questionnaire evaluation by expert reviewers. *Field Methods*, 22(4):295–318, 2010. doi: 10.1177/1525822X10379795.

K. Olson and T. D. Buskirk. "ChatBot" is a two syllable word...or is it?: Using generative AI for survey question readability assessments. *International Journal of Market Research*, 68(1):61–81, 2026.

OpenAI. GPT-4 technical report. Technical report, 2023. URL https://cdn.openai.com/papers/gpt-4.pdf.

S. Presser and J. Blair. Survey pretesting: Do different methods produce different results? *Sociological Methodology*, 24:73–104, 1994. doi: 10.2307/270979.

S. Presser, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, J. M. Rothgeb, and E. Singer. Methods for testing and evaluating survey questions. *Public Opinion Quarterly*, 68(1):109–130, 2004. doi: 10.1093/poq/nfh008.

J. Rothgeb, G. Willis, and B. Forsyth. Questionnaire pretesting methods: Do different techniques and different organizations produce similar results? *Bulletin de Méthodologie Sociologique*, 96(1):5–25, 2007. doi: 10.1177/075910630709600103.

D. Rothschild, J. Marlar, A. Amaya, S. Barari, T. Buskirk, C. Cobb, J. Gennai, D. Hillygus, R. Korlakai Vinayak, M. Krupenkin, S. Lee, D. Steiger, and B. Webb. Responsible AI integration in survey research. Technical report, American Association for Public Opinion Research, 2026. URL https://aapor.org/wp-content/uploads/2026/05/Responsible-AI-Integration-In-Survey-Research.pdf.

W. E. Saris and I. Gallhofer. *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. John Wiley & Sons, 2007.

A. Schaad, M. Jans, and M. Scott. Improving the Question Appraisal System (QAS). Proceedings of the Survey Research Methods Section (Paper 1505493). American Statistical Association, 2020. URL http://www.asasrms.org/Proceedings/y2020/files/1505493.pdf.

B. Schouten, F. Bais, and V. Toepoel. Estimating survey questionnaire profiles for measurement error risk. *Journal of Survey Statistics and Methodology*, 6(3):306–334, 2018. doi: 10.1093/jssam/smx027.

P. Sturgis, C. Roberts, and T. Robinson. LLMs for survey pretesting: How well can large language models identify flaws in survey questions? Technical Report 15, Survey Futures, Colchester, UK, 2026. URL https://surveyfutures.net/working-papers/.

R. Tao, R. Yang, G. Walejko, Y. Yang, and B. Groenhout. Using large language models (LLM) to pretest survey questions. Conference presentation, AAPOR Conference, Atlanta, GA, 2024. URL https://aapor.confex.com/aapor/2024/meetingapp.cgi/Paper/2861.

J. Timbrook, A. Griggs, J. Bollenbacher, and R. Stenger. Prompting for perfection: The role of LLMs in questionnaire design, evaluation and pretesting. Presentation at the 81st AAPOR Conference, Los Angeles, CA, 2026.

R. Tourangeau. Cognitive science and survey methods. In T. Jabine, M. Straf, J. Tanur, and R. Tourangeau, editors, *Cognitive Aspects of Survey Design: Building a Bridge Between Disciplines*, pages 73–100. National Academy Press, 1984.

R. Tourangeau and K. A. Rasinski. Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103(3):299–314, 1988. doi: 10.1037/0033-2909.103.3.299.

R. Tourangeau, A. Maitland, D. Steiger, and T. Yan. A framework for making decisions about question evaluation methods. In P. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. Willis, and A. Wilmot, editors, *Advances in Questionnaire Design, Development, Evaluation and Testing*, pages 47–73. Wiley, 2020.

United Nations Economic Commission for Europe (UNECE). Large language models for official statistics (HLG-MOS white paper). Technical report, 2023. URL https://unece.org/sites/default/files/2023-12/HLGMOS%20LLM%20Paper_Preprint_1.pdf.

