

Article

Breaking the Ceiling: Mitigating Extreme Response Bias in Surveys Using an Open-Ended Adaptive-Testing System and LLM-Based Response Analysis

Moshe Gish ¹ , Amit Nowominski ²  and Rotem Dror ^{2,*} 

¹ School of Environmental Sciences, Faculty of Social Sciences, University of Haifa, Abba Khoushy Ave 199, Haifa 3498838, Israel; mgish@univ.haifa.ac.il

² Department of Information Systems, Faculty of Computer and Information Science, University of Haifa, Abba Khoushy Ave 199, Haifa 3498838, Israel; amit.nowominski@gmail.com

* Correspondence: rdror@is.haifa.ac.il

Abstract

Assessments of extreme psychological constructs often face a persistent challenge: the ceiling effect, in which a significant proportion of respondents select the highest score on a scale, thus obscuring meaningful variation within the population. This effect may have profound consequences in studies of extreme psychological constructs. To address this limitation, we present a novel framework that integrates Multistage Testing (MST) with open-ended questions that are automatically analyzed by large language models (LLMs). This hybrid approach adapts the survey questions to the respondent while leveraging LLMs to efficiently and reliably interpret free-text answers from large-scale online surveys. Using a case study on aversion toward cockroaches, we show how our method can effectively eliminate extreme ceiling effects, revealing hidden data distributions that are often obscured by extreme responses to conventional Likert-type survey questions. We also validate our method by comparing LLM performance to expert human annotations. This demonstrates the consistency and reliability of LLMs in evaluating free-text answers. This framework offers a generalizable methodology that enables more precise and sensitive quantitative measurement of extreme psychological constructs, allowing researchers to study topics that until now were inaccessible due to significant, inherent ceiling effects.

Keywords: large language models; multistage testing; psychological assessment; extreme psychological constructs; ceiling effect; survey analysis; automated scoring



Academic Editor: Steven R. Livingstone

Received: 10 January 2026

Revised: 5 February 2026

Accepted: 10 February 2026

Published: 13 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Psychological measurements (the quantification of cognitive, emotional, and behavioral constructs) are fundamental in multiple research domains. In healthcare, psychological assessments guide diagnosis and treatment, directly affecting patient outcomes [1]. In education, assessments of both student-learning and teacher-instructional experiences are vital for both formative improvement and summative evaluation of teaching and learning [2]. In human–computer interaction, understanding users’ psychological states is key to user-centered design, as integrating psychological principles into system design boosts engagement, motivation, and overall user experience [3].

Studies employing self-report psychological measurements often incorporate open-ended questions, which allow individuals to express their perspectives in their own words rather than constraining their responses to predefined categories. This approach gives

researchers access to rich and complex insights into participants' perceptions and emotional experiences, facilitating the discovery of underlying processes, motivations, and context-specific phenomena that may not emerge from closed-ended items. However, these advantages come with notable trade-offs. In particular, interpreting and coding open-ended responses is labor-intensive, often resulting in smaller sample sizes and limited statistical generalizability [4–7]. In addition, the manual nature of this work can introduce subjectivity into the interpretation and classification of responses and may be susceptible to researcher bias (e.g., “confirmation bias”, where the interpretation of answers is subconsciously biased to confirm the researcher’s pre-existing beliefs or values), potentially compromising the reliability and validity of findings [8]. Furthermore, the absence of standardized response formats complicates cross-participant comparisons and poses challenges for integrating open-ended data with quantitative measures, often necessitating sophisticated mixed-methods approaches to improve the reliability of conclusions [5,9].

In contrast, the use of closed-ended items to measure psychological constructs enables researchers to collect data more efficiently, accommodate larger sample sizes, and obtain more standardized and precise measurements. This structured approach enhances the feasibility of statistical analyses and strengthens the ability to establish robust empirical relationships [5,10,11]. Although closed-ended items lack the depth and richness of open-ended formats, they are particularly suitable for large-scale internet surveys—a convenient and cost-effective means of data collection in the social sciences. Nonetheless, such items may introduce significant biases. Unless mitigated *a priori* or addressed *post hoc*, these items may be unsuitable for capturing certain types of knowledge.

One such bias is known as the “ceiling effect” (or the opposite “floor effect”), which occasionally occurs in surveys that use Likert scales, when more than 15% of respondents choose the highest (or lowest, in the case of a floor effect) possible score, resulting in reduced variability and limiting the ability to detect differences between individuals [12,13]. This clustering at the upper end of the measurement scale can mask true relationships between variables, thus compromising the accuracy and interpretability of statistical analyses [14,15]. If no existing methodology can adequately address the ceiling effect in a key parameter, research on affected topics may stall, rendering even straightforward research questions intractable. A good example of this is the case of entomophobia (the fear of insects) [16,17].

In a recent study on entomophobia, Gish et al. [18] hypothesized that a higher level of entomophobia leads to greater use of insecticides in homes. Despite the seemingly trivial nature of the research question and the large scale of the study, the authors appeared to have hit a measurement wall: self-ratings of entomophobia toward common household arthropods clustered at the top end of a 7-point scale, forming a clear ceiling effect with 47–50% of respondents choosing the highest score on the scale (Figure 1a). Even though entomophobia was prevalent and intense in the sampled population, and household insecticides were used frequently, no meaningful correlation between these two variables was detected. Rather than accepting this null result at face value, the authors suggested that the ceiling effect they observed had led to a Type II error (failure to reject a false null hypothesis). Consequently, since there was no other self-report method that can reliably and directly measure levels of phobia toward highly aversive creatures such as spiders and cockroaches, research on entomophobia’s causes and consequences could not be advanced.

One promising solution to the ceiling effect in studies of common and extreme phobias is computerized adaptive testing (CAT), an algorithm-based method of administering exams that adapts to the performance of the test-taker during the test itself [19,20]. One form of CAT, which takes a related but different approach, is the computerized multistage testing (MST): instead of selecting individual items according to item response theory (IRT), the algorithm routes respondents through fixed sets of items (modules), each calibrated to

a different level of ability [20]. In recent years, MSTs have been adopted in large-scale assessments, demonstrating comparable precision to classical CAT while offering greater content control and more predictable operation [21].

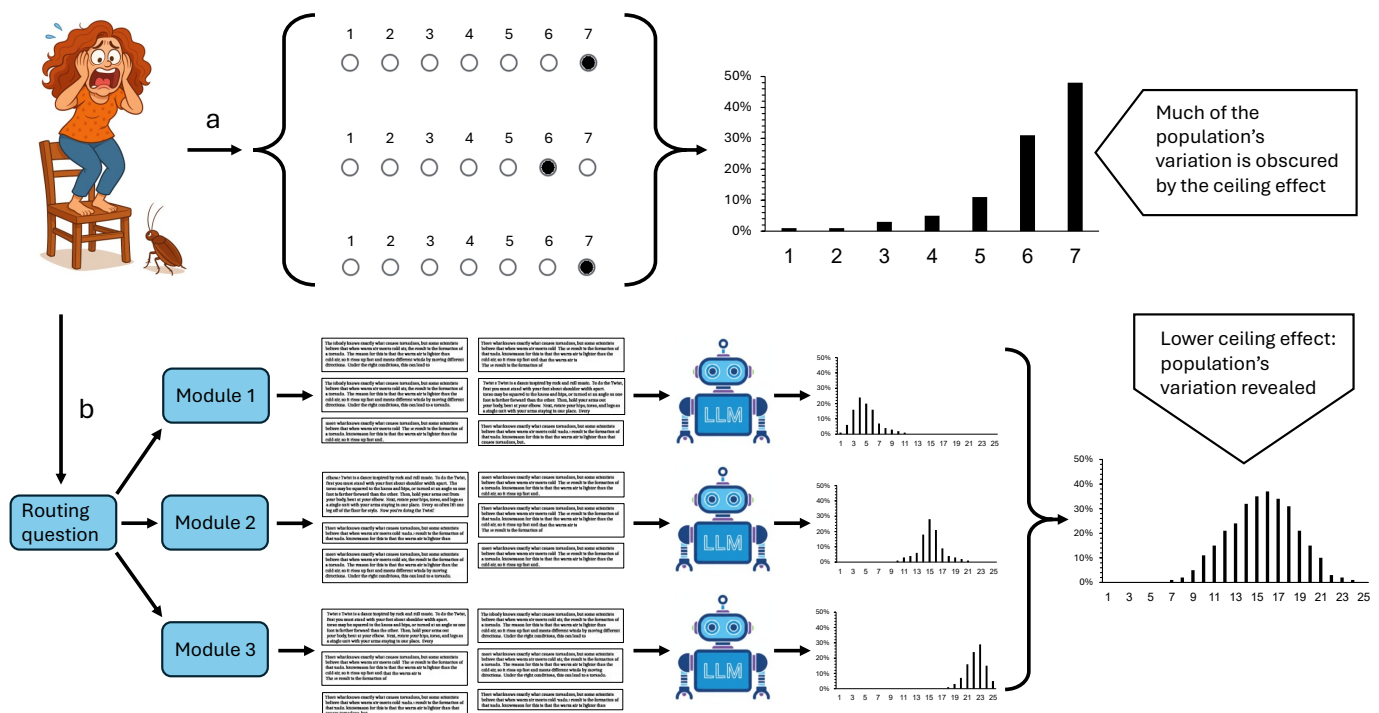


Figure 1. Framework of the proposed methodology. (a) How a ceiling effect may arise in a survey on extreme psychological constructs. (b) A hybrid survey design combining multistage testing (MST) with open-ended items, augmented by LLM-based response interpretation and quantification.

However, while improving measurement accuracy, an MST will not eliminate the tendency of respondents with high phobia to select the top category on the measurement scale. A practical solution could be to replace fixed-choice items within the MST with open-ended questions describing hypothetical situations, thus allowing participants to describe what their thoughts and feelings would be in those situations in a free-text form. This approach would make it more difficult for respondents to reach the highest possible score, and it would also eliminate the “reference group effect,” in which respondents assess themselves relative to their perceived peers rather than an absolute standard [22], since respondents are not asked to rate themselves on a measurement scale. The interpretation and coding of free-text responses could reveal important variations that would otherwise be hidden at the extreme end of the measurement scale. Therefore, integrating open questions into an MST could potentially mitigate data skewness caused by the ceiling effect, thus increasing the sensitivity of the test to nuances of extreme phobia.

At the same time, the high resource demands and other limitations and potential biases involved in turning free-text responses into quantitative data by human interpreters would make it difficult to employ a hybrid MST/open-ended design in decently replicated online surveys. To do this, an automated interpretation of the free-text responses would be required. We propose adopting a recently popular automatic methodology from the field of Natural Language Processing (NLP), for analyzing and classifying textual responses to open-ended questions, namely, LLM-as-a-judge [23,24].

Large language models (LLMs) have demonstrated remarkable abilities in interpreting, analyzing, and reasoning about textual data [25–28]. Specifically, LLM-as-a-judge is a

paradigm in which LLMs are used to annotate data and evaluate models, replacing human annotators and judges [29]. One key advantage of the LLM-as-a-judge paradigm is the scalability, speed and cost-effectiveness of LLMs. In our case, LLMs have the ability to quickly classify numerous responses to open-ended questions, based on the researcher's instructions outlined as a prompt submitted to the LLM. This action obviates the need to manually classify responses, a task that is time-consuming and could introduce inconsistencies and subjectivity biases, which often affect the reliability of large human-annotated datasets [30–32].

LLMs are increasingly being employed as reasoning engines, combining their encoded knowledge with expert knowledge inserted by prompt to support research in various domains [33,34]. By combining a knowledge-based LLM with a structured, open-ended MST, we establish an adaptive and automated psychological assessment framework capable of mitigating ceiling effects and capturing subtle variations in extreme psychological constructs within large population samples. Our proposed framework is illustrated in Figure 1b.

The contributions of this research are threefold. First, we develop an MST-based adaptive testing framework tailored specifically for open-ended responses, which mitigates the ceiling effect. Unlike traditional MST-based survey systems, which rely on predetermined item banks of closed-form questions, our approach uses open-ended questions and LLM-based analysis of free-text responses, thus supporting large-scale surveys. Second, we offer a methodological implementation and validation of LLMs as reliable evaluators of psychological constructs by benchmarking their performance against human judgments. Third, we demonstrate the applicability of the MST–LLM framework through a case study on cockroach aversion, demonstrating how this integration mitigates the ceiling effect, thus allowing better measurement precision and richer insights, compared to what conventional self-report methods offer when used for measuring extreme psychological constructs.

