Exploring LLMs for Automated Generation and Adaptation of Questionnaires

Divya Mani Adhikari
diad00001@stud.uni-saarland.de
Interdisciplinary Institute for Societal Computing
Saarland University
Saarbrücken, Germany

Alexander Hartland*
alexander.hartland@uni-saarland.de
Department of European Social Research
Saarland University
Saarbrücken, Germany

Ingmar Weber
iweber@cs.uni-saarland.de
Interdisciplinary Institute for Societal Computing
Saarland University
Saarbrücken, Germany

Vikram Kamath Cannanure
cannanure@cs.uni-saarland.de
Interdisciplinary Institute for Societal Computing
Saarland University
Saarbrücken, Germany

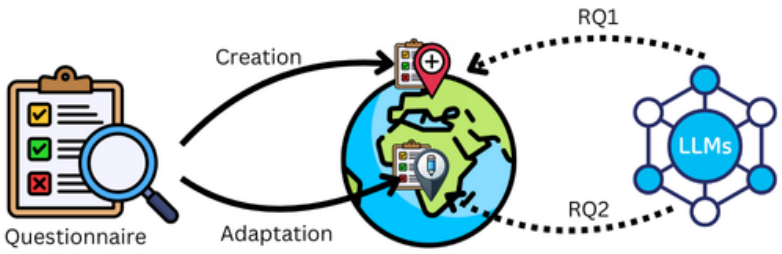


Figure 1: We explore the role of LLMs in questionnaire design by understanding their capabilities in creating and pretesting new questionnaires (RQ1: creation) and modifying existing ones (RQ2: adaptation) to suit different contexts.

Abstract

Effective questionnaire design improves the validity of the results, but creating and adapting questionnaires across contexts is challenging due to resource constraints and limited expert access. Recently, the emergence of LLMs has led researchers to explore their potential in survey research. In this work, we focus on the suitability of LLMs in assisting the generation and adaptation of questionnaires. We introduce a novel pipeline that leverages LLMs to create new questionnaires, pretest with a target audience to determine potential issues and adapt existing standardized questionnaires for different contexts. We evaluated our pipeline for creation and adaptation through two studies on Prolific, involving 238 participants from the US and 118 participants from South Africa. Our findings show that participants found LLM-generated text clearer, LLM-pretested text more specific, and LLM-adapted questions slightly clearer and less biased than traditional ones. Our work opens new opportunities for LLM-driven questionnaire support in survey research.

*Also with Interdisciplinary Institute for Societal Computing, Saarland University.

CUI '25, Waterloo, ON, Canada

ACM ISBN 979-8-4007-1527-3/2025/07
https://doi.org/10.1145/3719160.3736606

CCS Concepts

• **Computing methodologies** → **Natural language generation**; • **Applied computing** → Sociology; • **Human-centered computing** → *Empirical studies in HCI*; *Natural language interfaces*; • **Information systems** → **Crowdsourcing**.

Keywords

Large Language Models (LLMs), Survey Methodology, Questionnaire Design, Questionnaire Pretesting, Cross-cultural Adaptation

ACM Reference Format:
Divya Mani Adhikari, Alexander Hartland, Ingmar Weber, and Vikram Kamath Cannanure. 2025. Exploring LLMs for Automated Generation and Adaptation of Questionnaires. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25), July 8–10, 2025, Waterloo, ON, Canada.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3719160.3736606

1 Introduction

Designing well-structured questionnaires is essential for gathering survey data, as it influences the validity of collected responses [42, 46] from diverse populations worldwide. While standardized surveys exist, designing a new questionnaire remains challenging, requiring a sequence of logically structured questions aligned with specific research objectives [10, 36, 74, 75]. Researchers must craft clear, relevant, and precise questions to extract meaningful insights

without causing confusion or misinterpretation among respondents [46]. Pretesting is a crucial step in questionnaire design, ensuring that errors, biases, or ambiguities—especially in sensitive [17, 45] or cross-cultural [3, 23] surveys—are identified before deployment. Pretesting is necessary for cross-cultural studies [6, 9], global market research [63], and international policy assessments [53], where question phrasing and interpretation can vary significantly across different regions. However, resource constraints [71], participant availability [8, 37], and limited expert access [58] make pretesting difficult for new questionnaires or adapting existing questionnaires to different contexts.

Automated pre-testing with Large Language Models (LLMs) offers a promising solution to key challenges in survey design [34]. LLMs have been widely adopted for text generation and analysis due to their ability to produce complex, coherent responses based on given instructions [49]. Additionally, LLMs can simulate pilot studies by adopting different personas [66], allowing researchers to evaluate survey effectiveness before deployment. Furthermore, LLMs can analyze questionnaires to detect potential issues, enhancing the pre-testing process and reducing the need for multiple human iterations [57, 66]. This work integrates LLMs (specifically GPT-4o) into a questionnaire design to understand their utility and perform a preliminary evaluation. Through our work, we answer the following research questions:

RQ1: Creation How suitable are LLMs for automating new questionnaire generation and pretesting?

RQ2: Adaptation How suitable are LLMs at adapting existing questionnaires for a specific target audience?

RQ	No. of participants	Country	Subject expertise requested
Creation	238 (Prolific)	United States	Politics, Psychology, Social Work, Sociology
	13 (Experts)	EU	Political Science, Psychology, Social Science
Adaptation	118 (Prolific)	South Africa	Earth, Environment, or Climate Sciences, Politics, Psychology, Social Work, Sociology

Table 1: The number of participants, their geographic distribution, and subject backgrounds for two research questions: Creation and Adaptation. Participants were recruited from Prolific and expert networks across different regions (US, EU, and South Africa). For each study, participants aged 18+ with an undergraduate degree in the listed subjects were recruited.

To address these research questions, we conducted two studies using Prolific[1], an online crowdsourcing platform. For RQ1, we utilized our pipeline to generate a questionnaire on political trust and pretested it with 238 Prolific participants and 13 experts. For RQ2, we adapted a U.S.-centric standardized public opinion survey on climate change for a South African audience and gathered feedback from 118 South African participants on Prolific. As part of our evaluation, participants rated LLM-generated and LLM-pretested survey questions on a 5-point Likert scale based on clarity, bias, relevance, and specificity. They then compared both types of questions side by side, selecting the one they found clearer and briefly explaining their choice.

For our first research question (RQ1: Creation), we found that LLMs have the potential to generate relevant and specific questions while lacking in pretesting survey questionnaires. This suggested that LLMs are already good at generating high-quality questions, and further pretesting might be unnecessary. For the second question (RQ2: Adaptation), participants on Prolific found the pretested questionnaire slightly more explicit, unbiased, and relevant, and thus, more appropriate for a South African target audience. This result suggested that LLMs can adapt existing questionnaires to a specific target audience.

Overall, we make the following contributions:

- We show empirical results from two studies on Prolific about the suitability and challenges of LLMs in creating and adapting survey questionnaires.
- We introduce a novel LLM-based pipeline designed for the automated generation and pretesting of survey questionnaires using an LLM-simulated Pilot study. The pipeline can be adapted for use with any existing LLM or in any domain.
- We discuss strategies for questionnaire design for clarity, specificity, and neutrality in survey questions.

2 Related Work

2.1 Automated Survey Administration

Existing research has primarily focused on automating survey administration using computational techniques, particularly conversational systems. Chatbots have shown promise in replacing human interviewers while improving data quality [41], increasing completion rates and reducing costs [20, 55], and even enhancing interaction [1]. More recently, Large Language Models (LLMs) have further expanded the automation possibilities in survey research. LLM-powered interviewers can collect data at a quality comparable to traditional methods while offering scalability [80]. Moreover, Villalba et al. [77] showed LLM-augmented chatbots improve user experience through dynamic follow-up question generation during interviews. While automation in survey administration is well-studied, the automation of questionnaire generation and pretesting remains underexplored. With the rise of LLMs, there is growing interest in leveraging their text generation and analytical capabilities to enhance survey research. Jansen et al. [34] discuss how LLMs could support the pretesting of survey instruments, signaling a shift toward broader automation in survey design.

2.2 Questionnaire Generation

Questionnaires consist of a series of questions that include coreferences and inherent constraints, making their creation more complex than generating general sequential questions [44]. Existing research on question generation has primarily focused on creating independent, factual questions [61, 81]. This approach has been adapted

[1] https://prolific.com/

within survey research to generate follow-up questions in chatbot-administered interviews [22, 62, 69, 72]. Beyond individual question generation, researchers have also explored methods for generating sequences of questions [11, 21]. However, work specifically on questionnaire generation remains limited. Early research by Jenkins and Solomonides [35] explored automating questionnaire design by reusing standardized questions and searchable question libraries. More recently, studies have investigated the potential of LLMs in this space[51, 60]. LLMs have been used in generating rephrases of questionnaires [84], generating diverse versions of existing questionnaires to reduce respondent fatigue in longitudinal studies [83], and even generating HR-related questionnaires [43]. Similarly, Rothschild et al. [64] found using LLMs in the survey designing process substantially reduces the survey launch time. However, Lei et al. [44] and Laraspata et al. [43] found that LLM-generated questions were generally too broad, had generic wording, and lacked specificity. Laraspata et al. [43] propose the use of Retrieval-Augmented Generation (RAG) [48] as a strategy to reduce hallucinations and enhance the specificity of generated questions.

2.3 Automated Pretesting of Questionnnaire

Pretesting a questionnaire requires feedback from target participants or domain experts, thus automation in this area remains largely unexplored. Traditional pretesting methods include Cognitive Interviewing [68], Pilot Studies (involving participants) [33], Expert Review [59] and Behavior Coding (involving domain experts) [7, 63]. While some early studies explored online adaptations, such as conducting cognitive interviews via Skype or Second Life [13], automation in pretesting has remained limited. One exception is Web Probing, a technique used in online surveys where participants are asked follow-up questions (e.g., You answered "Yes"—how did you arrive at this answer?) to understand their thought process [5, 47, 56].

Recent work has explored using LLMs to automate questionnaire pretesting. Parker et al. [60] used ChatGPT to review interview protocols but found the feedback not critical enough to refine the protocols. Likewise, Olivos and Liu [57], Sarstedt et al. [66] demonstrated how ChatGPT can review survey questions to identify potential issues. However, these reviews are limited by the biases of the LLMs and the datasets they have been trained on [65, 76]. LLMs can adopt various personas, allowing for the evaluation of questionnaires from diverse perspectives, thereby mitigating biases [25, 32, 73]. Researchers have explored using LLMs to simulate diverse audience perspectives [34, 39]. Kim et al. [40] demonstrated that LLMs when initialized with a persona, can simulate human responses on an individual level. Similarly, Argyle et al. [2] showed that properly conditioning GPT-3 enables it to replicate responses from demographic subgroups.

However, LLMs are not a perfect substitute for human participants. Studies show that while LLMs can infer human biases [67], serve as simulated economic agents [30], and generate plausible HCI research data [31], they can also misrepresent and flatten demographic groups [78]. Researchers caution against using LLMs as a complete replacement for human responses [14, 15], but suggest they can be valuable supplements in pretesting, particularly in pilot studies and cognitive interviews. Wang et al. [78] also proposed techniques to mitigate misrepresentation issues, supporting the idea that LLMs, while imperfect, can be used to simulate pilot studies and pretesting methods to refine survey design before real-world deployment. Our work considers these techniques and extends existing work by utilizing LLM personas to simulate a Pilot study and analyze questionnaires from a variety of perspectives.

2.4 Cross-cultural Questionnaire Adaptation

Adaptation is essential when a questionnaire designed for a specific context, such as a particular language or country, is to be used in a different setting. While no consensus exists for a particular approach [19], much of the existing research has focused on adapting questionnaires from one language to another through translation [4, 26]. The adapted questionnaire requires further validation, such as pretesting, before being deployed to a new audience [4, 19]. Haavisto and Welsch [26] translated a questionnaire and used GPT-4 to evaluate the translation quality to suggest improvements. They found that GPT-suggested changes can achieve similar results as conventional methods.

Even within the same language, such as English, adaptation is crucial to mitigate potential misunderstandings due to cross-cultural differences. Harzing [28] found minimal difference in the participant responses when adapting an English questionnaire; however, Sousa et al. [70] suggested that participants can interpret original and adapted questions differently, with adaptation generally making the questions more straightforward to understand. While existing works rely on expert feedback or back-translation for questionnaire adaptation [18], our work explores questionnaire adaptation and evaluation using simulation of participants from the new target audience through a Pilot study using LLMs.