G. B. Willis. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. SAGE Publications, 2005.

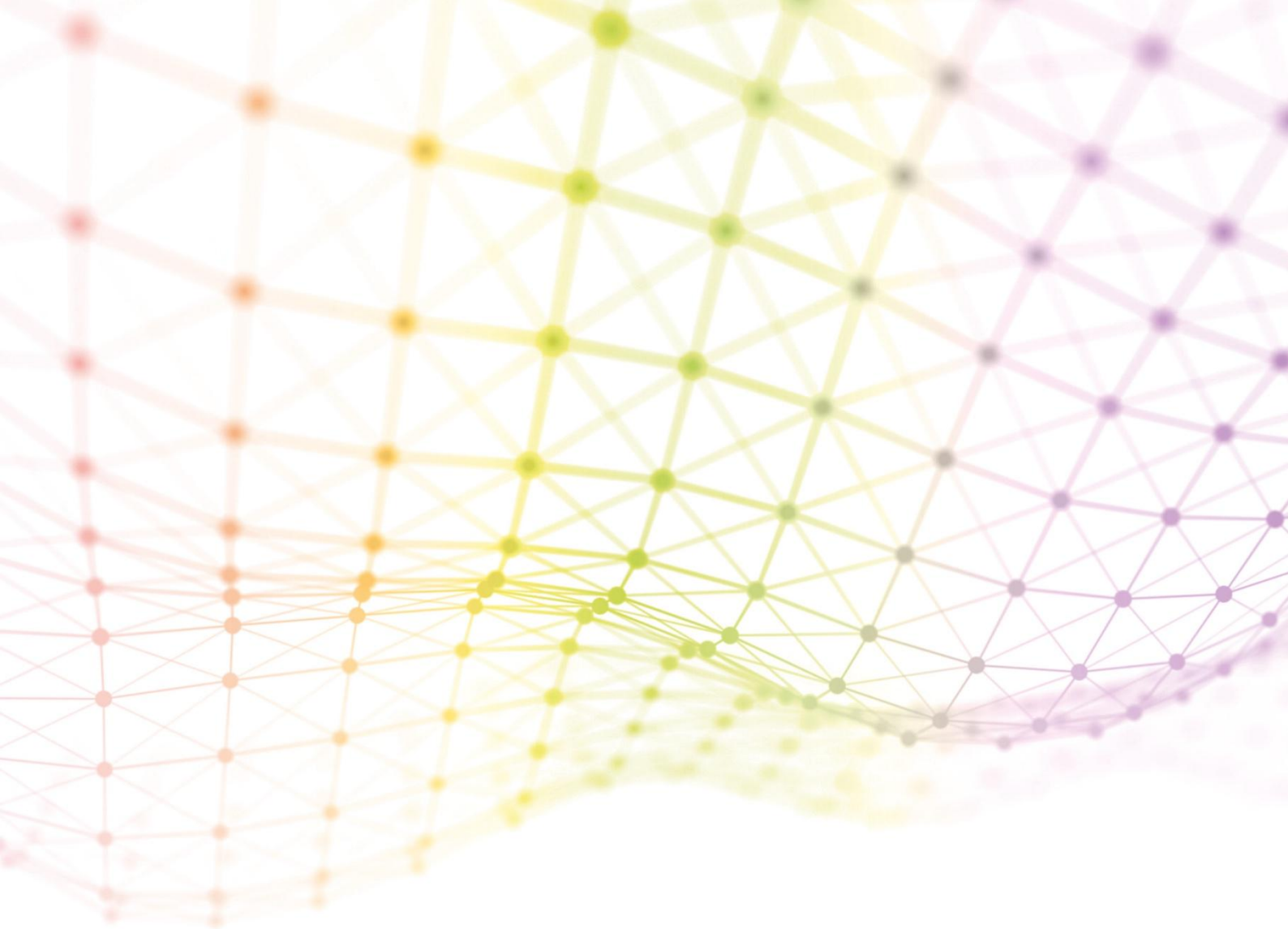
G. B. Willis. Questionnaire design, development, evaluation, and testing: Where are we, and where are we headed? In P. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. Willis, and A. Wilmot, editors, *Advances in Questionnaire Design, Development, Evaluation and Testing*, pages 1–34. John Wiley & Sons, 2020. doi: 10.1002/9781119263685.ch1.

G. B. Willis and J. Lessler. The BRFSS-QAS: A guide for systematically evaluating survey question wording. Technical report, Research Triangle Institute, 1999.

N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13:61, 2013. doi: 10.1186/1471-2288-13-61.

A. Wuttke, M. Aßenmacher, C. Klamm, M. M. Lang, Q. Würschinger, and F. Kreuter. AI conversational interviewing: Transforming surveys with LLMs as adaptive interviewers, 2025. URL https://arxiv.org/abs/2410.01824. arXiv preprint.

G. Zhang. Knowledge protocol engineering: A new paradigm for AI in domain-specific knowledge work, 2025. URL https://arxiv.org/abs/2507.02760. arXiv preprint.



University of Essex

University of Southampton

UKRI Economic and Social Research Council

NCRM National Centre for Research Methods

Office for National Statistics

UCL

Warwick The University of Warwick

Manchester 1824 The University of Manchester

City University of London Est 1894

National Centre for Social Research

LSE The London School of Economics and Political Science

Ulster University

Unil UNIL | Université de Lausanne

Ipsos

verian Formerly Kantar Public

www.surveyfutures.net