2. Related Work

In this work, we address two key challenges that could be encountered when trying to measure extreme psychological constructs. The first is the ceiling effect, which arises when traditional scales fail to capture meaningful variation because many respondents select the highest option. The second challenge arises from our proposed solution—using open-ended questions to mitigate the ceiling effect—which creates the need for analysis methods suited for large-scale studies, where manual annotation is impractical. To tackle the latter, we adopt the emerging LLM-as-a-judge framework, in which LLMs are used to automatically evaluate and interpret free-text responses.

2.1. The Ceiling Effect

The ceiling effect is occasionally found in surveys that use Likert scales, when more than 15% of respondents choose the highest possible score, rendering their scores undistinguishable from each other [13]. The ceiling effect could be a major concern, for example, in patient satisfaction surveys [15,35,36], or when researchers apply an existing measuring instrument to a new population with higher levels of the construct than the population for which the instrument was originally developed [37]. If unaccounted for, a ceiling effect can significantly lower the measuring instrument's validity (the degree to which an instrument truly measures what it purports to measure), its responsiveness (the ability to detect change), and its reliability (the degree to which the measurement is free of measurement error) [37]. When research data are subject to a ceiling effect, the resulting skewed distribution often violates the normality assumptions required for many statistical tests. As a consequence, non-parametric analysis methods, which generally have lower

statistical power, must be employed [38]. To mitigate ceiling effects, researchers use both design-based and analytical strategies.

On the design side, improving the response scale may allow, in some cases, finer distinctions among high-end responses [39]. This could be done by using an unbalanced scale with more positive labels than negative ones (a positively packed scale) or adding textual labels to all or most values on the scale [36,37,39–41]. Some researchers also report lower ceiling effects when replacing Likert-type items with visual analogue scales (a continuous measure that allows respondents to indicate any position on the scale that best represents their opinions) [42–44]. Another approach is to switch to a mixed-methods design, integrating fixed scales with open-ended questions that are manually interpreted and coded [45]. Alternatively, researchers may choose post-hoc analytical tools for dealing with ceiling effects after the data has been collected. Treating maximum scores as censored values permits the use of Tobit or other censored regression models to adjust for the restricted range and recover more accurate estimates [46,47].

However, these methods may fail in studies of mental health extremes or when research targets particularly sensitive, emotionally charged, or taboo subjects [48,49]. In these instances, altering the response scale is often insufficient to counteract the ceiling effect, since respondents tend to provide the highest possible ratings regardless of the number of points on the scale, which descriptors are used, or how positively skewed the options might be. In such cases, the reference group effect [22] should also compromise the accuracy of the measurement tool. For example, if two individuals have similar levels of a certain extreme psychological construct, they may rate themselves differently: one choosing 6 and the other 7 on a 1–7 scale, simply because the first knows someone with even stronger aversion, while the second does not. Furthermore, when intense emotions or affective attitudes give rise to significant ceiling effects, post-hoc analytical techniques cannot remedy the resulting skew, since the genuine variability across much of the population is never captured in the first place. Therefore, in some cases, ceiling effects can create significant barriers to knowledge acquisition that cannot be overcome with statistical methods or by improving the response scale.

More recently, it has been suggested that generative AI could assist in designing traditional questionnaires that adhere more strictly to psychometric principles. While utilizing LLMs to refine item wording may improve clarity and reduce ambiguity, these improvements do not resolve the fundamental structural limitation of fixed-choice scales: the finite set of response options. Even an optimally designed Likert item constrains the respondent to a discrete bin. In cases of extreme psychological traits, where the variation occurs at the tail of the distribution, discrete options, regardless of how well-written, fail to provide the necessary granularity.

2.2. Multistage Testing

MST is essentially a form of computerized adaptive testing (CAT), an algorithm-based method of administering exams that adapts to each test-taker's performance [20]. CAT dynamically selects test items based on previous answers, thus reducing test length and increasing measurement precision [50]. Unlike traditional linear tests, where all participants receive the same set of items, a CAT selects each subsequent question in response to the previous answer; correct answers are followed by more difficult questions that award a greater number of points, while incorrect answers lead to easier questions that yield fewer points [51]. MST, however, takes a related but different approach: instead of selecting individual items according to item response theory (IRT), the algorithm routes respondents through fixed sets of items (modules), each calibrated to a different level of ability [20]. In re-

cent years, MSTs have been adopted in large-scale assessments, demonstrating comparable precision to CAT while offering greater content control and more predictable operation [21].

While conventional MSTs are typically used to assess cognitive skills or abilities, our approach adapts the MST framework for measuring psychological constructs. In this adaptation, respondents are routed by an initial routing question to a module appropriate to their indicated level on the construct of interest. Rather than right-or-wrong questions, these modules employ open-ended questions. A direct consequence of implementing this hybrid MST/open-ended design in large-scale surveys is the generation of substantial volumes of free-text data that require systematic interpretation and scoring. For instance, in Gish et al. [18], if each respondent had provided six free-text answers (as in our proposed design), the resulting dataset would have comprised approximately 15,000 responses, rendering manual coding impractical. Thus, the feasibility of this methodology is contingent upon automated scoring of responses, making an LLM that serves as an automated evaluator an integral component of the proposed framework.

2.3. LLM-as-A-Judge

The rise of LLMs has also brought the advent of LLM-as-a-judge. Originally, the term “LLM-as-a-judge” referred to the use of LLMs for evaluating the outputs of other models within NLP tasks [23]. Over time, however, the approach has expanded to include a broader range of applications, such as general evaluation of models, annotation of datasets, and labeling of both text and image—tasks that were traditionally performed by humans [23,52–54]. This approach is applied not only in NLP. For example, Gilardi et al. [32] evaluate GPT as an annotator for political text classification and show it can match average annotator accuracy, and Argyle et al. [55] use LLMs to simulate survey respondents and to label attitude-relevant text.

In this work, we adopt the LLM-as-a-judge framework for a new purpose: the analysis of open-ended survey responses and the classification of respondents according to levels of a measured attribute. The use of LLMs for the analysis of survey responses is still limited [56,57]. Some similar applications use LLMs to analyze open-ended responses to test questions [58,59] and for qualitative analysis [60]. However, applications in psychological research are scarce [61], largely due to concerns about hallucinations, biases, and the difficulty of ensuring consistent and reliable judgments [62,63]. These concerns are especially pressing when working with open-ended responses, since LLMs are particularly vulnerable to hallucination, over-interpretation, and confirmation bias. Consequentially, they may impute sentiments or motives not evidenced in the text, collapse minority or ambivalent expressions into normative categories, or conflate sarcasm and figurative language with literal meaning [32,64,65].

Given these challenges, it is necessary to methodically test whether the LLM-as-a-judge performs well in the context of analyzing survey responses. In this work, we contribute to this agenda by outlining validation strategies; specifically, we adopt the alternative test outlined in Calderon et al. [29] to test if the LLM is a suitable replacement for a human judge. We also discuss the conditions under which LLM judgments can be trusted, and offer practical guidelines for prompt design and model selection that could help reduce bias and improve reliability of AI analysis.

3. Methods

The methodology presented in this paper combines a hybrid MST/open-ended design with LLM-based analysis of free-text responses. This analysis categorizes responses into predefined levels of a certain psychological construct, thus facilitating its quantitative measurement.

This methodology addresses the ceiling effect in two principal ways. First, by replacing Likert-type scale items with open-ended questions, we reduce the likelihood that respondents will easily choose the highest score. Attaining the highest rating would now require providing free-text responses that genuinely reflect the most extreme manifestations of the measured construct—a task that would be inherently challenging. Additionally, the open-ended format mitigates reference-group effects, as respondents are less able to benchmark their answers against perceived peer norms. Second, the MST structure increases measurement resolution at the upper end of the scale, and the tailoring of questions to each respondent's level enhances the reliability and validity of the assessment.

3.1. Modules

An MST is composed of modules, each of which corresponds to a distinct level of the underlying trait being assessed. For example, in the case of measuring fear, the modules might represent no fear, mild fear, and extreme fear. To ensure clarity and usability, it is crucial to define a limited number of modules, with each representing a meaningfully distinct level of the trait. If the number of modules is set too high, it is difficult to formulate questions that are clearly suitable for a given module, rather than an adjacent one. Consequently, the boundaries between the modules blur, and this reduces the reliability of the survey. On the other hand, setting the number of modules too low risks making the survey indistinguishable from a standard, linear format, thereby failing to address the ceiling effect that MST is designed to overcome. Therefore, we recommend using a structure of 3–5 modules to balance accuracy with effectiveness.

3.2. Routing Question for Module Assignment

To assign each respondent to the appropriate module, we recommend starting the survey with a closed-ended routing question, since an open-ended question would require real-time, unsupervised decision-making by the LLM, which may be unreliable. A closed-ended routing question works well here, as its purpose is not to measure the trait directly but rather to classify participants into the most suitable module. While this question may be subject to a ceiling effect, the impact of such an effect would be minimal because the subsequent open-ended questions inside each module are designed to further subdivide participants into finer-grained levels. If respondents misjudge themselves in the routing question (e.g., an individual with a high level of aversion that chooses an answer that indicates a medium level), then an overlap between adjacent modules (explained below) would still allow that respondent to achieve a higher final score.

3.3. Survey Questions

The survey is composed of open-ended questions describing various hypothetical scenarios, asking respondents to describe what their thoughts and reactions would be in such situations. The questions vary in the degree to which the described scenario aims to evoke the measured psychological construct (“harshness”). The harshness of the questions varies between modules, as a question that elicits a strong reaction from participants positioned at the higher end of the scale may generate little or no reaction from those at the lower end. Accordingly, question harshness is adapted to the module in which each question is included, ensuring that the questions are relevant and appropriately challenging for respondents at different levels of the measured construct. To allow variable performance within each module, the questions it contains should elicit a certain range of possible responses; for example, in our case study (Section 3.6), two questions in each module aimed to evoke relatively low levels of the measured construct, two were aimed at medium levels, and another two targeted high levels of the measured construct. This way,

each module contained a balanced set of six questions that together covered a wide range of the measured construct. An illustration of the survey's structure appears in Figure 2.

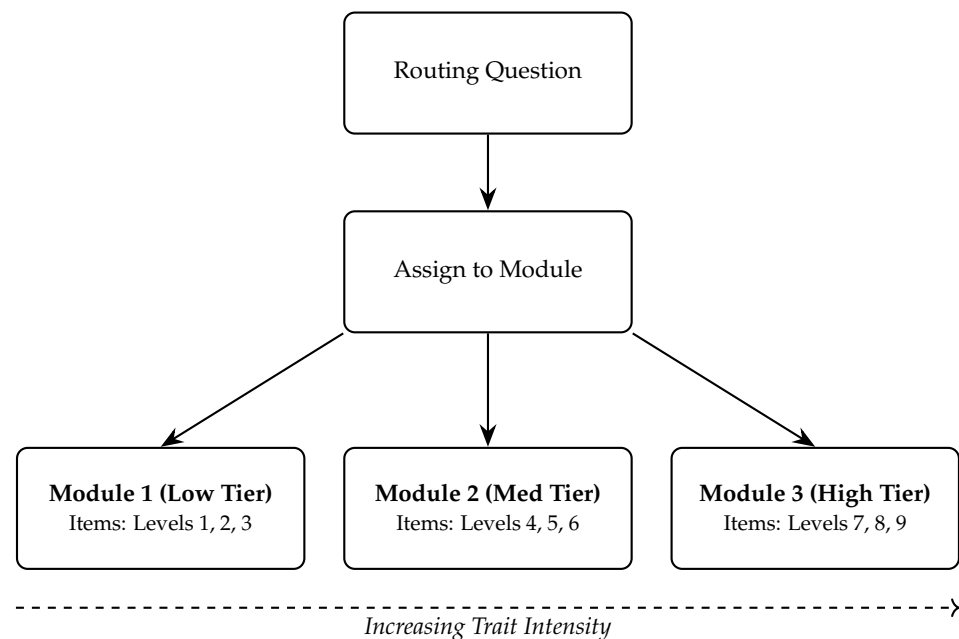


Figure 2. Survey structure. The routing question assigns respondents to a module corresponding to their estimated trait level. Note that “Low” items in Module 3 (e.g., Level 7) represent a higher absolute intensity than “High” items in Module 1 (e.g., Level 3).

While conventional MSTs are typically used to assess cognitive skills or abilities, our approach adapts the MST framework for measuring psychological constructs. In this adaptation, respondents are routed by an initial routing question to a module appropriate to their indicated level on the construct of interest. Rather than right-or-wrong questions, these modules employ open-ended questions. A direct consequence of implementing this hybrid MST/open-ended design in large-scale surveys is the generation of substantial volumes of free-text data that require automatic interpretation and scoring.