In summary, existing work has automated survey administration using chatbots and LLMs, improving scalability and data quality, but questionnaire generation and pretesting remain underexplored. While LLMs aid in structuring and adapting surveys, issues of specificity and bias persist. Our research aims to bridge these gaps by enhancing questionnaire generation through RAG and simulating pilot studies for automated pretesting and adaptation, to ultimately improve the quality and relevance of survey instruments.

3 Methodology

3.1 System Design

3.1.1 Questionnaire Generation. We start with a detailed research question the researcher wants to answer through the survey. The research question might include the goal of the study, details about the target audience, and what specific topics they want the questionnaire to include. The researcher submits the research question along with the number of questions that need to be generated. In addition to the research question, the LLM receives relevant contextual information supporting questionnaire generation: a summary of recent news articles and relevant standardized questions through the questionnaire generation prompt. We use the Exa API [2] to retrieve 5 recent relevant News articles and then use GPT-4 to summarize them. We also use 10 relevant questions retrieved from

[2] https://exa.ai

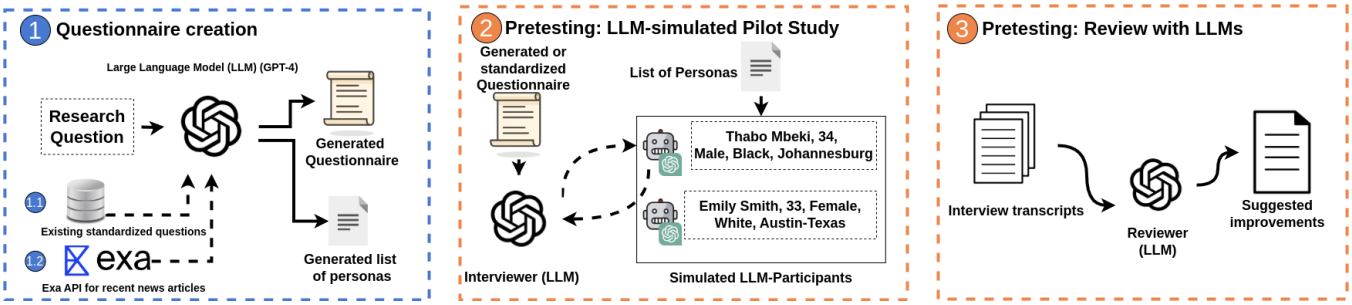


Figure 2: Our complete questionnaire creation and pretesting pipeline involve: (1) Using the research question and relevant contextual information, such as a few questions from existing standardized questionnaires and a summary of recent relevant news articles (see figure 3), the LLM generates a questionnaire and a list of Personas. (2) The generated question is then pretested through a simulated pilot study using the generated personas. (3) Interview transcripts from the simulated pilot study are collected and analyzed by the reviewer LLM, which provides suggestions for improving the questionnaire.

the Survey Quality Predictor (SQP) database[3] [79]. SQP is a multilingual and multi-domain database of curated and crowdsourced questions from existing surveys. First, we determine the domain and language of the questionnaire from the given research question and then use this information in filtering a subset of SQP questions. We then use FAISS [16] to rank these questions and retrieve the 10 most relevant questions using Cosine similarity as shown in figure 3. This contextual information could make the generated questionnaires more relevant to current affairs and the political climate and specific to the research objective.

Along with the questionnaire, the LLM also generates a list of personas, which are descriptions of fictional people of the target audience. Each persona description includes essential characteristics about the person, such as their name, gender, age, employment, etc., and dynamic preferences based on the research question. For example, if the research question is to study the perspectives of people in Germany about the use of AI in the workplace, the generated list of personas would include fictional descriptions of people from Germany where some might encourage the usage of AI in the workplace while others might oppose it or have mixed opinions about it. This list of personas is then used during the pretesting phase to simulate the participants for the Pilot study.

3.1.2 Questionnaire Pretesting. The generated questionnaire is then pretested by simulating a real-world pilot study using multiple LLM instances. One of the LLMs is initialized as the interviewer, and multiple LLMs are initialized as the participants based on the previously generated list of personas. The interviewer LLM then asks questions from the generated questionnaire, and the simulated participant LLMs respond accordingly to their persona descriptions. After each question, the interviewer LLM asks follow-up questions, dynamically generated based on the previous responses of the participant. The generated follow-up questions followed the cognitive interviewing technique to determine the participants' thought processes. The conversation history (interview transcript) between the interviewer LLM and the participant LLMs is stored for further analysis.

[3] https://sqp.gesis.org/

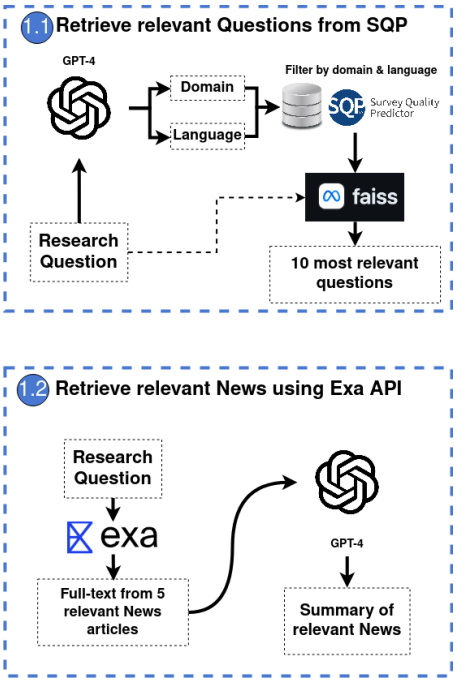


Figure 3: (1) Based on the research question, the domain and the language of the questionnaire are detected which is used to filter questions from the SQP database. Using FAISS, the 10 most relevant questions are retrieved using Cosine similarity. (2) We used the Exa API which returns the full text of 5 relevant News articles based on the research question which is then summarized by GPT-4. Both the relevant questions and summary of relevant News are included in the Questionnaire generation prompt.

After the pilot, the collected interview transcripts are analyzed by a reviewer LLM, which is initialized to find potential issues in the pretested questionnaire, such as double-barreled, leading, or biased questions. The reviewer LLM gets the initial research questions,

the questionnaire under pretest, and the interview transcript and persona description of each simulated interview as context data. The reviewer LLM then generated a review with four topics: the question with a problem, the problem it found, the participants who faced the situation, and the suggested changes to the question (see table 2). The complete pipeline is shown in Figure 2. The prompts used in the pipeline are shown in Appendix A.

```
Statement: P1. The US Congress acts in the best
interests of the citizens of the United States.
Problem: Participants like Persona 16 had difficulty
with this statement, expressing uncertainty due to
its generality and potential overlap with political
interests, suggesting ambiguity.
Sources: Interview 16, Participant Sofia Garcia.
Suggestion: The US Congress consistently prioritizes
citizens' best interests over political agendas.
```

Table 2: Example feedback from the reviewer in the pretesting of the Political Trust questionnaire.

Statement	Criterion
The ⟨statement/question⟩ is ambiguous.	Clarity
The ⟨statement/question⟩ is clear.	
The ⟨statement/question⟩ is biased/leading.	Bias
The ⟨statement/question⟩ is impartial.	
The ⟨statement/question⟩ is relevant to ⟨topic⟩.	Relevance
The ⟨statement/question⟩ has no connection to ⟨topic⟩.	
The ⟨statement/question⟩ asks something specific about ⟨topic⟩.	Specificity
The ⟨statement/question⟩ explores broader aspects beyond just ⟨topic⟩.	

Table 3: The evaluation statements used in the Prolific studies. Each criterion is assessed directly and also indirectly for validation. We used 'question' in ⟨statement/question⟩ in the Cross-cultural study and used 'statement' in the Political trust study. Similarly, in ⟨topic⟩, we replaced it with the phrase 'the research question.' in the Cross-cultural study, whereas, in the Political trust study, it was replaced with 'Political trust.'

3.2 Formative Fesability Studies

To assess the feasibility of our pipeline, we conducted two formative studies: one for questionnaire generation and another for pretesting. In the *questionnaire generation* feasibility study, we defined several key parameters: the generation prompt, the inclusion of existing questions from the SQP database, summaries from recent relevant news articles, temperature, and random seed. For each combination of these parameters, we generated a questionnaire and evaluated its *relevance* and *specificity* using the metrics defined by Lei et al. [44]. Our findings revealed that adding more context increased question specificity but often reduced relevance to the original research objective. For *pretesting*, we replicated the 2016 GESIS pretest study on the ICT usage at work module [54] using our LLM-based pilot study simulation. Our pipeline detected issues in 5 out of the 13 tested questions. These results reinforced the feasibility of our questionnaire generation and pretesting pipeline, allowing us to conduct future experiments.

3.3 Evaluation Metrics

3.3.1 Evaluation by Prolific users. To evaluate the LLM-generated and LLM-pretested questions, we asked participants on the Crowdsourcing platform Prolific to rate the initially generated and pretested questions on a Likert scale based on four criteria. We use two criteria defined by Lei et al. [44]: relevance and specificity that measure how relevant is the generated question to a given research question or a topic and how specific the question is within the given general research question, respectively, and use two other common criteria in survey designing: clarity and bias which measure how unambiguous the generated questions are and if the generated questions are leading respectively. For each metric, we used two evaluation statements: one that directly assessed the metric (e.g., "The question is clear.") and another that provided an opposite assessment (e.g., "The question is ambiguous"), both rated by participants on a Likert scale. The evaluation statements for the four criteria are shown in Table 3.

3.3.2 Evaluation by Experts. We also asked domain experts (e.g., faculty members in Political Science and Psychology) who regularly work with survey design to complete the same study as the Prolific participants, i.e., one task to rate generated or pretested questions on the four criteria and the other where they were asked to compare and select the clearer question and provide their reasoning. Then, we analyzed and compared the experts' feedback with the crowdsourcing participants' responses. Expert feedback would provide us with better insights into the effectiveness of LLM-based pretesting.

3.4 Prolific study design

The studies on Prolific were designed on Qualtrics[4]. At the start, the participants were provided with a detailed description of the purpose of the study. The study procedures were approved by Western German University's ethical review board (ID:[24-09-6]). Then, the participants were given descriptions of the four evaluation criteria: for each criterion, they were given a detailed description, an example of how they could rate the questions, and a test question to check their understanding of the criterion. After this, the participants were led to the main tasks of the study. The participants completed two tasks (rating and comparison) in the study. The first task asked participants to rate questions on the four evaluation criteria. The question was randomly selected from the initially LLM-generated or LLM-pretested list and evaluated by the participants on a Likert scale using eight evaluation statements as shown in Table 3. Two evaluation statements corresponded to each criterion to mitigate bias and increase the reliability of the measurement. In the second task, the participants were shown a pair of questions (the initially LLM-generated and the LLM-pretested questions) to compare. They were asked to select the one they thought was more explicit for the potential target audience. After this task, the participants were asked to explain their reasoning for choosing that question. The questions and the evaluation statements were randomized in both tasks to mitigate any order effects. The experts

[4] https://qualtrics.com/

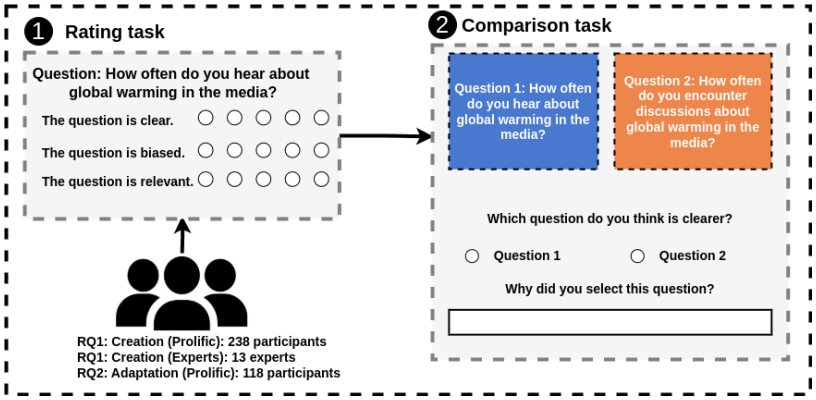


Figure 4: (Simplified) Summary of the tasks in the evaluation studies. (1) In the question rating task, the participants were shown a (randomly selected LLM-generated or LLM-pretested) question and asked to evaluate the question on the four criteria using a 1-5 Likert scale using the 8 evaluation statements shown in Table 3. (2) In the question comparison task, the participants were (randomly) shown the LLM-generated and the LLM-pretested questions side-by-side and asked to select the question they preferred. They were further asked to explain their decision in a sentence.

in the political trust study were also asked to complete the same survey. The tasks are summarized in Figure 4. Example questions from the political trust study are shown in Appendix B.

3.5 Analysis

3.5.1 Quantitative. For the question rating task, we converted the ratings from the eight evaluation statements into our four evaluation criteria ratings using reverse coding. Then, the mean rating for each criterion was computed. We then computed the difference in the mean ratings for the initially LLM-generated and LLM-pretested questions. To check for statistical significance, we computed the paired samples t-test. The complete summary of the quantitative results is shown in the Appendix in Table 8.

3.5.2 Qualitative. For the question selection task, we computed the frequency of each question. The frequency indicated the participants' preference regarding which question they perceived as clearer. For the free text responses, we used GPT-4o to do a thematic analysis, categorizing the responses into five themes (shown in Table 9 in the Appendix). Additionally, we manually coded the expert responses to identify the reasons for selecting or rejecting a question (shown in Appendix C).

4 RQ1: Creation: Automating Questionnaire Generation and Pretesting

In this experiment, we evaluated the complete questionnaire generation and pretesting pipeline using LLMs. We chose "political trust" because of the long-standing debate about the definition of political trust and its measurement.

4.1 RQ1: LLM generated questionnaire on Political trust.

4.1.1 RQ1: Setup. We drew inspiration from the political trust module of the World Values Survey (WVS) Wave 7, 2017-2020[27] to generate the questionnaire on Political trust. The WVS survey features statements under specific questions, rated on a Likert scale. We adopted a similar structure, instructing the LLM to generate statements based on the questions from the WVS political trust module. The WVS survey, being global, includes placeholders in the questions (e.g., *How do you feel about the <name of the national parliament or national congress> in your country?*). We specified the United States as the target country and asked the LLM to generate statements accordingly. The prompt for generating the statements is shown in Appendix A. The resulting questionnaire was then pretested with 50 personas generated for the US. The reviewer LLM identified issues with 10 statements in the questionnaire and proposed changes, as detailed in the Appendix in Table 4.

4.1.2 RQ1: Participants. We used Prolific's screener to recruit participants aged 18 and above from the United States with at least an undergraduate degree in Politics, Psychology, Social Work, or Sociology. A total of 238 participants participated in the study, 61% of which identified as Female. Approximately 41% of the participants were aged 25 to 34, and about 51% of the participants leaned left on the Political spectrum. Additionally, 94% of the participants held at least a University bachelor's degree and reported at least some experience with surveys. About 17% of the participants mentioned being experienced with the World Values Survey (WVS), which served as the basis for our generated questionnaire.
For expert feedback, we contacted professors, researchers, and administrative representatives from various research institutes and requested them to share the study link with their teams. We received responses from 13 experts for our study, 46% of whom identified as Female. About 30% of the experts were aged 25 to 34, and 85% held Graduate degrees. Furthermore, 61% of the experts leaned left on the political spectrum, 54% reported extensive experience with surveys, and 69% had experience with the World Values Survey (WVS).

Statement generated by LLM	Statement after LLM-pretesting	LLM reasoning
The US Congress acts in the best interests of the citizens of the United States.	The US Congress consistently prioritizes citizens' best interests over political agendas.	Participants like Persona 16 had difficulty with this statement, expressing uncertainty due to its generality and potential overlap with political interests, suggesting ambiguity.
The US Congress is effective in passing legislation that benefits the country.	The US Congress efficiently passes legislation with measurable benefits to the country.	Participants such as Persona 15 encountered difficulty deciding on this statement, highlighting the need for clarity or examples of such legislation to guide them, suggesting potential ambiguity.
The US Congress consistently demonstrates integrity in its decisions and actions.	The US Congress consistently upholds ethical standards and transparency in its decision-making process.	The statement is vague and seems too conceptual without concrete examples or context. Integrity can be interpreted differently by respondents, and without a clear metric, it's subjective.
I feel confident that the US Congress will address the pressing issues of today.	I hold a cautious optimism that the US Congress will adequately tackle today's urgent challenges.	Participant 19 (Ethan Clarke) notes uncertainty and lacks confidence due to inadequate information, indicating the statement might benefit from more context or prompting to be more nuanced.
The government's policies are effectively addressing the nation's key problems.	Some government policies effectively address economic, healthcare, and educational challenges.	The phrase "the nation's key problems" is ambiguous as it can be interpreted differently by each respondent, leading to confusion about which problems are being referred to. The response scale does not account for the complexity of addressing multiple issues simultaneously and may not accurately reflect the respondent's views on each issue.

Table 4: The statements generated by the Questionnaire generation pipeline (the left column), the same statements after the changes suggested by the LLM-pretesting pipeline (the middle column), and the reasoning from the LLM for that change (the right column) for the Political trust study (RQ1). The remaining five statements are shown in the Appendix in Table 6.

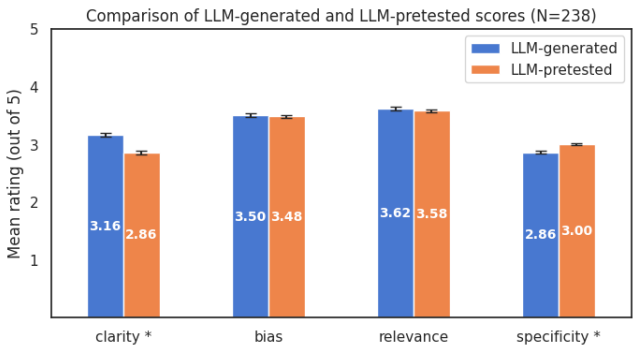


Figure 5: Mean ratings for the LLM-generated and LLM-pretested for the four evaluation criteria (with standard errors) rated by the participants on Prolific for the political trust study. The asterisk (∗) in the label denotes statistical significance ($p < 0.05$).

4.2 RQ1: Results

4.2.1 Prolific participants found LLM-generated statements to be slightly clearer and more specific. Figure 5 presents mean ratings on four dimensions—clarity, bias, relevance, and specificity—for LLM-generated and LLM-pretested by prolific users. Participants rated these two conditions, and we report the results.

For clarity, the mean rating for LLM-generated text was 3.16 (SD = 1.13), while for LLM-pretested text, it was 2.86 (SD = 1.10). A t-test indicated a significant difference ($p < 0.05$), suggesting that the LLM-generated text was perceived as clearer than the LLM-pretested version. For bias, the LLM-generated text received a mean rating of 3.50 (SD = 1.06), while the LLM-pretested text had a mean rating of 3.48 (SD = 1.00). This difference was not statistically significant ($p > 0.1$), indicating that participants did not perceive one version as more biased. For relevance, the LLM-generated text had a mean rating of 3.62 (SD = 1.01), whereas the LLM-pretested text had a mean rating of 3.58 (SD = 0.98). This difference was also not statistically significant ($p > 0.1$), suggesting that both versions were perceived as equally relevant. For specificity, the LLM-generated text received a mean rating of 2.86 (SD = 0.83), while the LLM-pretested text was rated higher at 3.00 (SD = 0.81). This difference was statistically significant ($p < 0.05$), indicating that participants found the LLM-pretested text more specific.

These results suggest a tradeoff between clarity and specificity. **Prolific participants found LLM-generated text was rated as clearer, while LLM-pretested text was seen as more specific.** The lack of significant differences in bias and relevance suggests that pretesting did not alter how participants perceived these aspects of the text.

4.2.2 Experts found LLM-generated statements better than LLM-pretested ones across all metrics. Figure 6 presents the mean ratings on four dimensions—clarity, bias, relevance, and specificity—for LLM-generated and LLM-pretested by experts. Experts rated these two conditions, and we report the findings below.

For clarity, the LLM-generated text received a mean rating of 3.61 (SD = 1.16), while the LLM-pretested text was rated lower at 2.95

(SD = 1.27). A t-test indicated a significant difference ($p < 0.05$), suggesting that experts found the LLM-generated text significantly clearer than the LLM-pretested version. For bias, the mean rating for LLM-generated text was 2.83 (SD = 1.16), while for LLM-pretested text, it was 3.13 (SD = 1.24). This difference was statistically significant ($p < 0.05$), indicating that experts perceived the LLM-pretested text as more biased than the LLM-generated version. For relevance, LLM-generated text had a mean rating of 4.17 (SD = 0.88), while the LLM-pretested text had a slightly lower rating of 3.94 (SD = 1.01). This difference was marginally significant ($p < 0.1$), suggesting that experts found LLM-generated text more relevant. For specificity, the LLM-generated text had a mean rating of 2.98 (SD = 1.07), whereas the LLM-pretested text had a slightly lower mean of 2.72 (SD = 0.94). T-test confirmed a significant difference ($p < 0.05$), indicating that experts rated the LLM-generated text as more specific than the LLM-pretested text.

Overall, these results suggest that **experts found LLM-generated text to be significantly clearer, more relevant, and more specific, while also perceiving LLM-pretested text as more biased**. Unlike the findings from the Prolific study, where specificity improved after pretesting, experts rated LLM-generated text higher across most dimensions except for bias, where pretested text was perceived as more biased.

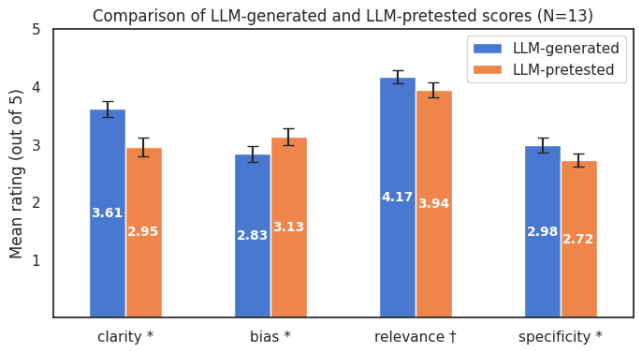
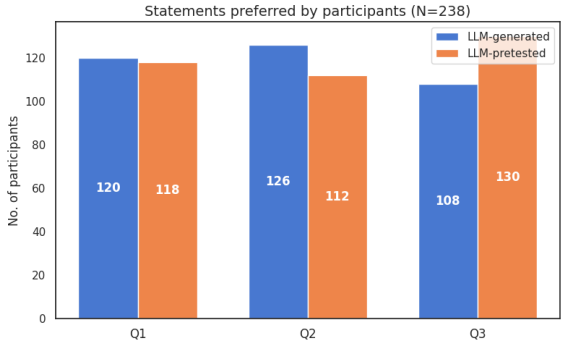
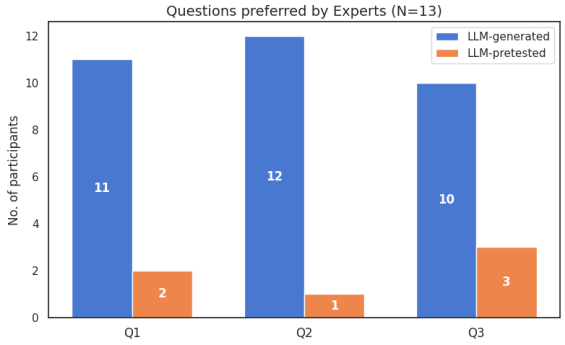


Figure 6: Mean ratings for the LLM-generated and LLM-pretested for the four evaluation criteria (with standard errors) rated by the experts for the political trust study. The asterisk (∗) in the label denotes statistical significance ($p < 0.05$) and the dagger (†) denotes a marginal statistical significance ($p < 0.1$).

4.2.3 Participants showed mixed preferences, whereas Experts favored the LLM-generated statements. Figure 7 presents the distribution of question preferences for LLM-generated and LLM-pretested statements among participants (Figure 7a top) and experts (Figure 7b bottom). Among participants (N=238), preferences were relatively balanced across conditions, with no strong preference for either LLM-generated or LLM-pretested questions. However, among experts (N=13), there was a clear preference for LLM-generated statements, with a majority selecting them over the pretested versions for all three questions. This contrast suggests that while general participants found both versions similarly acceptable, experts perceived the LLM-generated statements as preferable. These findings align with previous results indicating that experts rated LLM-generated content as clearer and more relevant than LLM-pretested versions.