3.4. Analysis of Survey Responses

The analysis of survey responses is conducted automatically with LLMs that classify each response into a specific level of the measured trait. The input to the LLM is a prompt that contains the survey question, the participant’s response, and instructions for classification together with the desired output format. The model’s output is a categorical decision corresponding to the evaluated trait intensity level.

3.4.1. Classification Instructions

An essential element of prompt design is the clarity of classification instructions provided to the LLM. The model should be explicitly directed to assign each response to one of the predefined categories, without generating additional interpretations or free-text elaborations unless specifically instructed. It should be noted that some models, such as GPT-4o, are optimized for step-by-step reasoning, which may not align with the strict adherence to instructions required in this context. To achieve optimal prompting, we recommend the following workflow (see example in Appendix B):

1. Define a small number of classification categories (e.g., low, medium, high).
2. Provide detailed descriptions of the types of responses that are appropriate for each category.

3. Clearly delineate category boundaries to minimize ambiguity and reduce the likelihood that a single response could fit multiple categories.
4. Incorporate keywords within the classification instructions whose presence may assist the LLM in distinguishing between categories.
5. Specify the desired output format, such as outputting only the label, or the label accompanied by a brief justification.
6. Prohibit any speculative inferences about the respondent's background, motivation, or intent beyond what is explicitly stated in the response's text.
7. Conduct a small-scale preliminary survey and refine classification instructions based on observed response content. At this stage, it is often necessary to further clarify the separation between categories.
8. Have an LLM classify the responses from the preliminary survey. Then, visualize the distribution of classifications for each question (see Appendix D). This would allow the researcher to detect unbalanced distributions or deviations from expected patterns, which may indicate flaws in the instructions.
9. Revise the instructions as needed in light of the outcomes of the previous stage. Occasionally, the wording of the survey questions themselves may require modification to support proper classification. This iterative process of re-running the LLM classification and revising should be repeated as necessary, until the classification distributions meet expectations.
10. Administer a second small-scale preliminary survey to ensure both the survey questions and classification instructions are optimally designed.

In our study, we also experimented with few-shot prompting [66], a method in which the prompt includes a few labeled examples to guide the LLM in subsequent classifications. However, since open-ended questions allow the respondents to express themselves openly and in diverse ways, we found that few-shot prompting offered no clear advantage over explicitly defining the label categories within the prompt. It is important to emphasize that classification is performed at the level of individual responses rather than across the full set of a participant's answers. Providing all responses to the LLM at once and requesting a single, overall classification would encourage the model to average or infer patterns across responses, potentially amplifying biases or overlooking meaningful variation between items. In contrast, classifying each response separately preserves the item-level distinctions that are fundamental to the MST design. This ensures that subtle expressions of the trait are not lost, enabling the detection of abnormalities and contradictions, thus allowing the scoring system to weigh responses appropriately across modules. The final score is therefore derived from an aggregation of these independent classifications, following the differentiated scoring scheme described below, which both respects the hierarchical module structure and enhances measurement precision.

3.4.2. Filtering Non-Human Responses—Bot and AI Detection

Ensuring data quality is critical when analyzing survey responses, particularly with large-scale surveys that are deployed online. Some responses may be generated by automated agents or AI systems rather than genuine human participants. To mitigate this risk, we implement a combination of heuristic and model-based approaches for bot detection. This procedure should be applied prior to analyzing the survey responses.

Researchers should conduct a post-survey screening to identify and exclude submissions that are automated, low-effort, or intentionally contrarian. This involves removing responses that are too short, irrelevant, not serious, off-topic, or unreasonably lengthy, as well as detecting answers that appear to be LLM-generated. For example, responses that end with generic prompts to elaborate further or that provide multiple possible answers

depending on a choice from several possible hypothetical personalities. Screening can be carried out both manually and with the help of AI tools to ensure greater accuracy.

When participants are recruited through online platforms such as Prolific or Amazon Mechanical Turk, researchers should also implement a verification step. Participants can be asked to report their age or another piece of demographic information stored in the platform's database, and the reported data should be cross-checked against the platform's metadata. Any discrepancies should be flagged as potential indicators that a participant is non-human or otherwise unreliable. Responses flagged using these methods can either be removed or subjected to manual inspection to confirm authenticity. This filtering step increases the chances that subsequent analysis will rely on high-quality, human-generated data.

3.4.3. Scoring Scheme

We recommend classifying each response into one of three levels—low, medium, or high—corresponding to the relative intensity of the measured trait within the context of a given module. However, these levels are not absolute across the entire survey; rather, their interpretation depends on the module to which they belong. For example, a response classified as “level 1” (low) in the extreme module does not carry the same meaning as “level 1” in the mild module. To address this, we assign differentiated numerical scores when aggregating responses across modules, thereby preserving the hierarchical structure of the MST design. Importantly, in the final calculation of the overall score, we allow overlaps between adjacent modules. This means that the highest level of one module may correspond numerically to, or even exceed, the lowest level of the next module, allowing some compensation for potential misrouting and retrospective dynamic movement between modules [67].

As an illustration of a scoring scheme, consider a participant routed to module $m \in \{1, 2, 3\}$ who provides six classified responses $r_1, \dots, r_6$. We define the mapping from the LLM label to numeric points as follows:

$$\text{map}(r) = \begin{cases} 0 & \text{if } r = A \text{ (low aversion)} \\ 1 & \text{if } r = B \text{ (moderate aversion)} \\ 2 & \text{if } r = C \text{ (high aversion)} \end{cases} \quad (1)$$

The within-module sum of points for participant i is

$$S_i = \sum_{j=1}^6 \text{map}(r_{i,j}), \quad S_i \in [0, 12]. \quad (2)$$

To allow higher modules to contribute more strongly to the final score, we add a module-specific offset O_m , defined as

$$O_m = 6 \cdot (m - 1), \quad (3)$$

so that $O_1 = 0$, $O_2 = 6$, and $O_3 = 12$. The participant's final score is therefore

$$F_i = S_i + O_m. \quad (4)$$

Accordingly, the ranges of possible final scores for three modules are

$$F_i \in \begin{cases} [0, 12] & \text{for } m = 1 \\ [6, 18] & \text{for } m = 2 \\ [12, 24] & \text{for } m = 3. \end{cases} \quad (5)$$

This design creates a six-point overlap between adjacent modules (e.g., scores 6–12 are possible in both modules 1 and 2), yielding a smoother distribution across the full scale. The global range of the final scores is $F_i \in [0, 24]$. This approach aligns the categorical judgments of the LLM with the continuous nature of levels of psychological constructs and enables finer-grained distinctions in measurement.

3.5. Evaluation of LLM-as-A-Judge

In this study, we adopted the methodology introduced by Calderon et al. [29] to assess the efficacy of using LLMs as evaluators of survey responses. Their framework, known as the Alternative Annotator Test (alt-test), provides a statistical approach to determine whether LLMs can reliably replace human annotators in specific tasks, including judgment tasks.

The alt-test involves comparing LLM-generated annotations against a small, curated subset of human-annotated examples. By analyzing the agreement between the LLM and human annotators, the procedure quantifies the reliability and alignment of the LLM's judgments. This method allows researchers to justify the use of LLMs as substitutes for human evaluators in contexts where human annotation is resource-intensive or impractical.

In analyzing open-ended responses to large-scale surveys, we recommend utilizing this framework to evaluate the performance of LLMs in classifying survey responses into predefined emotional levels. By applying the alt-test, we ensure that the LLM's classifications are statistically comparable to human judgments, thereby validating the LLM's effectiveness as a “judge” in this context.

Furthermore, this analysis allows for the evaluation and comparison of multiple LLMs to determine which model provides the most reliable and accurate judgments. An example output is presented in Figure 3. According to the alt-test, an LLM with a winning rate greater than 0.5 can be considered a valid substitute for a human judge. In this figure, GPT-4.1, GPT-5, and Gemini-2.5-Flash meet this criterion. To facilitate comparison across models, the advantage probability score is used. This measure calculates the probability that the LLM is as good or better than a randomly chosen annotator. In Figure 3, GPT-5 achieves the highest performance as a judge. In the following section, we illustrate our proposed methodology with a case study on measuring aversion toward cockroaches.

```
gpt-4.1 classification [PASSED]: Winning Rate=1.00 Advantage Probability=0.88
gpt-4o classification [FAILED]: Winning Rate=0.00 Advantage Probability=0.81
gpt-5 classification [PASSED]: Winning Rate=1.00 Advantage Probability=0.89
gemini-2.0-flash classification [FAILED]: Winning Rate=0.33 Advantage Probability=0.85
gemini-2.5-flash classification [PASSED]: Winning Rate=1.00 Advantage Probability=0.87
```

Figure 3. Example output from the alt-test methodology.

3.6. Case Study

To test the performance of our new methodology, we chose a psychological construct that is likely to demonstrate a pronounced ceiling effect—the aversion to cockroaches. Based on the findings of Gish et al. [18], we expected the data to show a ceiling effect when respondents rate their aversion toward cockroaches using a Likert-type question.

Although entomophobia and arachnophobia are generally defined as intense fears of insects and spiders, negative attitudes toward insects that do not sting or bite humans are more often driven by disgust than by fear [16]. In our study, the distinction between the fear component and the disgust component of entomophobia is immaterial, since it is the overall aversion and its consequences that are of interest. Therefore, instead of “entomophobia,” henceforth we use “cockroach aversion” and “insect aversion” to refer to general aversion toward cockroaches and insects. The next subsections describe the experimental setup and analysis models used to analyze the cockroach aversion survey responses.

3.6.1. Survey Administration

We distributed a computerized survey using the online platforms Prolific (www.prolific.com) and Qualtrics (www.qualtrics.com). We surveyed 510 adults (255 men, 255 women) between the ages of 25 and 80, residing in the contiguous United States. We chose only participants who fulfilled the following criteria: living with spouse/partner; past approval rate: 100%; weekly working hours on Prolific: 0–5 h; household size of 2–6 persons; no romantic partner on Prolific; fluent in English; primary language: English.

The survey was given the generic name “personal views and habits” and was vaguely described as part of a study on how people interact with their everyday environment. This was important for minimizing selection bias that could have been caused by the possibility that people with a high level of cockroach aversion would skip a survey if they knew, *a priori*, that it focused on cockroaches. Participants were informed that the survey would take no more than 10 min to complete, and they were encouraged to provide honest and detailed responses. Of 529 responses to the survey, we accepted 510. Ten respondents did not complete the survey, one was timed out, five were excluded due to very short responses, and three were identified as bots/AI by our non-human filtering procedure. The survey was conducted during June–August of 2025.

3.6.2. Questionnaire Structure

The questionnaire consisted of two parts. The first part was designed to assess and quantify participants’ level of cockroach aversion, following the methodology proposed in this paper. The second part served as a control test for this methodology, measuring cockroach aversion through a conventional Likert scale-based closed-ended question: On a scale of 1 to 7, how would you rate the level of aversion (disgust or fear) you feel toward cockroaches? Where 1 = No aversion at all and 7 = Extreme aversion.

The survey began with a routing question that asked “How aversive (scary or disgusting) do you find cockroaches?” for which respondents could choose one of three answers: not aversive at all, somewhat aversive, or very aversive. Based on their response to this question, they were directed to one of three different modules.

In each module, two questions were designed to invoke a relatively low level of aversion, two aimed at invoking a medium level of aversion, and another two were meant to invoke a high level of aversion (see summary in Table 1, and the full survey in Appendix A). Thus, the questions in each module covered a range of potential aversion levels.

3.6.3. Survey Development

The open-ended questions were composed and refined by two experts—an expert on entomophobia and an NLP expert. All parts of the cockroach aversion quantification section of the survey were iteratively tested, corrected, and refined through multiple stages of preliminary, small-scale surveys and performance analyses. These preliminary small-scale surveys included a total of 165 participants from the same target population, who were then excluded from the study. The response classification instructions provided to

the LLMs were composed and refined during the development phase using the responses collected in the small-scale preliminary surveys.

Table 1. The topics of the open-ended questions used in the study. Each question presented a hypothetical scenario and then asked the participants to describe what they would think or do in such a situation. The topics are arranged by module and by the level of aversion each pair of questions was designed to invoke.