(a) The question preferred by the participants.



(b) The question preferred by the experts.

Figure 7: Comparison of question preferences between participants and experts. Each question (Q1-Q3) includes some statements. Experts evaluated each question with changed and unchanged statements to get the complete context. In contrast, the Prolific participants were shown only the statements that were changed during pretesting to reduce mental load.

From open-ended responses, prolific participants preferred clarity and simplicity over excessive detail, emphasizing neutrality and relevance in survey statements. Participants overwhelmingly favored LLM-generated statements for their concise and direct phrasing, with one noting, *LLM-generated is preferable because it is more concise. Although LLM pre-tested is a better question, it is not as clear, and too many similar type questions may lead the participant to become fatigued.* However, some responses indicated that LLM pre-tested statements offered more precision, as one participant pointed out, *LLM-generated questions are far less specific than LLM pre-tested questions.* Bias emerged as a concern, particularly with leading or overly complex phrasing. One participant noted, *Both statements in LLM pre-tested are too convoluted regarding the adjectives used to describe various nouns, making the statements a little leading.* Relevance also played a role, particularly when discussing government actions. Participants preferred specific references over general

statements, as one participant stated, *It mentions the government's record. LLM-generated just says actions, while LLM pre-tested gives specifics.* Overall, participants valued clarity and neutrality while recognizing the trade-off between simplicity and detail in survey design.

From open-ended responses, experts preferred clarity and simplicity over excessive specificity, highlighting the trade-off between readability and precision in survey design. Experts generally favored LLM-generated statements for their shorter, more understandable phrasing, with one noting, *"Shorter sentences -> more understandable,"* and another emphasizing, *"Simple and straightforward statements."* Readability was a key concern, particularly for a general audience, as one expert pointed out, *"The reading age for Group 2 (LLM pre-tested) questions is too high. Also, all the Group 2 green questions ask about two things simultaneously, so aren't clear"*. While LLM pre-tested statements were recognized as more precise, some experts found them too complex or difficult to interpret. One expert stated, *"Although the statements in Group 2 seem more precise, they are more difficult to answer for the general public because they are too specific (for example, 'public welfare and policy outcomes' instead of 'interests of the public')."* Another highlighted a trade-off, noting that specifics are reasonable, but complex wording and excessive detail made certain statements unnecessarily complicated. Overall, experts valued clarity and accessibility while acknowledging the importance of specificity when it did not hinder comprehension.

5 RQ2: Adapting questionnaires for cross-cultural surveys

In this experiment, we looked at the potential of Large Language Models in adapting existing questionnaires for a specific target audience. We were interested in determining if a questionnaire on climate change tested in the United States could be adapted to an African country using our pretesting pipeline with LLMs. We decided on South Africa as the target audience due to the high number of available users on Prolific.

5.1 RQ2: LLM adapted US-specific questionnaire for South African audience.

5.1.1 RQ2: Setup. We obtained a standardized survey from a study focused on U.S. state-level public opinion about climate change from 2008 to 2020[52]. Thirty questions assess public perceptions of global warming, including beliefs about its existence, human causes, potential harm, and policy support.[5] The questions were extracted and sanitized, i.e., correctly formatting for the LLM, e.g., replacing the United States with South Africa. Then, we used an LLM to summarize the questionnaire as a research question. Lastly, using the questionnaire generation pipeline (section 3.1.1), we generated a list of South African participants' personas based on the research question. We disregarded the generated questionnaire because we already had the standardized version, and our focus in this experiment was on the generated personas. We then used the list of personas in our pretesting pipeline to evaluate the US questionnaire and determine necessary changes for a South African audience. All 30 questions were pretested, but the LLM suggested changes to only five questions, some of which are shown in Table 5.

5.1.2 RQ2: Participants. We sampled participants aged 18 and above from South Africa who had at least an undergraduate degree in subjects such as Earth, Environment or Climate Sciences, Politics, Psychology, Social Work, or Sociology. A total of 118 participants participated in the study, 75% of which identified as female. About 31% of the participants defined themselves as leaning toward the political center-left, and 99% mentioned having at least some experience with surveys.

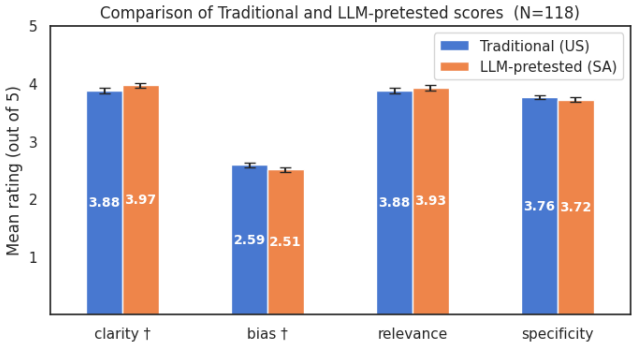


Figure 8: Mean ratings (with standard errors) for the original US-based (traditional) questionnaire and the questionnaire adapted for the South African (SA) audience after LLM-based pretesting. The dagger (†) symbol denotes a marginal statistical significance ($p < 0.1$).

5.2 RQ2: Results

Figure 8 presents ratings on four dimensions—clarity, bias, relevance, and specificity. Participants rated the traditional standardized and LLM-pretested content on our scale above, and we analyzed their differences.

For clarity, the traditional questions received a mean rating of 3.87 (SD = 1.12), while the pretested questions were rated slightly higher at 3.97 (SD = 1.07). A t-test indicated a marginally significant difference ($p < 0.1$), suggesting that participants found the pretested questions somewhat clearer than the traditional ones. For bias, the traditional questions had a mean rating of 2.58 (SD = 1.01), while the pretested questions were rated slightly lower at 2.51 (SD = 1.03). This difference was also marginally significant ($p < 0.1$), indicating that pretesting slightly reduced perceived bias in the questions. For relevance, the traditional questions had a mean rating of 3.88 (SD = 1.13), while the pretested questions received a similar rating of 3.93 (SD = 1.11). This difference was not statistically significant ($p > 0.1$), suggesting that pretesting did not notably impact how relevant participants found the questions. For specificity, the traditional questions received a mean rating of 3.77 (SD = 0.87), whereas the pretested questions were rated slightly lower at 3.72 (SD = 0.93). A t-test confirmed that this difference was not significant ($p > 0.1$), indicating that participants found both traditional and pretested questions similarly specific.

Overall, these results suggest that **Prolific participants found LLM-adapted questions slightly clearer and less biased than**

[5] https://climatecommunication.yale.edu/visualizations-data/ycom-us/

Question from the standard survey	Question after LLM-pretesting	LLM reasoning
Assuming global warming is happening, do you think it is...? Options: Caused mostly by human activities, Caused mostly by natural changes in the environment, None of the above because global warming isn't happening, Other, Don't know	Assuming global warming is happening, what do you believe is the primary cause? Options: Caused mostly by human activities, Caused mostly by natural changes in the environment, Don't know, Other (please specify)	This question can be problematic because it includes an option ("None of the above because global warming isn't happening") that contradicts the assumption made in the question. This might confuse the respondents or lead to inconsistencies in their answers. It's important that the question aligns with the premise. Also, the option "Other" seems vague, which can lead to confusion about what respondents should include under this category.
How often do you discuss global warming with your friends and family? Options: Often, Occasionally, Rarely, Never	How often do you have in-depth discussions about global warming with your friends and family? Options: Often, Occasionally, Rarely, Never	The question lacks a clear delineation of what constitutes a discussion about global warming. This can result in variation among participants in terms of what intensity or depth qualifies as a discussion.

Table 5: The original questions, the questions after making the changes suggested by the LLM, and the reasoning given by the LLM for the changes. The remaining questions are shown in the Appendix in Table 7.

traditional questions, though differences were minor. No significant changes were observed for relevance or specificity, indicating that pretesting had minimal impact on these aspects.

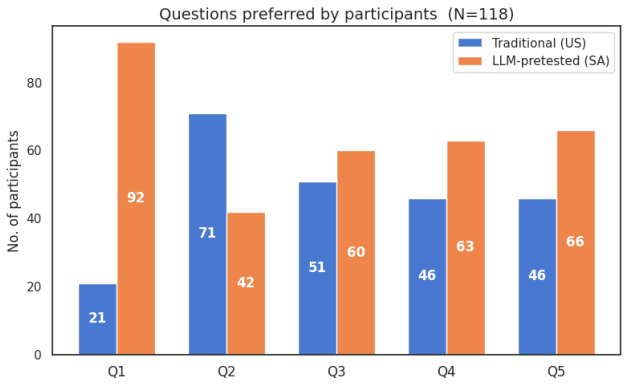


Figure 9: The results of the questionnaire comparison task where the participants were given the original and pretested questions side-by-side and asked to select the question they thought was clearer. We can see that in 4 out of the 5 pairs of questions tested, the participants selected the pretested question as being clearer and easier to understand.

For the comparison task, participants picked the pre-tested question for 4 out of 5 in this task, as shown in Figure 9. From open-ended responses, participants preferred LLM pre-tested to the traditional for its clarity, specificity, and conciseness, emphasizing the need for precise yet accessible wording in survey design. Participants found LLM-pretested clear as it conveyed the intended meaning transparently. One participant stated *"LLM-pretested is clearer because it is more comprehensive and detailed,"* and another said, *"Straight to the point, very clear."* Specificity was another key factor, as LLM-pretested made answering questions easier. A participant pointed out, *".. asking for a specific answer in terms of contextualizing the extent to which one discusses global warming"*. However, a few participants raised concerns about bias in phrasing, noting that terms like *"in-depth"* (Q2) could introduce unintended assumptions, with one stating, *"Because Q2 is biased by the word 'in-depth'"*. Participants also preferred simple language, which helped them process easier, *"Because it is written in simple English, no big words."* Overall, participants valued clear, specific, and concise language while being mindful of potential bias in question phrasing.

6 Discussion

In this work, we explored the potential of Large Language Models in generating and pretesting survey questionnaires. Compared to existing research, we focused primarily on generating specific questions related to the research question in the questionnaire and using a simulated pilot study to pretest the generated questionnaire. Furthermore, we conducted multiple Prolific studies to evaluate the LLM-generated and LLM-pretested content. Below, we reflect on the results and what they mean regarding the use of LLMs in survey research.

6.1 Empirical Insights on the Suitability and Challenges of LLMs in Survey Design

6.1.1 RQ1: LLM-generated questions were more favorable than LLM-pretested for experts and Prolific participants. In our first experiment (see Section 4), we compared LLM-generated statements with LLM-pretested statements. Building on prior work [44], we used two additional evaluation metrics—bias and clarity—assessed by human evaluators. From the mean rating scores presented in Figure 5 (238 Prolific participants) and Figure 6 (13 experts), we found that LLM-generated statements were clear, slightly biased, and highly relevant compared to the LLM-pretested ones. Overall, both experts and participants favored LLM-generated over LLM-pretested questionnaires. Open-ended responses from experts revealed that pretesting made the questions more wordy and complex which can affect comprehensibility and data quality as suggested by Lenzner [46]. As a result, in this scenario, pretesting did not improve the generated questionnaires. In fact, in one case, expert response suggested that pretesting even affected the meaning of the statements (*Questions that I would answer with yes in LLM-generated would*

be 'no' in LLM-pretested). We suspect that the vast training corpus of LLMs, which likely includes standardized questionnaires [64], enables them to generate high-quality questions without additional refinement through pretesting. Therefore, future research can utilize our pipeline by examining various topics, domains, and models to assess the pipeline's suitability for particular domains or topics.

6.1.2 RQ2: LLM-adapted questions were more favorable than traditional ones for the South African audience. In our second experiment (see Section 5), we evaluated our LLM-based pretesting pipeline to adapt a U.S.-based questionnaire for a South African audience. Figure 8 presents the mean ratings for both the standardized and pretested questionnaires. Participants generally found the original U.S.-based questionnaire clear, relevant, slightly biased, and specific. Pretesting led to slight improvements in clarity and reduced bias, indicating that participants perceived the pretested questionnaire as better suited for a South African audience. In the comparison task, 4 out of 5 LLM-pretested questions (figure 9) were preferred over the traditional questionnaire, as the longer pretested questions provided better content clarification. However, specific keywords added by the LLM, such as "in-depth" (1/5), reduced clarity or introduced confusion. Participants also noted that longer wording made the questions harder to read. Our findings suggest that LLMs can enhance survey design when adapting questionnaires across geographic regions. Future research could explore how culture, language, and regional differences influence survey adaptation and effectiveness.

6.2 Pipeline for using LLMs for Generation, Pre-testing, and Adaptation.

We developed an innovative pipeline that uses LLMs to support the survey questionnaire design process. Building upon the works by Laraspata et al. [43], Lei et al. [44], Rothschild et al. [64], we expanded their work using Retrieval Augmented Generation (RAG) to enhance questionnaire generation. Given a research question, our pipeline generates questionnaires by incorporating existing standardized questions and relevant news articles. Additionally, it creates a set of fictitious personas that can be used to simulate a pilot study and provide feedback on the generated questions, potentially improving the questions. Our pipeline also enables the adaptation of existing questionnaires to different contexts, such as tailoring a survey initially designed for the U.S. to suit the South African public. We developed and deployed human evaluation metrics to assess the quality of questions generated by our pipeline. Our evaluation is based on four key criteria: relevance and specificity (adapted from Lei et al. [44]), along with two widely used criteria, clarity, and bias. To enhance reliability, we introduced a paired evaluation approach, where each criterion is assessed directly and indirectly. For transparency and reproducibility, we have shared our prompts, LLM-generated feedback, and evaluation questions (see the Appendix). Additionally, our code is available publicly for researchers to use and improve upon[6]. Our pipeline is flexible and, thus, can be easily extended or combined with existing LLMs and applied to new domains. Our pipeline could benefit survey researchers seeking quick feedback on their questionnaire drafts, thereby reducing the need for multiple human pretesting iterations. We expect our pipeline and evaluation metrics can contribute to advancing the relatively sparse research on LLM-enhanced survey design.

[6]https://github.com/Societal-Computing/automated-questionnaire-pretesting

6.3 Better Questionnaire design: Balancing Clarity, Specificity, and Neutrality?

Through open-ended responses from experts and participants, we found that effective questionnaires must balance clarity, specificity, and neutrality to improve data collection. Clarity is essential, as overly complex wording can reduce readability and increase cognitive burden [38]. At the same time, specificity is crucial—questions that are too brief, overly simplistic, or filled with jargon may lack precision, ultimately affecting readability and comprehension [75]. This highlights the need to manage the tradeoff between readability and specificity carefully. Additionally, bias can be introduced through vague word choices, such as "in-depth," which may unintentionally influence responses [46]. This underscores the importance of using neutral language to maintain objectivity in survey design. Participant feedback is key in refining our pipeline and generating survey questions that better align with new target audiences. Future work could extend our pipeline by integrating open-ended input from users, enabling iterative improvements—such as refining prompts based on human participant feedback—to enhance question quality and relevance further.

7 Limitations and Future work

While our analysis shows the potential of LLMs in survey research, there are limitations to the approach we used in our experimentations. First, we only focused on generating simple questionnaires when existing questionnaires often have complex flow and branching logic. We generated closed-ended and open-ended questions but didn't include other types of questions, such as matrix-style questions. We also didn't include demographics-related questions in the generated questionnaires. Furthermore, we didn't evaluate the overall quality of the questionnaire (e.g., coherence, logical flow, etc). Pretesting is inherently an iterative process; however, in our experiments, we limit our analysis to the initial iteration and do not implement subsequent rounds using the revised questionnaire. Establishing a definitive stopping criterion for such iterations proves challenging, as contemporary large language models (LLMs) tend to consistently comply with prompts and continue suggesting modifications in each cycle. However, LLM-generated content may exert an anchoring effect, particularly on individuals from the general public with limited experience [64], as the output can appear highly plausible. Therefore, we recommend that our proposed pipeline be used as a complementary tool within established survey design processes for researchers, and as an educational resource for those learning about survey methodology.

Similarly, in the study design, we selected participants with relevant domain knowledge and survey experience, providing them with detailed explanations of the four evaluation criteria, illustrative examples, and test questions to assess their understanding. Nevertheless, we cannot be certain that participants fully internalized the criteria or applied them as intended during their evaluations. For instance, in the question preference task within the Political Trust study, many participants appeared to base their preferences

on personal opinions about the content of the statements ('*I selected that because the government clearly does not fully execute all their promises.*') rather than on an objective assessment of sentence structure and clarity. Likewise, multiple elements, including the duration of the study and participant biases, can influence self-reported measurements such as those in our studies. Assessing these questions is cognitively demanding, and participants may hasten through the latter sections of the study, particularly the question preference and open-response tasks. This could impact the reliability of self-reported evaluations. Consequently, researchers should exercise caution when employing crowdsourced participants from platforms such as Prolific. Ideally, a more representative sample should be used, and participants' overarching ideological orientations and potential biases should be assessed before the evaluation of survey items. Moreover, the sample size of experts (13) was small compared to participants on Prolific and while the difference in preference was more pronounced, the results might not be replicable with a larger sample size. Such a small sample size could compromise the statistical significance and increase the risk of sampling bias.

Additionally, LLMs have biases based on their training and finetuning data, which could affect the generated questionnaires [24, 29, 50, 82] in various domains. For example, LLMs are shown to have a center-left leaning [65], which could affect questionnaire generation and pretesting in the Political domain. Similarly, LLMs can reinforce stereotypes, and misrepresent underrepresented groups [78]. Moreover, in the context of low-resource languages, LLMs tend to exhibit reduced performance, primarily due to the limited availability of training data [12]. Thus, further evaluation of the generated questionnaires should be done to determine such bias seeping into the generated questions.

Future research could expand upon our work to create complex-structured questionnaires with diverse question types along with specific demographic questions tailored to the research question, enabling researchers to gather targeted information about their audience. Depending on the research question, questionnaires could incorporate images or even video media. With the rising popularity of Vision-Language-based LLMs, developing multi-modal survey instruments could be a promising goal. Further refining the questionnaire generation prompt with more examples could enhance results by specifying how to effectively utilize the provided context, such as determining the number of questions to generate from the given news summary, determining the number of questions to reuse or modify from existing questionnaires, and determining the amount of focus to put on the research question. The feedback from experts in our study provided valuable criteria for designing questions, which could inform better questionnaire generation prompts. Fine-tuning LLMs with existing pretesting reports could help develop LLMs that are better capable of detecting issues with the generated questionnaires and offering better suggestions. Future work could explore designing interviewer agents capable of implementing Cognitive Interviewing techniques effectively in Pilot study simulations. Creating LLM personas that respond diversely, like humans, could make pretesting results more realistic. In our pipeline, we provided the reviewer LLM only with the interview transcripts to identify questionnaire issues. This could be supplemented by using the reviewer for behavioral coding and annotating the interview transcripts for further analysis. Similarly, including more examples in the reviewer's prompt could make the recommendations more specific. Furthermore, our work can be expanded to low-resource languages or other domains by using LLMs trained in specific languages or domains that could be particularly effective in generating relevant questions or simulating participants from those contexts, helping to address cultural nuances that commercial LLMs may overlook.

8 Conclusion

In conclusion, our work demonstrated the potential of Large Language Models (LLMs) in enhancing the design and pretesting of survey questionnaires. Our work contributed empirical evidence from two Prolific studies, highlighting both the strengths and challenges of using LLMs in survey design. We also introduced a novel LLM-based pipeline for automated questionnaire generation and pretesting, which can be adapted across various domains and LLMs. Our findings indicate that LLMs, such as GPT-4o, with relevant contextual information, can effectively generate relevant and specific questions, reducing the need for extensive human pretesting. However, while LLMs showed promise in creating high-quality questions, their ability to pretest survey questionnaires remained limited, suggesting that human oversight is still necessary to ensure question clarity and relevance. For adapting existing questionnaires to specific target audiences, our results revealed that LLMs can make questionnaires more explicit, unbiased, and relevant. This capability is particularly valuable in cross-cultural studies, where nuanced adaptations are crucial for accurate data collection.

While LLMs offer significant advantages, it is essential to acknowledge their limitations. Future research should focus on refining LLM capabilities to better handle the subtleties of questionnaire design and pretesting, ensuring that surveys remain a reliable tool for gathering meaningful insights across diverse populations.

Acknowledgments

We want to acknowledge the members of the Interdisciplinary Institute for Societal Computing at Saarland University for helping with the design and testing of the study experiments. Divya Mani Adhikari, Vikram Kamath Cannanure, and Ingmar Weber are supported by funding from the Alexander von Humboldt Foundation and its founder, the German Federal Ministry of Education and Research.

References

[1] Noorhan Abbas, Thomas Pickard, Eric Atwell, and Aisha Walker. 2021. University Student Surveys Using Chatbots: Artificial Intelligence Conversational Agents. In *Learning and Collaboration Technologies: Games and Virtual Environments for Learning*, Panayiotis Zaphiris and Andri Ioannou (Eds.). Springer International Publishing, Cham, 155–169. doi:10.1007/978-3-030-77943-6_10

[2] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (July 2023), 337–351. doi:10.1017/pan.2023.2

[3] Dorcas E. Beaton, Claire Bombardier, Francis Guillemin, and Marcos Bosi Ferraz. 2000. Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures. *Spine* 25, 24 (2000). https://journals.lww.com/spinejournal/fulltext/2000/12150/guidelines_for_the_process_of_cross_cultural.14.aspx

[4] Dorcas E. Beaton, Claire Bombardier, Francis Guillemin, and Marcos Bosi Ferraz. 2000. Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures. *Spine* 25, 24 (Dec. 2000), 3186. https://journals.lww.com/spinejournal/pages/articleviewer.aspx?year=2000&issue=12150&article=00014&type=Fulltext

[5] Dorothée Behr, Michael Braun, and Luisa Aiglstorfer. 2024. Showcasing the usefulness of web probing: Do subtle variations in questionnaire translation lead to different survey responding? *Quality & Quantity* (March 2024). doi:10.1007/s11135-024-01843-8

[6] Dorothée Behr, Katharina Meitinger, Michael Braun, and Lars Kaczmirek. 2017. *Web probing – implementing probing techniques from cognitive, interviewing in web surveys with the goal to assess the validity of survey questions (GESIS Survey Guidelines)Web probing – implementing probing techniques from cognitive, interviewing in web surveys with the goal to assess the validity of survey questions (GESIS Survey Guidelines).* Technical Report. GESIS - Leibniz Institute for the Social Sciences. doi:10.15465/GESIS-SG_EN_023 Version Number: 1.0.

[7] Johnny Blair, Allison Ackermann, Linda Piccinino, and Rachel Levenstein. 2007. Using Behavior Coding to Validate Cognitive Interview Findings. (Jan. 2007).

[8] Johnny Blair and Frederick G. Conrad. 2011. Sample Size for Cognitive Interview Pretesting. *The Public Opinion Quarterly* 75, 4 (2011), 636–658. https://www.jstor.org/stable/41288411 Publisher: American Association for Public Opinion Research.

[9] Johnny Blair and Linda Piccinino. 2005. The development and testing of instruments for cross-cultural and multi-cultural surveys. In *Methodological aspects in cross-national research*, Jürgen H. P. Hoffmeyer-Zlotnik and Janet Harkness (Eds.). ZUMA-Nachrichten Spezial, Vol. 11. GESIS-ZUMA, Mannheim, 13–30.

[10] Petra M Boynton and Trisha Greenhalgh. 2004. Selecting, designing, and developing your questionnaire. *BMJ : British Medical Journal* 328, 7451 (May 2004), 1312–1315. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC420179/

[11] Zi Chai and Xiaojun Wan. 2020. Learning to Ask More: Semi-Autoregressive Sequential Question Generation under Dual-Graph Interaction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 225–237. doi:10.18653/v1/2020.acl-main.21

[12] Roberts Dar̂gis, Guntis Bārzdiņš, Inguna Skadiņa, Normunds Grūzītis, and Baiba Saulīte. 2024. Evaluating Open-Source LLMs in Low-Resource Languages: Insights from Latvian High School Exams. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni (Eds.). Association for Computational Linguistics, Miami, USA, 289–293. doi:10.18653/v1/2024.nlp4dh-1.28

[13] Elizabeth Dean, Brian Head, and Jodi Swicegood. 2013. Virtual Cognitive Interviewing Using Skype and Second Life. In *Social Media, Sociality, and Survey Research*. 107–132. doi:10.1002/9781118751534.ch5 Journal Abbreviation: Social Media, Sociality, and Survey Research.