Aversion Level the Questions were Intended to Invoke	Question #	Topics of the Questions in Module 1	Topics of the Questions in Module 2	Topics of the Questions in Module 3
Low	1, 2	Seeing a live cockroach in a place where food is prepared	Being with a dead cockroach in an indoor space	Being exposed to symbolic representations of cockroaches
Medium	3, 4	Picking up or holding a live cockroach	Seeing live cockroaches outdoors	Hearing or reading about someone else's encounter with cockroaches
High	5, 6	Ingesting a cockroach or food that came in contact with a cockroach	Being with a live cockroach in an indoor space	Looking at photography of cockroaches on a screen

3.6.4. LLM Candidates

To identify the most suitable LLM for analyzing the survey responses, we first sampled 108 responses from all open-ended survey questions (2–3 responses per question across all modules). According to Calderon et al. [29], a sample of 50–100 is sufficient for testing the LLM's performance as a judge in various applications. In addition, we made sure that the sampled responses were from different respondents to generate a representative sample of responses and levels of aversion. The authors of this paper manually annotated these responses into three categories based on their subjective judgment: Type A for minimal to no aversion, Type B for moderate aversion, and Type C for extreme aversion. The classification into types was carried out with consideration of the module context, namely the initial aversion level that participants declared in the routing question. We evaluated the following LLMs: GPT-4.1 (OpenAI), GPT-4o (OpenAI), GPT-5 (OpenAI), Gemini-2.0-Flash (Google), Gemini-2.5-Flash (Google). Although more reasoning models are available, we avoided them after observing that during preliminary testing these models tended to over-interpret responses, occasionally inferring intentions or emotional subtext that did not align with a literal reading of the text. The full prompt provided to the LLMs appears in Appendix B.

We evaluated the LLMs' performance using the methodology defined in Calderon et al. [29], which was specifically designed to evaluate LLM-as-a-judge models. The performance outcomes are presented in Figure 3. Based on these results, we selected GPT-4.1 and GPT-5 as judging models for the full survey analysis.

3.6.5. Data Analysis

Each participant's open-ended responses were evaluated by the two selected LLMs. Both models were accessed via API in zero-shot mode with a temperature of 0, to ensure deterministic outputs. Responses were submitted individually. Each prompt included the survey question, the participant's answer, and a rubric outlining detailed classification rules. The rubric instructed the LLMs to assign responses to one of three categories: Type A (low aversion), Type B (moderate aversion), or Type C (high aversion). In addition, the LLMs could classify a response as *Irrelevant* if it did not address the question, or as *Not Sure* if it was too vague or contradictory. In these two cases, the models were required to provide

an explanation. Such responses were flagged and subsequently reclassified independently by one of the researchers according to the same rubric. We evaluated the prompt and the questions in the survey by comparing the classifications of the LLMs to the researchers' classifications on an independent set of users who answered the survey in its pilot version.

3.6.6. Score Calculation

In our case study, each module consisted of three item pairs, corresponding to three levels of cockroach aversion. We implemented the 6-point overlap scoring scheme described in Section 3.4.3. We also tested 5- and 7-point overlap variants, but found no substantial differences among them that would favor one over the others in a goodness-of-fit test (Kolmogorov-Smirnov) compared to a skewed-normal distribution. We selected the skewed-normal as a reference because, when surveying a large population about their fear of cockroaches, the distribution of responses is expected to approximate normality due to the central limit theorem, while also exhibiting slight skewness. This skewness arises because many respondents report some aversion to cockroaches, resulting in most scores clustering toward the upper end of the scale, with fewer but more extreme values extending toward the lower end.

3.6.7. Bot and LLM Response Detection

To ensure that responses reflected genuine human input, we conducted a post hoc screening as recommended in Section 3.4.2. This screening was done both manually, and with the help of ChatGPT (the prompt used for bot detection is included in Appendix C). In addition, at the end of the survey, we asked participants to report their age. We compared the reported age with their registered age on Prolific, and discrepancies were flagged as possible indicators of non-human participants. All flagged responses were reviewed manually by the authors to decide if the participants who provided them should be excluded.

4. Results

4.1. Ceiling Effect in Likert Scale Rating

As noted earlier, we sought to confirm that Likert-type questions produce a ceiling effect when measuring cockroach aversion. Figure 4 reveals a pronounced ceiling effect: the majority of participants reported very high aversion, with most responses concentrated at values 6 and 7. While the cutoff for a ceiling effect is typically 15%, in our case 23% of the total sample selected the highest rating (7), and 45% selected 6 or 7. Among men, 14.5% selected 7, and 32.5% selected 6 or 7. Among women, 33% selected 7, and 57% selected 6 or 7.

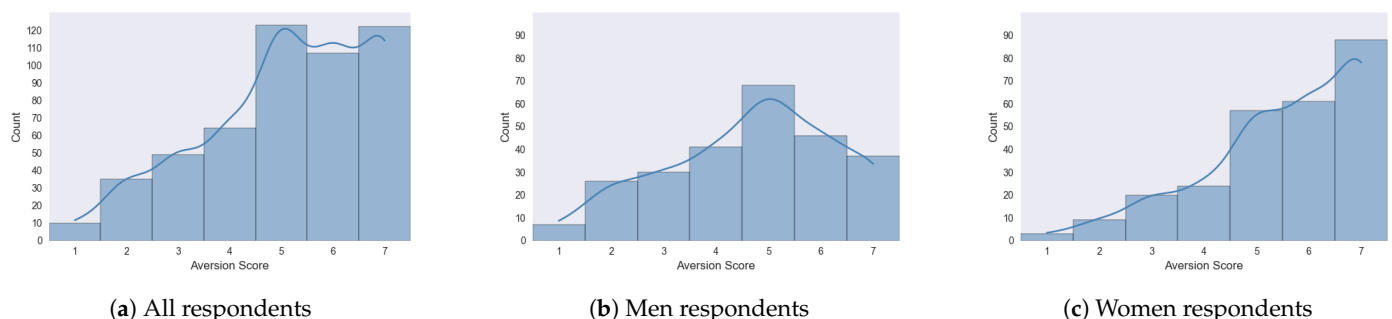


Figure 4. Histograms of Likert scale self ratings of cockroach aversion in response to the question: On a scale of 1 to 7, how would you rate the level of aversion (disgust or fear) you feel toward cockroaches? Where 1 = No aversion at all and 7 = Extreme aversion.

4.2. Ceiling Effect Mitigation with Open-Ended MST

Figure 5 presents the distribution of scores assigned by the LLMs to all survey respondents. As illustrated, the distributions of scores resembled a normal distribution with a slight right skew. Scores were generally concentrated at the upper-middle range of the scale (roughly between 15–17 for the whole sample, 15–16 for men and 16–18 for women). Importantly, the distributions show no evidence of a ceiling effect, as few respondents were classified at the extreme top values of the scale.

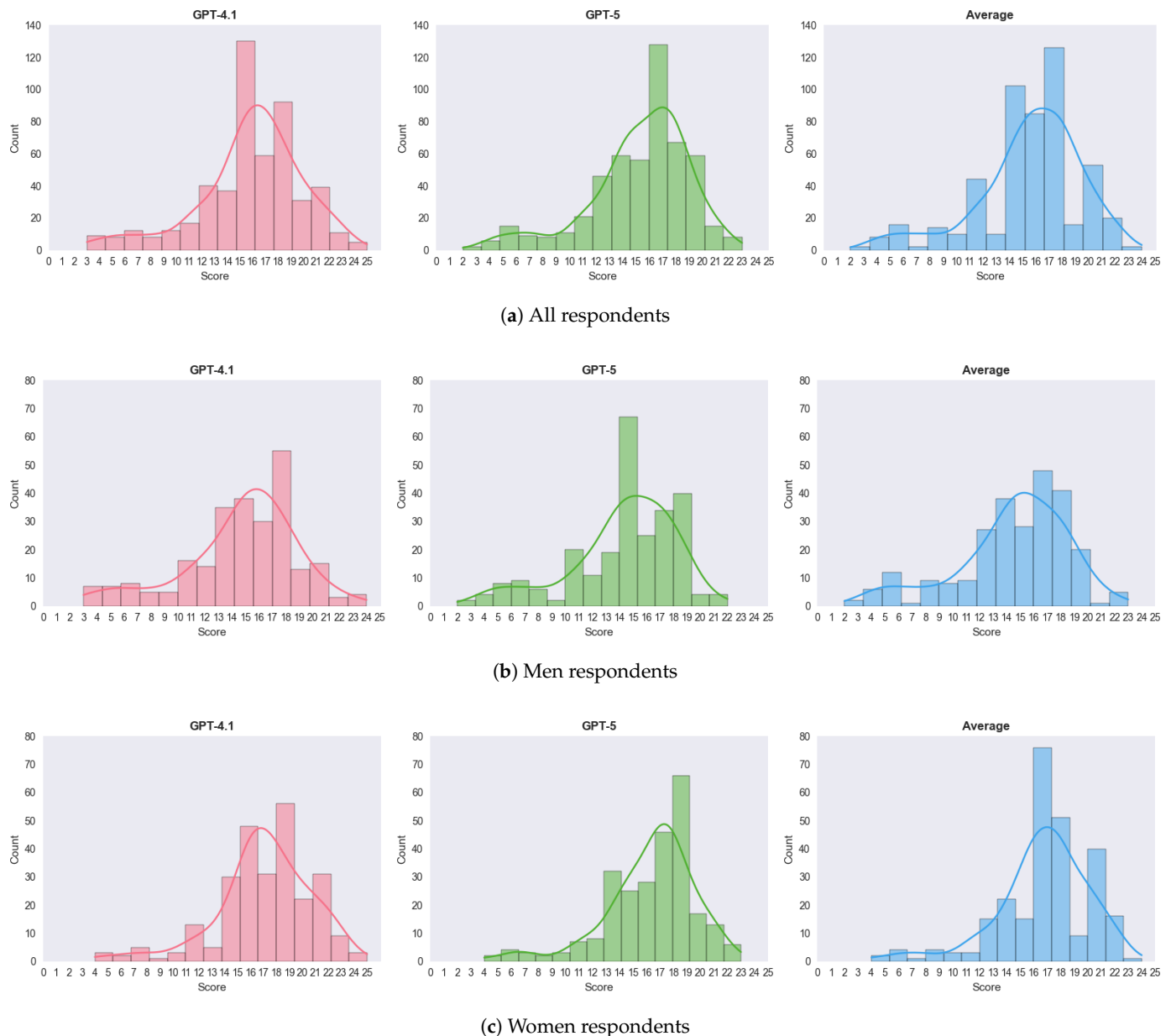


Figure 5. Histograms of scores of cockroach aversion levels calculated with the formula in Section 3.4.3 based on LLM classifications of responses to open-ended questions.

4.3. LLM Performance

The performance of the LLMs as judges was tested on the 108 samples that were classified by human experts. In addition to the outcomes from the alternative test (Figure 3), we also report agreement measures for these classifications. Specifically, we computed Cohen's kappa between each LLM and the majority vote derived from the expert annotations,

and calculated the overall multi-rater reliability using Fleiss's kappa across all annotators, including the LLMs (Table 2).

Table 2. Agreement between LLMs and human judges. Cohen's kappa is computed between each LLM and the majority vote of the human judges. Fleiss' kappa includes all three human annotators and the LLM.

LLM	Cohen's Kappa vs. Majority Vote	Fleiss' Kappa (All Annotators)
GPT-4.1	0.750	0.782
GPT-5	0.719	0.780

Overall, the two LLMs, GPT-4.1 and GPT-5, classified 3060 responses to the questions in the survey, of which only 118 (3.8%) were classified as *Irrelevant* or *Not Sure* by at least one LLM. The counts of these responses appear in Table 3. Upon manual inspection, we identified 4 cases where GPT-4.1 should have assigned a label (A, B, or C) and 7 cases where GPT-5 should have assigned a label (i.e., the *Irrelevant* or *Not Sure* labels were not justified).

Table 3. Counts of responses classified as *Irrelevant* or *Not Sure* by each LLM and jointly by both models.

LLM	Irrelevant	Not Sure
GPT-4.1	20	33
GPT-5	13	63
Both GPT-4.1 & GPT-5	12	16

Cost-Effectiveness Analysis

Beyond reliability and scalability, the MST-LLM framework demonstrated substantial cost efficiency. The total API cost for processing and classifying the entire dataset of responses in this study, using both GPT-4.1 and GPT-5, was \$69.78 (USD); see Figure 6. Therefore, running an analysis with only one chosen model will cost around \$40. To contextualize this figure, we estimated the cost of equivalent human labor. Assuming a trained research assistant could code open-ended responses at a sustained rate of 60 responses per hour (1 min per response) at a standard hourly rate of \$25, manually coding the same dataset would require approximately 50+ hours of labor, totaling over \$1250. Thus, the LLM-driven approach reduced costs by a factor of approximately 18, while also eliminating the logistical overhead of hiring, training, and managing human raters. For larger datasets (e.g., $N = 15,000$), this disparity would likely widen, making the difference between a few hundred dollars in API credits versus tens of thousands of dollars in human labor.