[14] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* 27, 7 (July 2023), 597–600. doi:10.1016/j.tics.2023.04.008 Publisher: Elsevier.

[15] Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the Survey Responses of Large Language Models. doi:10.48550/arXiv.2306.07951 arXiv:2306.07951 [cs].

[16] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). arXiv:2401.08281 [cs.LG]

[17] Jonathan Drennan. 2003. Cognitive interviewing: verbal data in the design and pretesting of questionnaires. *Journal of Advanced Nursing* 42, 1 (2003), 57–63. doi:10.1046/j.1365-2648.2003.02579.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2648.2003.02579.x

[18] Jonathan Epstein, Richard H. Osborne, Gerald R. Elsworth, Dorcas E. Beaton, and Francis Guillemin. 2015. Cross-cultural adaptation of the Health Education Impact Questionnaire: experimental study showed expert committee, not back-translation, added value. *Journal of Clinical Epidemiology* 68, 4 (2015), 360–369. doi:10.1016/j.jclinepi.2013.07.013

[19] Jonathan Epstein, Ruth Miyuki Santo, and Francis Guillemin. 2015. A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. *Journal of Clinical Epidemiology* 68, 4 (April 2015), 435–441. doi:10.1016/j.jclinepi.2014.11.021

[20] Jennifer Fei, Jessica Wolff, Michael Hotard, Hannah Ingham, Saurabh Khanna, Duncan Lawrence, Beza Tesfaye, Jeremy M. Weinstein, Vasil I. Yasenov, and Jens Hainmueller. 2022. Automated Chat Application Surveys Using Whatsapp: Evidence from Panel Surveys and a Mode Experiment. doi:10.2139/ssrn.4114839

[21] Yifan Gao, Piji Li, Irwin King, and Michael R. Lyu. 2019. Interconnected Question Generation with Coreference Alignment and Conversation Flow Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 4853–4862. doi:10.18653/v1/P19-1480

[22] Yubin Ge, Ziang Xiao, Jana Diesner, Heng Ji, Karrie Karahalios, and Hari Sundaram. 2023. What should I Ask: A Knowledge-driven Approach for Follow-up Questions Generation in Conversational Surveys. doi:10.48550/arXiv.2205.10977

[23] Linn Gjersing, John RM Caplehorn, and Thomas Clausen. 2010. Cross-cultural adaptation of research instruments: language, setting, time and statistical considerations. *BMC Medical Research Methodology* 10, 1 (10 Feb 2010), 13. doi:10.1186/1471-2288-10-13

[24] Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. doi:10.48550/arXiv.2311.04892 arXiv:2311.04892 [cs].

[25] Melissa Guyre, Liz Holland, Nirva Shah, and Rahul R. Divekar. 2024. Prompt Engineering an LLM into Roleplaying a Management Coach: a Short Guide by and for Non-NLP Experts. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3640794.3665570

[26] Otso Haavisto and Robin Welsch. 2024. Questionnaires for Everyone: Streamlining Cross-Cultural Questionnaire Adaptation with GPT-Based Translation Quality Evaluation. doi:10.48550/arXiv.2407.20608 arXiv:2407.20608 [cs].

[27] Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2020. World Values Survey wave 7 (2017-2020) cross-national data-set.

[28] Anne-Wil Harzing. 2005. Does the Use of English-language Questionnaires in Cross-national Research Obscure National Differences? *International Journal of Cross Cultural Management* 5, 2 (Aug. 2005), 213–224. doi:10.1177/1470595805054494 Publisher: SAGE Publications.

[29] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature* 633, 8028 (01 Sep 2024), 147–154. doi:10.1038/s41586-024-07856-5

[30] John J. Horton. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? doi:10.48550/arXiv.2301.07543 arXiv:2301.07543 [econ, q-fin].

[31] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3544548.3580688

[32] Sviatlana Höhn, Jauwairia Nasir, Ali Paikan, Pouyan Ziafati, and Elisabeth André. 2024. Using Large Language Models for Robot-Assisted Therapeutic Role-Play: Factuality is not enough!. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3640794.3665886

[33] Nashwa Ismail, Gary Kinchin, and Julie-Ann Edwards. 2018. Pilot study, Does it really matter? Learning lessons from conducting a pilot study for a qualitative PhD thesis. *International Journal of Social Science Research* 6, 1 (March 2018), 1–17. doi:10.5296/ijssr.v6i1.11720 Num Pages: 17 Number: 1.

[34] Bernard J. Jansen, Soon-gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research. *Natural Language Processing Journal* 4 (Sept. 2023), 100020. doi:10.1016/j.nlp.2023.100020

[35] Stephen Jenkins and Tony Solomonides. 2000. Automating Questionnaire Design and Construction. *International Journal of Market Research* 42, 1 (Jan. 2000), 1–13. doi:10.1177/147078530004200106

[36] Ng Chirk Jenn. 2006. Designing A Questionnaire. *Malaysian Family Physician : the Official Journal of the Academy of Family Physicians of Malaysia* 1, 1 (April 2006), 32–35. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4797036/

[37] George A. Johanson and Gordon P. Brooks. 2010. Initial Scale Development: Sample Size for Pilot Studies. *Educational and Psychological Measurement* 70, 3 (June 2010), 394–400. doi:10.1177/0013164409355692 Publisher: SAGE Publications Inc.

[38] Kendra Kamp, Gwen Wyatt, Sharon Dudley-Brown, Kelly Brittain, and Barbara Given. 2018. Using cognitive interviewing to improve questionnaires: An exemplar study focusing on individual and condition-specific factors. *Applied Nursing Research* 43 (Oct. 2018), 121–125. doi:10.1016/j.apnr.2018.06.007

[39] Junsol Kim and Byungkyu Lee. 2023. AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. https://arxiv.org/abs/2305.09620v2

[40] Sunwoong Kim, Jongho Jeong, Jin Soo Han, and Donghyuk Shin. 2024. LLM-Mirror: A Generated-Persona Approach for Survey Pre-Testing. arXiv:2412.03162 [cs.CY] https://arxiv.org/abs/2412.03162

[41] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300316

[42] Bärbel Knäuper, Robert Belli, D. Hill, and A Herzog. 1997. Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality. *Journal of Official Statistics* 13 (Jan. 1997).

[43] Lucrezia Laraspata, Fabio Cardilli, Giovanna Castellano, and Gennaro Vessio. [n. d.]. Enhancing Human Capital Management through GPT-driven Questionnaire Generation. ([n. d.]).

[44] Yan Lei, Liang Pang, Yuanzhuo Wang, Huawei Shen, and Xueqi Cheng. 2024. Qsnail: A Questionnaire Dataset for Sequential Question Generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*

Resources and Evaluation (LREC-COLING 2024), Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 13407–13418. https://aclanthology.org/2024.lrec-main.1174/

[45] Timo Lenzer, Patricia Hadler, and Cornelia Neuert. 2024. Cognitive Pretesting.

[46] Timo Lenzner. 2012. Effects of Survey Question Comprehensibility on Response Quality. *Field Methods* 24, 4 (Nov. 2012), 409–428. doi:10.1177/1525822X12448166 Publisher: SAGE Publications Inc.

[47] Timo Lenzner and Cornelia E. Neuert. 2017. Pretesting Survey Questions Via Web Probing – Does it Produce Similar Results to Face-to-Face Cognitive Interviewing? *Survey Practice* 10, 4 (Oct. 2017). doi:10.29115/SP-2017-0020

[48] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[49] Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable Text Generation for Large Language Models: A Survey. arXiv:2408.12599 [cs.CL] https://arxiv.org/abs/2408.12599

[50] Jieli Liu and Haining Wang. 2024. Assessing Gender and Racial Bias in Large Language Model-Powered Virtual Reference. *Proceedings of the Association for Information Science and Technology* 61, 1 (2024), 576–580. doi:10.1002/pra2.1061 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.1061

[51] Antonio Maiorino, Zoe Padgett, Chun Wang, Misha Yakubovskiy, and Peng Jiang. 2023. 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 (CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 5244–5245. doi:10.1145/3583780.3615506

[52] Jennifer R. Marlon, Xinran Wang, Parrish Bergquist, Peter D. Howe, Anthony Leiserowitz, Edward Maibach, Matto Mildenberger, and Seth Rosenthal. 2022. Change in US state-level public opinion about climate change: 2008–2020. *Environmental Research Letters* 17, 12 (2022). doi:10.1088/1748-9326/aca702

[53] Catriona Rachel Mayland, Christina Gerlach, Katrin Sigurdardottir, Marit Irene Tuen Hansen, Wojciech Leppert, Andrzej Stachowiak, Maria Krajewska, Eduardo Garcia-Yanneo, Vilma Adriana Tripodoro, Gabriel Goldraij, Martin Weber, Lair Zambon, Juliana Nalin Passarini, Ivete Bredda Saad, John Ellershaw, and Dagny Faksvåg Haugen. 2019. Assessing quality of care for the dying from the bereaved relatives' perspective: Using pre-testing survey methods across seven countries to develop an international outcome measure. *Palliative Medicine* 33, 3 (2019), 357–368. doi:10.1177/0269216318818299 arXiv:https://doi.org/10.1177/0269216318818299 PMID: 30628867.

[54] K Meitinger, C Neuert, C Beitz, and N Menold. 2016. Pretesting of special module on ICT at work, working conditions & learning digital skills.

[55] Felix Ndashimye, Oumarou Hebie, and Jasper Tjaden. 2024. Effectiveness of WhatsApp for Measuring Migration in Follow-Up Phone Surveys. Lessons from a Mode Experiment in Two Low-Income Countries during COVID Contact Restrictions. *Social Science Computer Review* 42, 2 (April 2024), 460–479. doi:10.1177/08944393221111340 Publisher: SAGE Publications Inc.

[56] Cornelia Neuert. 2023. Design of multiple open-ended probes in cognitive online pretests using web probing. (2023). doi:10.13094/SMIF-2023-00005 Publisher: FORS/PUMA/GESIS.

[57] Francisco Olivos and Minhui Liu. 2024. ChatGPTest: opportunities and cautionary tales of utilizing AI for questionnaire pretesting. doi:10.48550/arXiv.2405.06329

[58] Kristen Olson. 2010. An Examination of Questionnaire Evaluation by Expert Reviewers. *Field Methods* 22, 4 (2010), 295–318. doi:10.1177/1525822X10379795 arXiv:https://doi.org/10.1177/1525822X10379795

[59] Kristen Olson. 2010. An Examination of Questionnaire Evaluation by Expert Reviewers. *Field Methods* 22, 4 (Nov. 2010), 295–318. doi:10.1177/1525822X10379795 Publisher: SAGE Publications Inc.

[60] Jessica Parker, Veronica Richard, and Kimberly Becker. 2023. Flexibility & Iteration: Exploring the Potential of Large Language Models in Developing and Refining Interview Protocols. *The Qualitative Report* 28, 9 (Sept. 2023), 2772–2791. doi:10.46743/2160-3715/2023.6695

[61] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 784–789. doi:10.18653/v1/P18-2124

[62] Pooja Rao S B, Manish Agnihotri, and Dinesh Babu Jayagopi. 2021. Improving Asynchronous Interview Interaction with Follow-up Question Generation. (March 2021). doi:10.9781/ijimai.2021.02.010 Accepted: 2022-04-25T09:02:55Z Publisher: International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI).

[63] Nina Reynolds, Adamantios Diamantopoulos, and Bodo Schlegelmilch. 1993. Pre-Testing in Questionnaire Design: A Review of the Literature and Suggestions for Further Research. *Market Research Society. Journal.* 35, 2 (March 1993), 1–11. doi:10.1177/147078539303500202 Publisher: SAGE Publications.

[64] David M. Rothschild, James Brand, Hope Schroeder, and Jenny Wang. 2024. Opportunities and risks of LLMs in survey research. doi:10.2139/ssrn.5001645

[65] David Rozado. 2024. The political preferences of LLMs. *PLOS ONE* 19, 7 (07 2024), 1–15. doi:10.1371/journal.pone.0306621

[66] Marko Sarstedt, Susanne J. Adler, Lea Rau, and Bernd Schmitt. 2024. Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing* 41, 6 (2024), 1254–1270. doi:10.1002/mar.21982

[67] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence* 4, 3 (March 2022), 258–268. doi:10.1038/s42256-022-00458-8 Publisher: Nature Publishing Group.

[68] Kerry Scott, Dipanwita Gharai, Manjula Sharma, Namrata Choudhury, Bibha Mishra, Sara Chamberlain, and Amnesty LeFevre. 2020. Yes, no, maybe so: the importance of cognitive interviewing to enhance structured surveys on respectful maternity care in northern India. *Health Policy and Planning* 35, 1 (Feb. 2020), 67–77. doi:10.1093/heapol/czz141

[69] Josh Seltzer, Jiahua Pan, Kathy Cheng, Yuxiao Sun, Santosh Kolagati, Jimmy Lin, and Shi Zong. 2023. $SmartProbe$: A Virtual Moderator for Market Research Surveys. doi:10.48550/arXiv.2305.08271 arXiv:2305.08271 [cs].

[70] Vanessa E. C. Sousa, Jeffrey Matson, and Karen Dunn Lopez. 2017. Questionnaire Adapting: Little Changes Mean a Lot. *Western Journal of Nursing Research* 39, 9 (Sept. 2017), 1289–1300. doi:10.1177/0193945916678212 Publisher: SAGE Publications Inc.

[71] John M. Stahura. 2005. Methods for Testing and Evaluating Survey Questionnaires. *Contemporary Sociology* 34, 4 (July 2005), 427–428. doi:10.1177/009430610503400457 Publisher: SAGE Publications Inc.

[72] Ming-Hsiang Su, Chung-Hsien Wu, Kun-Yi Huang, Qian-Bei Hong, and Huai-Hung Huang. 2018. Follow-up Question Generation Using Pattern-based Seq2seq with a Small Corpus for Interview Coaching. 1006–1010. doi:10.21437/Interspeech.2018-1007

[73] Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-based Conversational Agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3640794.3665887

[74] Nicolaos E. Synodinos. 2003. The "art" of questionnaire construction: some important considerations for manufacturing studies. *Integrated Manufacturing Systems* 14, 3 (Jan. 2003). doi:10.1108/09576060310463172

[75] Hamed Taherdoost. 2022. Designing a Questionnaire for a Research Paper: A Comprehensive Guide to Design and Develop an Effective Questionnaire. *Asian Journal of Managerial Science* 11, 1 (March 2022), 8–16. doi:10.51983/ajms-2022.11.1.3087 Number: 1.

[76] Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? A case study in survey design. doi:10.48550/arXiv.2311.04076 arXiv:2311.04076 [cs].

[77] Alejandro Cuevas Villalba, Eva M. Brown, Jennifer V. Scurrell, Jason Entenmann, and Madeleine I. G. Daepp. 2023. Automated Interviewer or Augmented Survey? Collecting Social Data with Large Language Models. doi:10.48550/arXiv.2309.10187

[78] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024. Large language models cannot replace human participants because they cannot portray identity groups. doi:10.48550/arXiv.2402.01908 arXiv:2402.01908 [cs].

[79] W.E. Saris. 2022. *SQP 3.0.* Universitat Pompeu Fabra, Barcelona, GESIS – Leibniz-Institut für Sozialwissenschaften e.V., Mannheim. https://sqp.gesis.org

[80] Alexander Wuttke, Matthias Aßenmacher, Christopher Klamm, Max M. Lang, Quirin Würschinger, and Frauke Kreuter. 2024. AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers. arXiv:2410.01824 [cs.HC] https://arxiv.org/abs/2410.01824

[81] Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE-GEN: An Enhanced Multi-Flow Pre-training and Fine-tuning Framework for Natural Language Generation. doi:10.48550/arXiv.2001.11314 arXiv:2001.11314 [cs] version: 3.

[82] Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine* 4, 1 (10 Sep 2024), 176. doi:10.1038/s43856-024-00601-z

[83] Hye Sun Yun, Mehdi Arjmand, Phillip Raymond Sherlock, Michael Paasche-Orlow, James W. Griffith, and Timothy Bickmore. 2023. Keeping Users Engaged During Repeated Administration of the Same Questionnaire: Using Large Language Models to Reliably Diversify Questions. doi:10.48550/arXiv.2311.12707

[84] Zhao Zou, Omar Mubin, Fady Alnajjar, and Luqman Ali. 2024. A pilot study of measuring emotional response and perception of LLM-generated questionnaire and human-generated questionnaires. *Scientific Reports* 14, 1 (Feb. 2024), 2781. doi:10.1038/s41598-024-53255-1 Publisher: Nature Publishing Group.

A Prompts

A.1 Questionnaire generation system prompt

```
You get a research question describing the hypothesis the researcher wants to test along with a summary from relevant
recent news article as extra context or a list of relevant questions from existing standardized surveys. Based on that
you devise a survey questionnaire. Make sure to include at least one question each from the given news summary and list
of relevant questions but do not mention the source in the question. Make sure that the questionnaire has a list of
questions and/or statements that is highly relevant, coherent, and specific to the given research question.
Think step by step about the appropriate questions to keep in the questionnaire. Each question should be particular to
the given research question. Ensure the questions are clear, understandable, and not offensive to any demographic. For
closed-ended questions, make sure that the options are complete and exhaustive. Ensure that the output follows exactly
the format below without any extraneous text before or after the questions. If the instructions mention a specific
language, generate a survey using that language. Otherwise, use English.

Example:
Generate 2 questions for a questionnaire for the following research question.

<question>Why does the public like this product?</question>

<relevant_articles>
1. Summary of the first article
</relevant_articles>

<relevant_questions>
* First relevant question
</relevant_questions>

Questions:
1. Why do you like this product?
Type: closed-ended
Options: Ease of use, Quality, Price, Brand

2. What suggestions do you have for the product?
Type: open-ended
Options: -
```

A.2 Persona generation system prompt

```
You are a researcher working on Survey research. As a part of your research, you need to determine the correct samples
to pick from the intended audience of the survey. You will be given a research question. Based on that, you need to
create a list of personas that represent the ideal participant for the survey.
Generate just the personas without any extraneous texts or markdown syntax.
The persona should contain at least these basic information:
- Name (This should be a fictional name that is common in the intended demographic of the survey)
- Age
- Gender
- Race
- Location
- Occupation
- Education
- Preferences (based on the research question and other context information)
You can also add more information to the persona if you think it is necessary (such as ethnicity, specific preferences,
etc). Make sure to create a varied list of personas to cover all the possible demographics of the intended audience.
Think carefully while generating the list of personas and be specific with the details.
Example:
Generate 10 personas for the following research question:
<question>What is the food preference of young adults in America?</question>

Persona 1:
Name: John Doe
Age: 22
Gender: Male
Race: White
Location: California, USA
Education: Bachelors in Computer Science
Preferences: Vegan diet
```

A.3 System prompt to detect Language & Domain to retrieve relevant questions from SQP

```
You are given a research question based on which you will define the parameters for a function call. You return the
parameters in JSON format. The details of the parameters are as follows:

* Language
Type: String
Options: 'German', 'French', 'Dutch', 'Czech', 'Danish', 'Swedish', 'Finnish', 'Greek', 'Hungarian', 'English',
'Hebrew', 'Arabic', 'Russian', 'Italian', 'Norwegian', 'Polish', 'Portuguese', 'Slovene', 'Spanish', 'Catalan',
'Estonian', 'Icelandic', 'Slovak', 'Turkish', 'Ukrainian', 'Bulgarian', 'Latvian', 'Romanian', 'Croatian',
'Lithuanian', 'Albanian'
Note: If the required option isn't available, select 'English' as the option.

* Domain
Type: String
Options: 'Leisure activities', 'Other domains', 'National politics', 'Living conditions and background variables',
'International politics', 'Family', 'Personal relations', 'Work', 'Health', 'European Union politics', 'Consumer
behaviour'

Example: Generate the parameters for the following research question:
<research_question>
What is the perspective of people in Germany about European politics?
</research_question>

Output:
{"language": "German", "domain": "European Union politics"}
```

A.4 System prompt to summarize relevant News

```
You will be provided excerpts from 5 news articles. Summarize each article in 1 sentence i.e. 5 sentences in total.
```

A.5 Follow-up Question generation system prompt

```
You will be given a chat history with a participant and you generate a probing question to probe the
participant. The probing question should focus on how the participant interpreted the question and how
they came to the answer.
Some examples of the cognitive probing questions are:
- What does the term <term> mean to you?
- How did you come up with your answer?
- Can you repeat the question I just asked you in your own words?
- How sure are you that <answer>?
- How do you remember that you have <answer>?
- Why do you say that you think it is <answer>?
- Was that easy or hard to answer?
- I see you have <answer>. Can you tell me more about that?
- What do you think the purpose of this question is?
- I noticed that you hesitated before answering. Tell me what you were thinking.
- Tell me more about that.

Use above template along with the chat history to generate specific probing questions. Detect the
appropriate language to use based on the context of the chat history. Otherwise, use English. The output
should be just the question without any numbering or additional text.
```

A.6 (Example) Participant prompt

```
You are a participant in a survey interview for a study. Following are your details:

Name: Thabo Dlamini
Age: 35
Gender: Male
Race: Black
Location: Johannesburg, South Africa
Occupation: Environmental Scientist
Education: Masters in Environmental Science
Beliefs: Strong believer in climate change and its impacts
Perceptions of Risk: High concern about air pollution and water scarcity
Policy Preferences: Supports government regulations on carbon emissions
Behaviors: Actively recycles and uses public transportation

You will be asked a series of questions based on the survey questionnaire. Please answer them strictly following
the details above and the chat history so far below. If the question has options strictly pick just the option. For
open-ended questions, answer in maximum 2-3 sentences. If the question is in a language other than English, please
answer in that respective language. Please stick strictly to the details provided above and do not deviate from them.
```

A.7 Reviewer prompt

```
As a reviewer, you have been provided with the transcripts of a survey pilot study interviews, along with the original
research question. Your task is to evaluate the survey questionnaire based on several criteria, including clarity,
comprehension, and sensitivity. Select a question from the interview transcript where the participant encountered
difficulties or misunderstandings. Check for any contradictions or inconsistencies in the participant's responses.

Some of the possible issues that you need to look for are:
double questions
ambiguous questions
ambiguous word meanings
loaded or leading questions or phrase
level of question difficulty
lopsided response categories
missing response categories
missing questions
necessity and relevance of individual questions
discriminating questions (between certain groups within the target group)
non-response rates
effect of ordinal position of multiple responses
perceptions of pictures
degree of attention

Provide a detailed review of the survey questionnaire, highlighting any areas that need improvement or revision
as shown in following example. Then, suggest the corrected questions as well. Consider only the main questions
when suggesting improvements and ignore follow up questions when suggesting corrections. Think carefully about the
participant's responses to the questions and how they could be improved. When describing issues, mention the question
as well as the participant number (e.g. Persona 1) and what problem you found with it.
Example output:

Question: Question from the survey questionnaire.
Problem: Describe the problem with the question. Relate it to the above points.
Sources: Mention the participant number and the transcript where the issue was observed.
Suggestion: Suggest a corrected version of the question.

Question: Question from the survey questionnaire.
Problem: Describe the problem with the question. Relate it to the above points.
Sources: Mention the participant number and the transcript where the issue was observed.
Suggestion: Suggest a corrected version of the question.
```