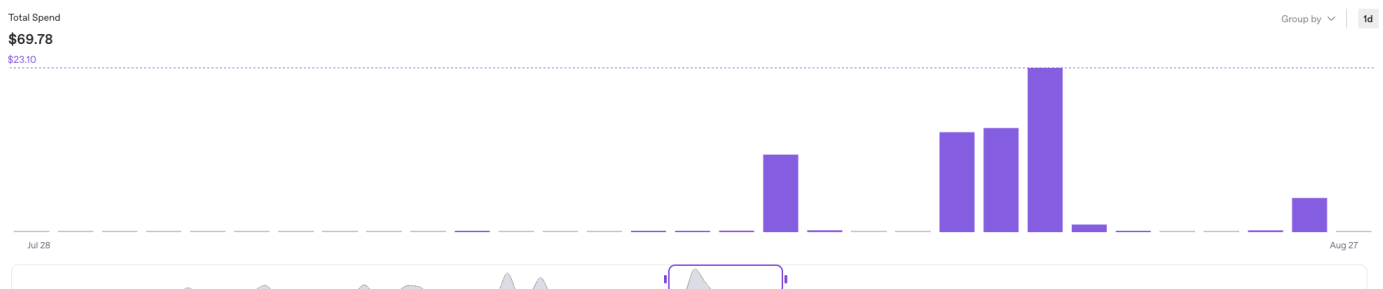


Figure 6. Costs of API usage for analyzing the survey responses.

4.4. Prompting Methodologies

To assess the sensitivity of our results to the specific formulation of the classification prompt, we conducted additional experiments using two alternative prompting strategies:

(i) a one-shot prompting approach that includes a single illustrative example, and (ii) a rephrased version of the original prompt that preserves its semantic intent while modifying surface-level wording (using GPT 5 as a rephraser). These experiments are intended to evaluate the robustness of the scoring distribution to minor prompt variations.

Figure 7a,b present the resulting score distributions for GPT 5 under the one-shot and rephrased prompting conditions, respectively. The rephrased prompt yields a distribution that is highly consistent with that obtained using the original prompt formulation, both in terms of central tendency and overall shape. This indicates that the model's scoring behavior is largely invariant to modest linguistic rewording, suggesting that the underlying classification rubric defines a stable decision boundary.

By contrast, the one-shot prompting condition produces a noticeably less smooth score distribution, with greater variance across score bins. Although the overall score range remains comparable, the increased variability suggests that relying on a single example provides insufficient contextual grounding for the model to consistently apply the classification criteria at scale.

At the same time, the one-shot condition results in a substantially lower number of uncertainty-related responses. Specifically, the number of “not sure” classifications decreases from 87 under the rephrased prompt to 39 under one-shot prompting, while “irrelevant” responses decrease from 9 to 6. This highlights a trade-off between decisiveness and distributional stability: one-shot prompting encourages more decisive classifications, but at the cost of increased noise in the aggregate score distribution.

Overall, these results demonstrate that the proposed scoring methodology is robust to moderate prompt rephrasing, while also underscoring the importance of prompt structure and contextual richness in producing stable score distributions. For this reason, we adopt the original prompting strategy in our main experiments, as it offers the best balance between robustness and consistency.

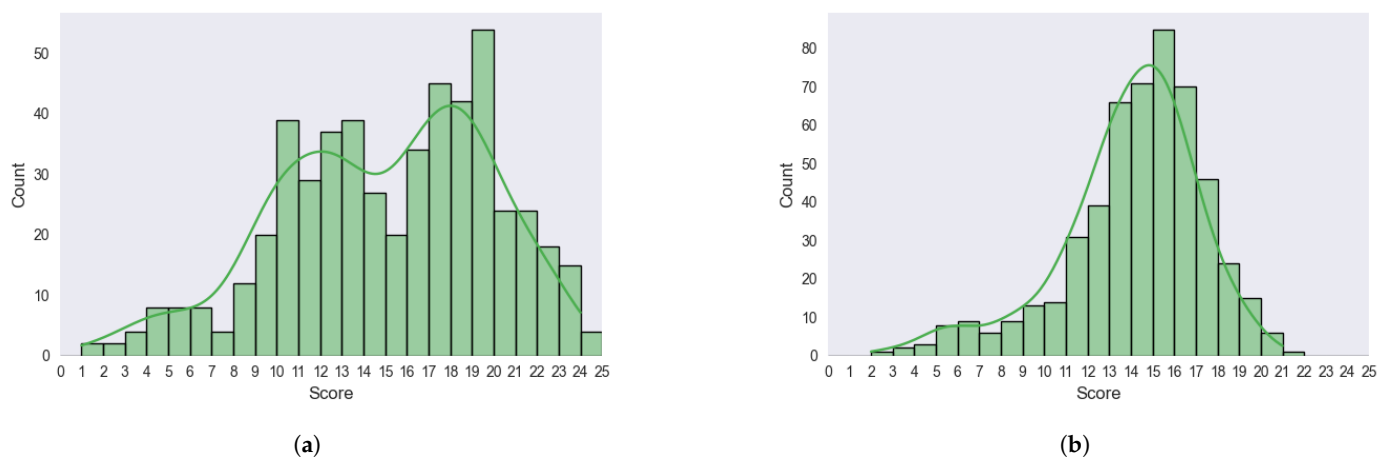


Figure 7. Histograms of scores for all respondents of cockroach aversion levels calculated with the formula in Section 3.4.3 based on LLM classifications of responses to open-ended questions. (a) LLM scores generated by GPT 5 with one shot prompting. (b) LLM scores generated by GPT 5 with rephrased prompt.

5. Discussion

In this study, we addressed a persistent challenge in psychological measurement: the ceiling effect, which can occur when traditional self-report numeric scales are used to quantify extreme emotions and affective attitudes [48,49]. This measurement artifact may obscure the real variation within a population, leading to inaccurate conclusions and hindering scientific progress [12–15]. To overcome this limitation, we introduced a

novel framework that combines a multistage testing design with open-ended questions and leverages the analytical capabilities of LLMs for scalable, automated response analysis. Since this framework depends on interpreting free-text responses, its feasibility for medium- and large-scale studies requires the use of automated text evaluators, such as LLM-as-a-judge models.

Our case study on cockroach aversion allowed us to test the ability of our new framework to mitigate a clear and extreme ceiling effect, one that until now, no study had been able to suppress. While the conventional Likert-type question produced a highly skewed distribution with many respondents selecting the highest possible rating, the MST-LLM framework revealed a distribution of the data that was likely obscured by the strong ceiling effect in the Likert-type question (as seen in Figure 4). This result demonstrates that the proposed method can reveal meaningful variability that fixed-choice items mask. In doing so, it supports our central hypothesis that replacing fixed-choice items with carefully designed open-ended questions enables a more realistic and fine-grained representation of the nuances of extreme psychological constructs.

The validation process further confirmed the viability of this approach. Using the alt-test, we demonstrated that LLMs can serve as reliable and consistent substitutes for human judges when classifying open-ended survey responses. This was also supported by traditional agreement measures such as Cohen's and Fleiss' kappas, which showed substantial agreement with expert judgments (Table 2). The strong agreement between LLM-generated classifications and expert annotations provides empirical support for using LLMs as evaluators in this context.

Moreover, across a dataset of more than three thousand automatically classified responses, only a small number were labeled as *Irrelevant* or *Not Sure* by at least one model. Manual review of these flagged responses, in which the model was not sure in its classification, showed that almost all of them reflected true misclassifications, suggesting that the models generally interpreted respondents' intentions and levels of aversion with minimal uncertainty. Taken together, these results suggest that the selected LLMs were capable of maintaining both precision and restraint, avoiding overinterpretation while still capturing meaningful distinctions in free-text responses.

It is important to note, however, that the success of this framework likely depends not only on the capabilities of the LLMs themselves but also on the thoughtful design and refinement of both the survey questions and the classification prompts. In our study, these components were iteratively developed through multiple rounds of small-scale testing. Questions that elicit sub-optimal responses and prompts that allow overly permissive interpretation could undermine classification reliability. This highlights that the application of AI in this context is not merely a plug-and-play solution but requires a dedicated development phase where prompt instructions are systematically engineered and validated to ensure they function as precise measurement tools.

6. Conclusions

This study introduces the MST-LLM framework, a novel methodological approach that integrates discriminative questioning with scalable AI-driven evaluation. By moving beyond traditional fixed-choice items, we demonstrated that carefully engineered LLM pipelines can effectively mitigate the ceiling effect, revealing meaningful variability in the measurement of extreme psychological constructs, which until now would have been obscured.

Our findings confirm that when guided by systematic prompt constraints, LLMs can serve as reliable, scalable evaluators that align closely with expert human judgment. This offers a robust solution to the limitations of manual qualitative analysis, enabling

researchers to process large-scale datasets with the nuance of open-ended inquiry and the speed and accuracy of quantitative methods.

The implications of this work extend beyond the specific case of cockroach aversion. We provide a generalizable AI application for researchers in psychology, mental health, and political science—fields where extreme responding sometimes compromises data quality. By validating this framework, we offer a pathway to improved measurement precision, demonstrating how applied AI can resolve long-standing statistical bottlenecks in social science research. Future work should focus on expanding this framework to other affective domains while further refining the evaluation protocols to maximize generalizability across different LLM architectures.

Author Contributions: Conceptualization, M.G. and R.D.; methodology, M.G. and R.D.; software, A.N.; validation, A.N., M.G. and R.D.; formal analysis, R.D.; investigation, A.N., M.G. and R.D.; data curation, A.N.; writing—original draft preparation, M.G. and R.D.; writing—review and editing, M.G. and R.D.; visualization, A.N.; supervision, M.G. and R.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of The University of Haifa (protocol code [030/25] and date of approval [20 January 2025]).

Informed Consent Statement: All participants gave their written informed consent to participate in online surveys when signing up to the Prolific platform.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MST	Multistage Testing
LLM	Large Language Model
NLP	Natural Language Processing
CAT	Computerized Adaptive Testing
IRT	Item Response Theory
AI	Artificial intelligence
alt-test	Alternative Annotator Test

Appendix A. Open-Ended Questions—Full Version

Appendix A.1. Module 1

1. You've just finished making yourself dinner. You place the plate on the table, but before sitting down, you go to the bathroom. When you return, you see a cockroach standing on your food. As you approach, it scurries off and upon reaching the edge of the table, it spreads its wings and flies out the window. What do you do with your dinner, and why do you do that?
2. In some countries, fried cockroaches are regarded as a delicacy. Imagine a friend returning from one of these countries with a bag of fried cockroach snacks and offering you a taste. Would you try one? Under what conditions might you be willing to give it a try?
3. Your friend has gone on vacation and asked you to feed their pet hissing cockroaches. When you arrive at their house, you notice that, despite the winter cold making the

cockroaches lethargic, one has managed to escape and is sitting motionless near the terrarium. What would you do with the escaped cockroach, and what factors would influence your choice?

4. While visiting a natural history museum with your child, you come across an exhibit featuring live insects. In one display, an open terrarium allows visitors to pick up and hold cockroaches. Your child wants to try holding one but asks you to go first. How would you feel about it? What could affect your decision?
5. Imagine you're sitting in a fancy restaurant, waiting for your dinner to arrive, when you notice a cockroach scurrying across the floor before disappearing under a counter. Would you take any immediate or delayed action in response? What thoughts or emotions do you think you would think or feel?
6. It's late in the evening, and as you walk into the kitchen, you notice two long antennae peeking out from behind a jar. Suspecting it might be a cockroach, do you take any action? Why do you think you would behave that way?

Appendix A.2. Module 2

1. You're seated in a movie theater, waiting for the movie to begin, when you spot two cockroaches scurrying on the floor nearby. Moments later, the lights dim and the movie begins. Do you take any action in response? If so, what do you do and what motivates your choice of action?
2. You've just returned home in the evening. As you step into the living room and turn on the light, you spot a cockroach scurrying across the floor before it disappears under the couch. What do you do? Why would you choose that course of action?
3. While driving through town, you stop at a red light and glance to your left, noticing a bunch of cockroaches crawling out of a storm drain. Does this sight evoke any feelings in you? If so, what emotions do you think you might experience?
4. While hiking through the woods, you pause to rest on a tree stump, taking in the natural surroundings. As you look around, you notice a cockroach crawling on a nearby log. How does this sight make you feel, and how do you think you would respond after noticing it?
5. Imagine you're alone in your living room, watching a movie on your television, when you suddenly notice a dead cockroach lying on its back in the corner of the room. Would you take any action? If so, would you address it immediately or wait until later, and what would influence the timing of your response?
6. Suppose you are visiting a large museum, and one of the exhibits focuses on the natural world, featuring a collection of various preserved cockroach species displayed safely behind glass. Would you choose to explore this exhibit? If not, why would you avoid it? If you do decide to go in, what would be your motivation to enter?

Appendix A.3. Module 3

1. Imagine you're at a party, and someone begins sharing their passion for insect photography, showing photographs of cockroaches they've captured on their phone. How would you respond to this? Why do you think you would act this way?
2. Imagine you are browsing your TV alone, looking for something to watch, and come across a highly rated documentary about the life of cockroaches, highlighting their complex social interactions and sophisticated communication skills. How would you feel about watching this documentary? What factors might influence your decision?
3. While you are at work, your partner calls to tell you they just spotted a cockroach in the kitchen. They managed to catch it and throw it out the window. What would be your reaction to this news? What would you say or ask?

4. Imagine scrolling through your social media feed and coming across a post from a friend describing how they discovered a small family of cockroaches living in the cabinet under their kitchen sink, and got rid of them. If you were to respond to their post, what would you write in your response to the post? Why would that be your response?
5. Suppose you are at a friend's house, and their child asks you to play with them using a collection of realistic plastic toy cockroaches. How would you handle the situation, and what feelings might it evoke in you?
6. Imagine coming across a children's book featuring friendly, cartoonish cockroach characters designed to teach kids about nature. Would you consider reading this book to your child? Why or why not?

Appendix B. Prompt for LLM Classification

Task Overview

You will analyze a response to a survey question and classify it into one
 → of five categories: A, B, C, Irrelevant, or Not Sure. You will be given
 → the question, the response, and my specific instructions on how to
 → analyze and label the answer to the question.

A, B, or C - Based on compatibility with provided category descriptions.

Irrelevant - If the response does not address the question.

Not Sure - If the response is too ambiguous or does not clearly fit any
 → category.

You will receive specific classification guidelines for each question. If a
 → response falls between two categories, apply the provided rules to
 → determine the best fit.

Classification Guidelines

Literal First: Prioritize a literal reading unless clear figurative or
 → emotional intent is present.

Tone Sensitivity: If the response contains emotional cues (e.g.,
 → enthusiasm, reluctance, disgust), prioritize tone over strict literal
 → meaning.

Colloquial Interpretation: Assume informal, conversational language.

Contradictory Responses: If the response expresses both agreement and
 → refusal (e.g., "I wouldn't eat it, but maybe I'd try a bite"), classify
 → as Not Sure.

Example Explanation: "Conflicting intent makes classification unclear."

Figurative Language: Classify based on the intended meaning, not literal
 → words.

Not Sure Responses: Use this category if the response is too vague,
 → ambiguous, or doesn't fit the defined classifications.

Special Language Handling:

"Most likely" / "Probably" → Definitive Yes

"Most likely not" / "Probably not" → Definitive No

Expected Output Format (JSON)

```
{
  "classification": "A/B/C/Irrelevant/Not Sure",
  "explanation": "Concise reason for classification"
```

```
}
```

The question and the example

Question: ...

Answer type A: ...

Answer type B: ...

Answer type C: ...

Classification Task:

Please classify the following response according to the provided guidelines

→ and output it in JSON format.

Response to classify:

Answer: ...

Prompt for responses from Module 1

The Questions and Examples

Question 1:

You've just finished making yourself dinner. You place the plate on the

→ table, but before sitting down, you go to the bathroom. When you
→ return, you see a cockroach standing on your food. As you approach, it
→ scurries off and upon reaching the edge of the table, it spreads its
→ wings and flies out the window. What do you do with your dinner, and
→ why do you do that?

Type A: I eat it the way it is, without expressing any disgust at all.

Type B: I eat only the part that did not contact the cockroach. Or, I eat
→ the food but I express some discomfort.

Type C: I throw out all the food without eating any of it.

Question 2:

In some countries, fried cockroaches are regarded as a delicacy. Imagine a

→ friend returning from one of these countries with a bag of fried
→ cockroach snacks and offering you a taste. Would you try one? Under
→ what conditions might you be willing to give it a try?

Type A: I will definitely try it. I express enthusiasm toward tasting it.

Type B: Maybe. I cannot say for sure if I will try it or not. Or, I would
→ try it under certain conditions.

Type C: There is a low or no chance that I will try it.

Question 3:

Your friend has gone on vacation and asked you to feed their pet hissing

→ cockroaches. When you arrive at their house, you notice that, despite
→ the winter cold making the cockroaches lethargic, one has managed to
→ escape and is sitting motionless near the terrarium. What would you do
→ with the escaped cockroach, and what factors would influence your
→ choice?

Type A: Eventually I pick it up and put it back into the terrarium. I do
 → not say that I would use an object that is not my bare hand to help me
 → pick it up. I express concern about the wellbeing of the escaped
 → cockroach or the other cockroaches in the terrarium.

Type B: Eventually I pick it up and put it back into the terrarium. I do
 → not say that I would use an object that is not my bare hand to help me
 → pick it up. I do not express concern about the wellbeing of the escaped
 → cockroach or the other cockroaches in the terrarium.

Type C: Eventually I pick it up and put it back into the terrarium, but I
 → use an object that is not my bare hand to help me pick it up. Or, I do
 → not pick it up by myself.

Question 4:

While visiting a natural history museum with your child, you come across an
 → exhibit featuring live insects. In one display, an open terrarium
 → allows visitors to pick up and hold cockroaches. Your child wants to
 → try holding one but asks you to go first. How would you feel about it?
 → What could affect your decision?

Type A: I would gladly and happily hold the cockroach. I express enthusiasm
 → toward holding a cockroach.

Type B: I would do it, but I do not express enthusiasm about holding it.

Type C: I am not sure if I would do it. Or, there is a low or no chance
 → that I will do it.

Question 5:

Imagine you're sitting in a fancy restaurant, waiting for your dinner to
 → arrive, when you notice a cockroach scurrying across the floor before
 → disappearing under a counter. Would you take any immediate or delayed
 → action in response? What thoughts or emotions do you think you would
 → think or feel?

Type A: Eventually, I eat that dinner in the restaurant, or I do not
 → mention that I won't. I also do not mention any effect on my future
 → inclination to dine in that restaurant again. I do not say that I would
 → feel uncomfortable eating there.

Type B: Eventually, I eat that dinner in the restaurant, or I do not
 → mention that I won't, but I do mention that I will not dine in that
 → restaurant again or that I will not recommend it to others. I might
 → express moderate discomfort or concern about hygiene.

Type C: Eventually, I leave the restaurant without eating that dinner. If I
 → don't say that I would leave, I express significant discomfort,
 → aversion or concern about the hygienic conditions in that restaurant.

Question 6:

It's late in the evening, and as you walk into the kitchen, you notice two
 → long antennae peeking out from behind a jar. Suspecting it might be a
 → cockroach, do you take any action? Why do you think you would behave
 → that way?

Type A: I let it be. Or, I do not make any special efforts to catch it.

Type B: I focus only on dealing with that individual cockroach myself, and
 → I do not express any concern about the possibility of having other
 → cockroaches in my home. I do not take additional prophylactic measures
 → against a potential infestation.

Type C: I am concerned about the possibility of having more cockroaches in
 → the house, and I take measures to rule out or treat a potential
 → infestation.

Prompt for responses from Module 2

The Questions and Examples

Question 1:

You're seated in a movie theater, waiting for the movie to begin, when you
 → spot two cockroaches scurrying on the floor nearby. Moments later, the
 → lights dim and the movie begins. Do you take any action in response? If
 → so, what do you do and what motivates your choice of action?

Type A: My answer indicates that eventually I will remain seated in the
 → same seat throughout the movie. I do not express significant
 → discomfort.

Type B: My answer indicates that eventually I will remain seated in the
 → same seat throughout the movie, but I do express significant discomfort
 → or efforts to avoid contact with the cockroaches.

Type C: Any answer that indicates a high level of aversion toward
 → cockroaches, and does not fit the above descriptions of Type A or Type
 → B.

Question 2:

You've just returned home in the evening. As you step into the living room
 → and turn on the light, you spot a cockroach scurrying across the floor
 → before it disappears under the couch. What do you do? Why would you
 → choose that course of action?

Type A: I try to catch or kill that individual cockroach. I do not take
 → further action, and I do not express concerns about a possible
 → infestation.

Type B: On top of trying to catch or kill that individual cockroach, I am
 → concerned about the possibility of having more cockroaches in the
 → house.

Type C: Any answer that indicates a high level of aversion toward
 → cockroaches, and does not fit the above descriptions of Type A or Type
 → B.

Question 3:

While driving through town, you stop at a red light and glance to your
 → left, noticing a bunch of cockroaches crawling out of a storm drain.
 → Does this sight evoke any feelings in you? If so, what emotions do you
 → think you might experience?

Type A: I look at them with some interest or curiosity, or with no negative
 → emotions at all.

Type B: I feel slightly uneasy.

Type C: Any answer that indicates a high level of aversion toward
 → cockroaches, and does not fit the above descriptions of Type A or Type
 → B.

Question 4:

While hiking through the woods, you pause to rest on a tree stump, taking
 → in the natural surroundings. As you look around, you notice a cockroach
 → crawling on a nearby log. How does this sight make you feel, and how do
 → you think you would respond after noticing it?

Type A: I express positive feelings of interest in the cockroach.

Type B: It doesn't bother me much. I do not move farther away from the
 → cockroach.

Type C: My answer indicates a medium or high level of aversion toward
 → cockroaches or insects, and it does not fit the above descriptions of
 → Type A or Type B.

Question 5:

Imagine you're alone in your living room, watching a movie on your
 → television, when you suddenly notice a dead cockroach lying on its back
 → in the corner of the room. Would you take any action? If so, would you
 → address it immediately or wait until later, and what would influence
 → the timing of your response?

Type A: The dead cockroach does not bother me enough for me to pause or
 → stop the movie in order to take care of it promptly. Therefore, I will
 → take care of the dead cockroach after I finish watching.

Type B: My answer indicates that I would pause or stop the movie in order
 → to take care of it promptly. Or, I do not specify the timing of my
 → action. I do not express concern about the possible presence of
 → additional cockroaches in the house.

Type C: Any answer that indicates a high level of aversion toward
 → cockroaches, and does not fit the above descriptions of Type A or Type
 → B.

Question 6:

Suppose you are visiting a large museum, and one of the exhibits focuses on
 → the natural world, featuring a collection of various preserved
 → cockroach species displayed safely behind glass. Would you choose to
 → explore this exhibit? If not, why would you avoid it? If you do decide
 → to go in, what would be your motivation to enter?

Type A: I would definitely enter the exhibit. I express enthusiastic
 → interest specifically about cockroaches or a great passion toward
 → insects, animals or nature.

Type B: I would enter the exhibit. I express moderate interest in
 → cockroaches or a general interest in insects, animals or nature.

Type C: Any answer that indicates some level of aversion toward cockroaches
 → or insects, and does not fit the above descriptions of Type A or Type
 → B.

Prompt for responses from Module 3

The Questions and Examples

Question 1:

Imagine you're at a party, and someone begins sharing their passion for
→ insect photography, showing photographs of cockroaches they've captured
→ on their phone. How would you respond to this? Why do you think you
→ would act this way?

Type C: I will absolutely not look at any of the pictures, no matter what.
→ I express a clear objection to being exposed to these pictures. I
→ prefer to disengage altogether, since I view that person as weird or
→ creepy.

Type B: I try to avoid looking at the pictures. I may look at a few and
→ then change the subject or find an excuse to avoid the uncomfortable
→ situation.

Type A: Any answer that indicates a medium to low level of aversion toward
→ cockroaches, and does not fit the above descriptions of Type C or Type
→ B.

Question 2:

Imagine you are browsing your TV alone, looking for something to watch, and
→ come across a highly rated documentary about the life of cockroaches,
→ highlighting their complex social interactions and sophisticated
→ communication skills. How would you feel about watching this
→ documentary? What factors might influence your decision?