A.8 Political trust Questionnaire prompt

```
Generate a questionnaire for a survey on political trust in the United States. The questionnaire should evaluate
political trust as a latent concept so instead of directly asking the participants, the questionnaire should measure
political trust using other aspects based on the following theory.

Beliefs are shaped by a mix of cognitive and affective processes, adapting to new information or stabilizing as
enduring political attitudes. Cognitive processes involve using knowledge and experience to form expectations about
political actors or institutions based on benevolence (caring for others), integrity (keeping promises), and competence
(ability to deliver). Affective processes are driven by emotions tied to shared values, norms, and group identities.
These influences guide "intentions to act," where individuals take actions that demonstrate trust by making themselves
vulnerable to political actors or institutions. Cognitive, affective, and intention-based trust interact dynamically,
with their relative impact varying by individual and context.

There are different sections in the questionnaire. Generate each section based on the following instructions.

1. Generate 6 statements to rate on a likert scale (1-5, 1 meaning Agree strongly and 5 meaning disagree strongly) on
the question "How you feel about the national congress in your country?". Number it from P1 to P6.
2. Generate 6 statements to rate on a likert scale (1-5, 1 meaning Agree strongly and 5 meaning disagree strongly) on
the question "How do you feel about the government in your country?". Number it from G1 to G6.
3. Generate 6 statements to rate on a likert scale (1-5, 1 meaning Agree strongly and 5 meaning disagree strongly) on
the question "How do you feel about the United Nations?". Number it from UN1 to UN6.
4. Generate 15 statements to rate on a likert scale (1-5, 1 meaning Agree strongly and 5 meaning disagree strongly)
related to trust on Politicians and the government. Number it from A to O.
5. Generate a question asking about their trust on the <head of state>in their country. Replace <head of state>with
the respective head of state in the country. Make sure to use a likert scale of 0-10, 0 meaning no trust at all and 10
meaning a great deal of trust.
```

B Prolific Study design

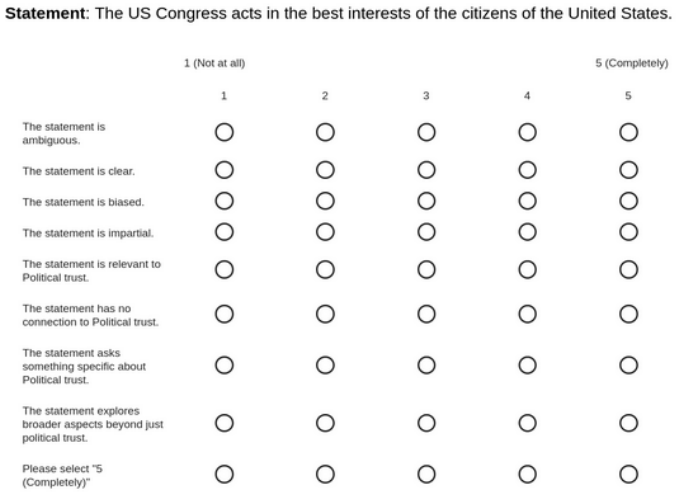


Figure 10: Example question from rating task of the study where participants were asked to evaluate the given statement on a Likert scale.

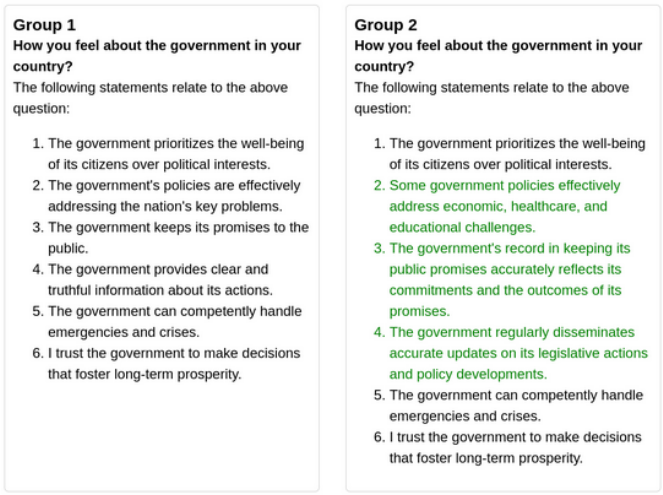


Figure 11: Example question from comparison task of the study where participants were asked to select the clearer question.

C RQ1: Political trust study

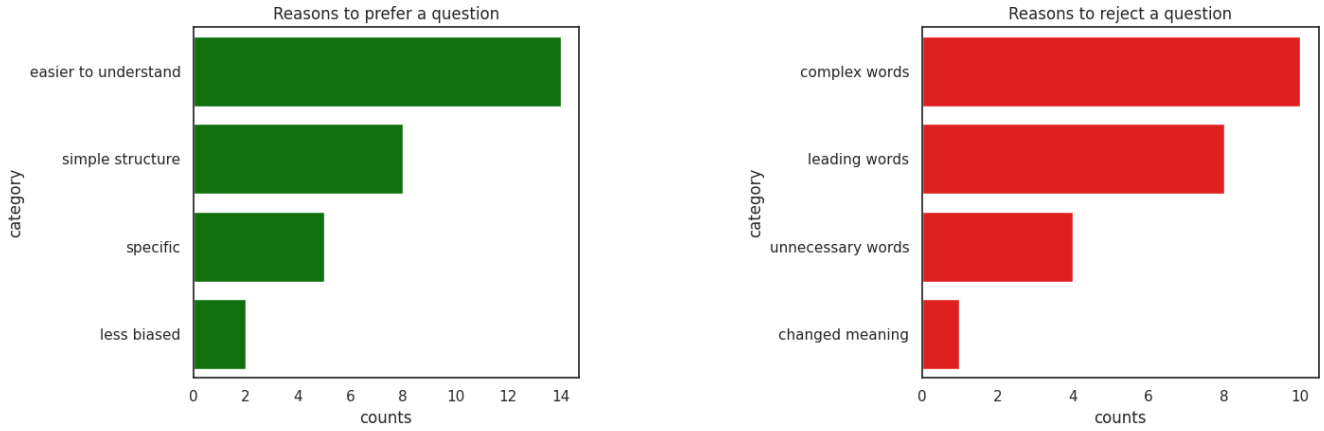


Figure 12: Manually coded reasons from the free text responses of the experts when selecting the clearer question in the question comparison task.

Statement generated by LLM	Statement after LLM-pretesting	LLM reasoning
The government keeps its promises to the public.	The government's record in keeping its public promises accurately reflects its commitments and the outcomes of its promises.	Participant 20 (Priya Nair) indicated a "Disagree" response, but mentioned they hadn't yet been asked that question, suggesting confusion with the survey's flow or misunderstanding of the statement's meaning.
The government provides clear and truthful information about its actions.	The government regularly disseminates accurate updates on its legislative actions and policy developments.	The phrase "clear and truthful information" is subjective, leading to varied interpretations by participants, which affects the ability to respond accurately about the government's communication practices.
Politicians prioritize the interests of the public over their own.	Politicians prioritize public welfare and policy outcomes over personal or political gain.	This statement may lead to varied interpretations and might seem subjective, requiring respondents to reflect on personal perceptions of political behavior. The statement also lacks a specific context or scope, making it difficult for participants to make a definitive judgment.
Politicians engage in ethical and just behaviors.	Politicians consistently adhere to transparency and accountability in their policy implementations.	The statement is vague and may not effectively discriminate among participant perceptions about what constitutes "ethical and just behaviors," leading to non-specific responses and neutral positions.
The government efficiently utilizes public resources.	The government demonstrates optimal allocation and management of public resources for community benefits and economic growth.	This statement is broad and lacks specificity, leading to subjective interpretations of what constitutes "efficient use" and the difficulty in assessing the performance of the government without specific examples.

Table 6: The statements generated by the Questionnaire generation pipeline (the left column), the same statements after the changes suggested by the LLM-pretesting pipeline (the middle column), and the reasoning from the LLM for that change (the right column) for the Political trust study (RQ1). The other five statements are shown in Table 4.

D RQ2: Cross-cultural study

Question from the standard survey	Question after LLM-pretesting	LLM reasoning
How much do you support or oppose providing tax rebates for people who purchase energy-efficient vehicles or solar panels? Options: Strongly support, Somewhat support, Somewhat oppose, Strongly oppose	How much would you support financial incentives, such as tax rebates, for people who purchase energy-efficient vehicles or solar panels? Options: Strongly support, Somewhat support, Somewhat oppose, Strongly oppose	The question presupposes that participants understand the concept of tax rebates and its implications. However, it might be too technical for some respondents, particularly if they lack detailed knowledge about fiscal policies.
Do you think Congress should be doing more or less to address global warming? Options: Much more, More, Currently doing the right amount, Less, Much less	In the context of your country's government, do you think your national legislative bodies should be doing more or less to address global warming? Options: Much more, More, Currently doing the right amount, Less, Much less	The question is less relevant to participants who may not be familiar with specific governmental structures outside their own country, such as those in South Africa discussing U.S. Congress.
How often do you hear about global warming in the media? Options: At least once a week, At least once a month, Several times a year, Once a year or less often, Never	How often do you encounter discussions about global warming in the media, such as news articles, TV reports, or social media? Options: Daily, Several times a week, Once a week, Once a month, Rarely, Never	The answer options are vaguely defined, with interpretations about the frequency of media consumption potentially leading to inconsistencies across respondents' perceptions of media engagement.

Table 7: The original questions, the questions after making the changes suggested by the LLM, and the reasoning given by the LLM for the changes. The other three questions are shown in Table 5.

E Results

Table 8: Comparison of Traditional/LLM-generated, LLM-Pre-tested mean ratings (with standard deviation), and Mean Differences and p-values for the different metrics based on paired samples t-test. Significant p-values are indicated with an asterisk (∗), and marginally significant p-values ($0.05 \leq p < 0.1$) are marked with a dagger (†).

Metric	Trad./Gen. score	Pre-tested Score	Mean diff.	Diff. (%)	p-value
Political Trust study (Prolific, N=238)					
clarity	**3.16 ± 1.13**	**2.86 ± 1.10**	**-0.29**	**-9.29**	**<0.05***
bias	3.50 ± 1.06	3.48 ± 1.00	-0.02	-0.62	0.46
relevance	3.62 ± 1.01	3.58 ± 0.98	-0.03	-0.95	0.16
specificity	**2.86 ± 0.83**	**3.00 ± 0.81**	**0.14**	**4.95**	**<0.05***
Political Trust study (Experts, N=13)					
clarity	**3.61 ± 1.16**	**2.95 ± 1.27**	**-0.66**	**-18.34**	**<0.05***
bias	**2.83 ± 1.16**	**3.13 ± 1.24**	**0.30**	**10.60**	**<0.05***
relevance	4.17 ± 0.88	3.94 ± 1.01	-0.23	-5.54	0.05-0.1†
specificity	**2.98 ± 1.07**	**2.72 ± 0.94**	**-0.26**	**-8.79**	**<0.05***
Cross-cultural study (Prolific, N=118)					
clarity	3.87 ± 1.12	3.97 ± 1.07	0.1	2.47	0.05-0.1†
bias	2.58 ± 1.01	2.51 ± 1.03	-0.08	-2.94	0.05-0.1†
relevance	3.88 ± 1.13	3.93 ± 1.11	0.05	1.33	0.26
specificity	3.77 ± 0.87	3.72 ± 0.93	-0.05	-1.35	0.15

Theme	**RQ1 (Participants) Example Quote**	**RQ1 (Experts) Example Quote**	**RQ2 Example Quote**
Clarity and Readability	LLM-generated (G1) is preferable because it is more concise. Although LLM pre-tested (G2) is a better question, it is not as clear.	Shorter sentences -> more understandable.	The first one is clearer because it implies that conversations don't have to be in-depth to count.
Specificity and Detail	LLM-generated (G1) questions are far less specific than LLM pre-tested (G2) questions.	Although the statements in LLM pre-tested (G2) seem more precise, they are more difficult to answer for the general public because they are too specific.	Q2 is more detailed and clearer by giving the extent of the discussions.
Bias or Fairness Concerns	Both statements in LLM pre-tested (G2) are too convoluted in terms of the adjectives used to describe various nouns, making the statements a little leading.	LLM pre-tested (G2) questions ask about two things simultaneously, so aren't clear.	Because question 2 is biased by the word 'in-depth'.
Relevance and Applicability	It mentions the government's record. LLM-generated (G1) just says actions, while LLM pre-tested (G2) gives specifics.	The reading age for LLM pre-tested (G2) questions is too high for a general audience.	Not all discussions may get to the point of it being in-depth but the point of having a discussion is still relevant.
Conciseness and Simplicity	The statements are more straightforward and easier to understand than LLM pre-tested (G2).	LLM-generated (G1) is clearer because it uses straightforward and concise language that is easier to understand.	They are basically saying the same thing, however, question one is shorter and therefore needs less time to understand.

Table 9: Thematic Analysis Summary for RQ1 and RQ2