Type C: I will absolutely not watch it, nor will I give it a chance. I
→ express a very clear objection to watching something like this. In
→ addition, my answer must include explicit repulsion and disgust toward
→ such a television show, including the mentioning of nausea or fear. I
→ also specifically say that I do not think it is appropriate to
→ broadcast such programs at all.

Type B: Even though I am not 100% sure, I think that there is a very low
→ chance that I will watch it.

Type A: Any answer that indicates a mild or low level of aversion toward
→ cockroaches, and does not fit the above descriptions of Type C or Type
→ B.

Question 3:

While you're at work, your partner calls to tell you they just spotted a
→ cockroach in the kitchen. They managed to catch it and throw it out the
→ window. What would be your reaction to this news? What would you say or
→ ask?

Type C: I think it is absolutely terrible and I express explicit, extreme
→ repulsion and distress. I also demand significant pest control actions
→ in our entire home. My answer indicates that I will absolutely not
→ tolerate even the slightest possibility that there might be more
→ cockroaches in the house, and that I would consider such an incident as
→ an unbearable crisis. I make it clear that I will not come down until
→ the entire home is thoroughly treated against cockroaches.

Type B: In my response, I express intense aversion toward the cockroach or
 → toward other cockroaches that might be in the house. I also express
 → significant concern about the possibility of having additional
 → cockroaches in the house. However, my response is less extreme than
 → what should be considered as a type C response.

Type A: Any answer that indicates a moderate or low level of aversion
 → toward cockroaches, and does not fit the above descriptions of Type C
 → or Type B.

Question 4:

Imagine scrolling through your social media feed and coming across a post
 → from a friend describing how they discovered a small family of
 → cockroaches living in the cabinet under their kitchen sink-and got rid
 → of them. If you were to respond to their post, what would you write in
 → your response to the post? Why would that be your response?

Type C: My post expresses extreme, visceral emotions of aversion or extreme
 → fear. In addition, from the way I phrased my response it is clear that
 → I perceive my friend's situation as a huge, horrific and unbearable
 → crisis.

Type B: My post expresses significant emotions of aversion or significant
 → relief that the infestation was dealt with.

Type A: Any answer that indicates a medium or low level of aversion toward
 → cockroaches, and does not fit the above descriptions of Type C or Type
 → B.

Question 5:

Suppose you are at a friend's house, and their child asks you to play with
 → them using a collection of realistic plastic toy cockroaches. How would
 → you handle the situation, and what feelings might it evoke in you?

Type C: I will absolutely not play with the plastic cockroaches, no matter
 → what. My answer must include expressions of deep aversion and disdain.

Type B: There is a high chance that I would play with the plastic
 → cockroaches, but I clearly indicate that I would feel some discomfort.

Type A: Any answer that indicates a low level of aversion toward
 → cockroaches, and does not fit the above descriptions of Type C or Type
 → B.

Question 6:

Imagine coming across a children's book featuring friendly, cartoonish
 → cockroach characters designed to teach kids about nature. Would you
 → consider reading this book to your child? Why or why not?

Type C: I will absolutely not read that book to my child. My answer must
 → include extreme expressions of deep aversion and disgust. I also say
 → that I think it is inappropriate to read such books to children.

Type B: Maybe. I cannot say for sure if I will read it or not. Or, I would
 → read it despite feeling mild discomfort.

Type A: Any answer that indicates a mild or low level of aversion toward
 → cockroaches, and does not fit the above descriptions of Type C or Type
 → B.

Appendix C. Bot and AI Detection Prompt

Prompt for per_answer_checks:

""You are a careful, literal grader. Judge ONLY the given text.
If uncertain, set booleans to false.

PERSONALITY-CONDITIONAL FRAMING (strict):

Count as TRUE only if the answer uses a hypothetical self-identity framing
→ like:

- "If I were an anxious person, I'd do...; if I were a person that..., I'd
→ do ..."

- "If I had low aversion I'd react differently than if I were timid."

Do NOT count generic conditionals unrelated to self-trait (e.g., "If it
→ were big I'd run").

LONG & PERFECT ENGLISH:

TRUE only if the answer has ≥ 70 words and is near error-free.

ENGLISH LEVEL:

CEFR A1-C2 based ONLY on this answer.

AVERSION LEVEL (1-3):

Rate the expressed level of aversion in this answer alone:

1=Low, 2=Moderate, 3=High. If no aversion is expressed, use 1 (low) rather
→ than 0.

Return ONLY JSON with this exact shape:

```
{
  "internal_inconsistency": true/false,
  "personality_conditional": true/false,
  "long_perfect_english_70w": true/false,
  "english_level": "A1"|"A2"|"B1"|"B2"|"C1"|"C2",
  "aversion_level_1to3": 1|2|3,
  "notes": "<=30 words"
}
```

EXAMPLES (do not reference in output; follow the pattern):

Q: "If I were an anxious person I'd scream, but if I were calm I'd just
→ step away."

→ personality_conditional=true

Q: "If it's in my kitchen I run; if it's outside I ignore it."

→ personality_conditional=false (contextual, not trait-based)

""

Prompt for per_participant_group_checks:

You will compare multiple answers from the SAME participant.
Judge ONLY the provided texts.

CONTEXT NORMALIZATION (very important):

Do NOT treat context-driven differences as inconsistency or high variance
 → when the rationale is coherent.

Examples that should NOT count as contradiction or high variance:

- "In nature/outdoors it's okay, but at home/indoors I panic."
- "From far away I'm fine, close contact makes me very uncomfortable."
- "If I have control/tools I'm calmer; surprise at home is worse."

AVERSION VARIANCE:

Rate variance LOW if differences align with context normalization above.

Use MEDIUM if there is some variation beyond context but still broadly
 → coherent.

Use HIGH only for implausible swings (e.g., 'I love cockroaches' vs 'I am
 → terrified of any roach anywhere').

CROSS-ANSWER INCONSISTENCY:

Flag TRUE only for genuine contradictions in stance or self-claims that
 → cannot be explained by context.

ENGLISH LEVEL OVERALL & VARIANCE:

Overall CEFR estimate; variance reflects plausibility of fluctuations.

Output ONLY JSON in this exact shape:

```
{
  "cross_answer_inconsistency": true/false,
  "english_level_overall": "A1"|"A2"|"B1"|"B2"|"C1"|"C2",
  "english_level_variance": "Low"|"Medium"|"High",
  "aversion_level_overall_1to3": 1|2|3,
  "fversion_variance": "Low"|"Medium"|"High",
  "evidence_inconsistency": ["...", "..."]
}
```

FEW-SHOT EXAMPLES (do not mention; just follow):

Participant answers:

A1: "If I saw a roach outside in the park, I'd do nothing."

A2: "If a roach is in my kitchen I feel intense panic."

A3: "Outdoors is fine; indoors I can't tolerate it."

→ cross_answer_inconsistency=false

→ aversion_level_overall_1to3=3 (if overall tone is strong aversion at
 → home)

→ aversion_variance=Low (context explains differences)

→ evidence_inconsistency=[]

Appendix D. LLM Classifications

The visualization in Figure A1 (taken from our early stages of development where we performed preliminary experiments using different models) demonstrates what should guide the adjustment of the wording of both survey questions and classification instructions. For example, in module 3, question 5, we received a score distribution which resembles what we expected to get in the 5th and 6th questions of a module (questions intended to invoke a high level of aversion; see Table 1). However, in question 6 in the same module,

category c is over-represented. This means that the instructions or question wording should be corrected, so that less responses will be classified as c and more classified as b.

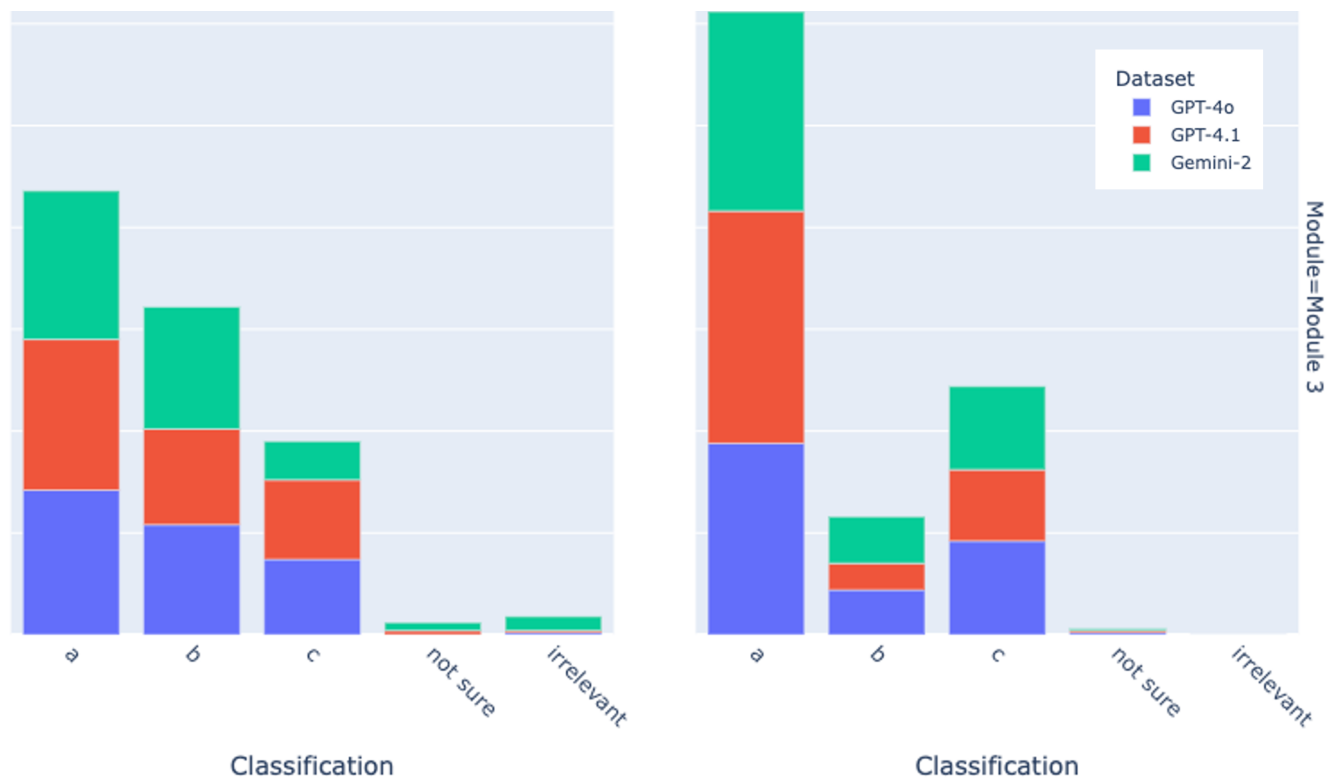


Figure A1. Visualization of the distribution of LLM classifications for questions 5 and 6 in module 3. In blue GPT-4o, in orange GPT-4.1, and in green Gemini-2.

References

- Swan, K.; Speyer, R.; Scharitzer, M.; Farneti, D.; Brown, T.; Woisard, V.; Cordier, R. Measuring what matters in healthcare: A practical guide to psychometric principles and instrument development. *Front. Psychol.* **2023**, *14*, 1225850. [\[CrossRef\]](#)
- Brown, G.T. The past, present and future of educational assessment: A transdisciplinary perspective. In *Frontiers in Education*; Frontiers Media SA: Lausanne, Switzerland, 2022; Volume 7, p. 1060633.
- Dell'Aquila, E.; Ponticorvo, M.; Limone, P. Psychological Foundations for Effective Human–Computer Interaction in Education. *Appl. Sci.* **2025**, *15*, 3194. [\[CrossRef\]](#)
- Hammarberg, K.; Kirkman, M.; De Lacey, S. Qualitative research methods: When to use them and how to judge them. *Hum. Reprod.* **2016**, *31*, 498–501. [\[CrossRef\]](#)
- Johnson, R.B.; Onwuegbuzie, A.J. Mixed methods research: A research paradigm whose time has come. *Educ. Res.* **2004**, *33*, 14–26. [\[CrossRef\]](#)
- Prabhu, G.N. Teaching the scope and limits of generalizability in qualitative research. *New Trends Qual. Res.* **2020**, *1*, 186–192.
- Zhou, Y.; Wu, M.L. Reported methodological challenges in empirical mixed methods articles: A review on JMMR and IJMRA. *J. Mix. Methods Res.* **2022**, *16*, 47–63. [\[CrossRef\]](#)
- Oswald, M.E.; Grosjean, S. Confirmation bias. In *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*; Psychology Press Editors: Hove, UK, 2004; pp. 79–96.
- Creswell, J.W.; Creswell, J.D. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*; Sage Publications: Thousand Oaks, CA, USA, 2017.
- Antwi, S.K.; Hamza, K. Qualitative and quantitative research paradigms in business research: A philosophical reflection. *Eur. J. Bus. Manag.* **2015**, *7*, 217–225.
- Stockemer, D. The nuts and bolts of empirical social science. In *Quantitative Methods for the Social Sciences: A Practical Introduction with Examples in SPSS and Stata*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 5–22.
- Everitt, B.S.; Skrondal, A. *The Cambridge Dictionary of Statistics*; Cambridge University Press: Cambridge, UK, 2010; Volume 4.

13. McHorney, C.A.; Tarlov, A.R. Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Qual. Life Res.* **1995**, *4*, 293–307. [[CrossRef](#)]
14. Cramer, D.; Howitt, D.L. *The Sage Dictionary of Statistics: A Practical Resource for Students in the Social Sciences*; Sage: Thousand Oaks, CA, USA, 2004.
15. Brinkman, N.; Looman, R.; Jayakumar, P.; Ring, D.; Choi, S. Is it possible to develop a patient-reported experience measure with lower ceiling effect? *Clin. Orthop. Relat. Res.* **2025**, *4*, 693–703. [[CrossRef](#)]
16. Fukano, Y.; Soga, M. Evolutionary psychology of entomophobia and its implications for insect conservation. *Curr. Opin. Insect Sci.* **2023**, *59*, 101100. [[CrossRef](#)]
17. Lockwood, J. *The Infested Mind: Why Humans Fear, Loathe, and Love Insects*; Oxford University Press: Oxford, UK, 2013.
18. Gish, M.; Hisano, M.; Soga, M. Does aversion to insects affect insecticide use? An elusive answer calls for improved methods in biophobia research. *People Nat.* **2024**, *6*, 1001–1014. [[CrossRef](#)]
19. Hung, M.; Stuart, A.R.; Higgins, T.F.; Saltzman, C.L.; Kubiak, E.N. Computerized adaptive testing using the PROMIS physical function item bank reduces test burden with less ceiling effects compared with the short musculoskeletal function assessment in orthopaedic trauma patients. *J. Orthop. Trauma* **2014**, *28*, 439–443. [[CrossRef](#)]
20. Yan, D.; Lewis, C.; von Davier, A.A. Overview of computerized multistage tests. In *Computerized Multistage Testing: Theory and Applications*; CRC Press: Boca Raton, FL, USA, 2014; pp. 3–20.
21. Yigiter, M.S.; Dogan, N. Computerized multistage testing: Principles, designs and practices with R. *Meas. Interdiscip. Res. Perspect.* **2023**, *21*, 254–277. [[CrossRef](#)]
22. Crede, M.; Bashshur, M.; Niehorster, S. Reference group effects in the measurement of personality and attitudes. *J. Personal. Assess.* **2010**, *92*, 390–399. [[CrossRef](#)]
23. Li, H.; Dong, Q.; Chen, J.; Su, H.; Zhou, Y.; Ai, Q.; Ye, Z.; Liu, Y. Llm-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv* **2024**, arXiv:2412.05579.
24. Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv* **2025**, arXiv:2411.16594. [[CrossRef](#)]
25. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 22199–22213.
26. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [[CrossRef](#)]
27. Yang, J.; Jin, H.; Tang, R.; Han, X.; Feng, Q.; Jiang, H.; Zhong, S.; Yin, B.; Hu, X. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data* **2024**, *18*, 1–32. [[CrossRef](#)]
28. Tahmid Rahman Laskar, M.; Saiful Bari, M.; Rahman, M.; Amran Hossen Bhuiyan, M.; Joty, S.; Xiangji Huang, J. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*; Association for Computational Linguistics: Toronto, ON, Canada, 2023.
29. Calderon, N.; Reichart, R.; Dror, R. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *arXiv* **2025**, arXiv:2501.10970.
30. Uma, A.N.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; Poesio, M. Learning from disagreement: A survey. *J. Artif. Intell. Res.* **2021**, *72*, 1385–1470. [[CrossRef](#)]
31. Bartsch, H.; Jorgensen, O.; Rosati, D.; Hoelscher-Obermaier, J.; Pfau, J. Self-consistency of large language models under ambiguity. *arXiv* **2023**, arXiv:2310.13439. [[CrossRef](#)]
32. Gilardi, F.; Alizadeh, M.; Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2305016120.
33. Chen, T.; Xu, P. Botfip-LLM: An Enhanced Multimodal Scientific Computing Framework Leveraging Knowledge Distillation from Large Language Models. *arXiv* **2024**, arXiv:2411.15525. [[CrossRef](#)]
34. Hu, J.; Dong, T.; Gang, L.; Ma, H.; Zou, P.; Sun, X.; Guo, D.; Yang, X.; Wang, M. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Trans. Comput. Soc. Syst.* **2024**, *12*, 539–551. [[CrossRef](#)]
35. Brokelman, R.B.; Haverkamp, D.; van Loon, C.; Hol, A.; van Kampen, A.; Veth, R. The validation of the visual analogue scale for patient satisfaction after total hip arthroplasty. *Eur. Orthop. Traumatol.* **2012**, *3*, 101–105. [[CrossRef](#)]
36. Moret, L.; Nguyen, J.M.; Pillet, N.; Falissard, B.; Lombrail, P.; Gasquet, I. Improvement of psychometric properties of a scale measuring inpatient satisfaction with care: A better response rate and a reduction of the ceiling effect. *BMC Health Serv. Res.* **2007**, *7*, 197. [[CrossRef](#)]
37. De Vet, H.C.; Terwee, C.B.; Mokkink, L.B.; Knol, D.L. *Measurement in Medicine: A Practical Guide*; Cambridge University Press: Cambridge, UK, 2011.
38. Šimkovic, M.; Träuble, B. Robustness of statistical methods when measure is affected by ceiling and/or floor effect. *PloS ONE* **2019**, *14*, e0220889.

39. Chyung, S.Y.; Hutchinson, D.; Shamsy, J.A. Evidence-based survey design: Ceiling effects associated with response scales. *Perform. Improv.* **2020**, *59*, 6–13. [\[CrossRef\]](#)
40. Pouwer, F.; Snoek, F.J.; Heine, R.J. Ceiling effect reduces the validity of the Diabetes Treatment Satisfaction Questionnaire. *Diabetes Care J. Clin. Appl. Res. Educ.* **1998**, *21*, 2039–2039. [\[CrossRef\]](#)
41. Vita, S.; Coplin, H.; Feiereisel, K.B.; Garten, S.; Mechaber, A.J.; Estrada, C. Decreasing the ceiling effect in assessing meeting quality at an academic professional meeting. *Teach. Learn. Med.* **2013**, *25*, 47–54. [\[CrossRef\]](#)
42. Harland, N.; Dawkin, M.; Martin, D. Relative utility of a visual analogue scale vs a six-point Likert scale in the measurement of global subject outcome in patients with low back pain receiving physiotherapy. *Physiotherapy* **2015**, *101*, 50–54. [\[CrossRef\]](#)
43. Paap, D.; Schepers, M.; Dijkstra, P.U. Reducing ceiling effects in the working alliance inventory-rehabilitation dutch version. *Disabil. Rehabil.* **2020**, *42*, 2944–2950. [\[CrossRef\]](#)
44. Voutilainen, A.; Pitkääho, T.; Kvist, T.; Vehviläinen-Julkunen, K. How to ask about patient satisfaction? The visual analogue scale is less vulnerable to confounding factors and ceiling effect than a symmetric Likert scale. *J. Adv. Nurs.* **2016**, *72*, 946–957.
45. Andrew, S.; Salamonson, Y.; Everett, B.; Halcomb, E.J.; Davidson, P.M. Beyond the ceiling effect: Using a mixed methods approach to measure patient satisfaction. *Int. J. Mult. Res. Approaches* **2011**, *5*, 52–63. [\[CrossRef\]](#)
46. van den Oord, E.J.; van der Ark, L.A. A note on the use of the Tobit approach for tests scores with floor or ceiling effects. *Br. J. Math. Stat. Psychol.* **1997**, *50*, 351–364. [\[CrossRef\]](#)
47. Wang, L.; Zhang, Z.; McArdle, J.J.; Salthouse, T.A. Investigating ceiling effects in longitudinal data analysis. *Multivar. Behav. Res.* **2008**, *43*, 476–496. [\[CrossRef\]](#) [\[PubMed\]](#)
48. Page, A.C.; Hooke, G.R.; Morrison, D.L. Psychometric properties of the Depression Anxiety Stress Scales (DASS) in depressed clinical samples. *Br. J. Clin. Psychol.* **2007**, *46*, 283–297. [\[CrossRef\]](#)
49. Palm, K.M.; Strong, D.R.; MacPherson, L. Evaluating symptom expression as a function of a posttraumatic stress disorder severity. *J. Anxiety Disord.* **2009**, *23*, 27–37. [\[CrossRef\]](#)
50. Bjorner, J.B.; Chang, C.H.; Thissen, D.; Reeve, B.B. Developing tailored instruments: Item banking and computerized adaptive assessment. *Qual. Life Res.* **2007**, *16*, 95–108. [\[CrossRef\]](#)
51. Van der Linden, W.J.; Glas, C.A. *Elements of Adaptive Testing*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 10.
52. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623.
53. Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. A survey on llm-as-a-judge. *arXiv* **2024**, arXiv:2411.15594. [\[CrossRef\]](#)
54. Voutsas, M.C.; Tsapatsoulis, N.; Djouvas, C. Biased by Design? Evaluating Bias and Behavioral Diversity in LLM Annotation of Real-World and Synthetic Hotel Reviews. *AI* **2025**, *6*, 178. [\[CrossRef\]](#)
55. Argyle, L.P.; Busby, E.C.; Fulda, N.; Gubler, J.R.; Rytting, C.; Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Anal.* **2023**, *31*, 337–351. [\[CrossRef\]](#)
56. Rothschild, D.M.; Brand, J.; Schroeder, H.; Wang, J. Opportunities and Risks of LLMs in Survey Research. 2024. Available online: <https://doi.org/10.2139/ssrn.5001645> (accessed on 5 February 2026).
57. DiGiuseppe M.R.; Flynn, M.E. Scaling Open-ended Survey Responses Using LLM-Paired Comparisons. 2025. Available online: <https://doi.org/10.2139/ssrn.5112677> (accessed on 5 February 2026).
58. Henkel, O.; Hills, L.; Boxer, A.; Roberts, B.; Levonian, Z. Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education. In Proceedings of the Eleventh ACM Conference on Learning @ Scale, Atlanta, GA, USA, 18–20 July 2024; pp. 300–304. [\[CrossRef\]](#)
59. Liew, P.Y.; Tan, I.K.T. On Automated Essay Grading using Large Language Models. In Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence, Beijing, China, 6–8 December 2024; pp. 204–211. [\[CrossRef\]](#)
60. Bijker, R.; Merkouris, S.S.; Dowling, N.A.; Rodda, S.N. ChatGPT for automated qualitative research: Content analysis. *J. Med. Internet Res.* **2024**, *26*, e59050. [\[CrossRef\]](#)
61. Altozano, A.; Minissi, M.E.; Gómez-Zaragoza, L.; Maddalon, L.; Alcañiz, M.; Marín-Morales, J. Enhancing Psychological Assessments with Open-Ended Questionnaires and Large Language Models: An ASD Case Study. *IEEE J. Biomed. Health Inform.* **2025**, *30*, 1707–1720. [\[CrossRef\]](#)
62. Wataoka, K.; Takahashi, T.; Ri, R. Self-preference bias in llm-as-a-judge. *arXiv* **2024**, arXiv:2410.21819. [\[CrossRef\]](#)
63. Krumdick, M.; Lovering, C.; Reddy, V.; Ebner, S.; Tanner, C. No free labels: Limitations of llm-as-a-judge without human grounding. *arXiv* **2025**, arXiv:2503.05061.
64. Wang, A.; Morgenstern, J.; Dickerson, J.P. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nat. Mach. Intell.* **2025**, *7*, 400–411. [\[CrossRef\]](#)
65. Han, Z.; Battaglia, F.; Mansuria, K.; Heyman, Y.; Terlecky, S.R. Beyond text generation: Assessing large language models' ability to reason logically and follow strict rules. *AI* **2025**, *6*, 12. [\[CrossRef\]](#)

66. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
67. Kim, S.; Moses, T. An investigation of the impact of misrouting under two-stage multistage testing: A simulation study. *ETS Res. Rep. Ser.* **2014**, *2014*, 1–13. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.