RESEARCH-ARTICLE

From Narrative to Numbers: Evaluating Survey Questionnaires with Large Language Models

AKASHDEEP CHAKRABORTY, University of Central Florida, Orlando, FL, United States

JOSEPH J. LAVIOLA JR., University of Central Florida, Orlando, FL, United States

Published: 22 March 2026

IUI '26: 31st International Conference on Intelligent User Interfaces
March 23 - 26, 2026
Paphos, Cyprus

Conference Sponsors:
SIGCHI
SIGAI

IUI '26: Proceedings of the 31st International Conference on Intelligent User Interfaces (March 2026)
https://doi.org/10.1145/3742413.3789071
ISBN: 9798400719844

From Narrative to Numbers: Evaluating Survey Questionnaires with Large Language Models

Akashdeep Chakraborty
Interactive Systems and User Experience Lab
University of Central Florida
Orlando, Florida, USA
ak958558@ucf.edu

Joseph J. LaViola Jr.
Computer Science
University of Central Florida
Orlando, Florida, USA
jjl@cs.ucf.edu

Abstract

As the field of intelligent interfaces is evolving, there is also a growing need for feedback mechanisms that are both expressive for participants and also contain reliable and useful information for the researcher conducting the study for survey data collection. Our study involves exploring two different kinds of survey methods: a standardized slider scale for web-based surveys and a free-form text input with a Large Language Model (LLM) acting as a backbone. The experiment includes 36 participants completing a 4×4 sliding-tile game at two different levels (easy and hard) with difficulty standardized via Manhattan-distance targets. The response mode order was counterbalanced across two sequences. This task aimed to evaluate the accuracy and quality of participant responses through different survey methodologies. Our key findings are that the LLM survey results are equivalent to the ones reported through the Web-Based slider scale questionnaire method. Our contribution is an intelligent framework that allows text-based reflections within an adaptive survey interface, helping both participants to express their experiences naturally and also researchers to gain valuable information about their system.

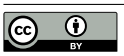
CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **Natural language interfaces**; **User studies**; **Interaction design**; **User interface design**; • **Computing methodologies** → **Natural language processing**.

Keywords

User experiments and studies, LLMs - End-user interaction with LLMs, agents, and multimodal models (e.g., chatbots, image generation), LLMs - Prompt engineeering

ACM Reference Format:
Akashdeep Chakraborty and Joseph J. LaViola Jr.. 2026. From Narrative to Numbers: Evaluating Survey Questionnaires with Large Language Models. In *31st International Conference on Intelligent User Interfaces (IUI '26), March 23–26, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3742413.3789071



IUI '26, Paphos, Cyprus

ACM ISBN 979-8-4007-1984-4/26/03
https://doi.org/10.1145/3742413.3789071

1 Introduction

Traditional survey methodologies with a fixed Likert scale often miss out the context or reasoning [30]. For example, while participating in a study a participant may feel that their experience was not smooth and selects a point on the Likert scale. With that single point, we can say it was challenging for the participant, but we often overlook the reasons behind the difficulty. Traditional Likert and slider scales offer efficiency but introduce response biases such as central tendency and straight-lining effects [11]. A rating on the mental demand dimension lets us assume a "5" means someone had to focus a bit more than average to complete the task, but for someone else, it can be due to other distractions, which can be the reason for the higher mental demand. These nuances can be essential for researchers and can improve understanding of their systems [27]. Yet researchers persist with numbers because they need quantifiable, comparable data. The core gap we address is preserving nuances to understand the participants' perspective and the Likert values for further analytical reasoning. We propose an intelligent system that captures both.

Although the richest context typically comes from open-ended interviews they are costly and very difficult to scale for larger samples. Free-form text responses, promises richer context and more expressive feedback than Likert scales; however, they have drawbacks. Manual coding of numerous comments can take months of labor [3]. Rule-based classifiers can categorize themes, but they still reduce text to coarse labels rather than numeric scores. This indicates that, so far, we can capture rich insights; however, we still lack an efficient method to convert these insights into the quantitative data required for analysis.

Recent rapid developments in Large language models (LLMs) suggest that they are a promising path forward. Fagbohun et al. [5] showed that LLMs can now grade essays and short answers with high consistency, along with personalized feedback. Lee et al. developed a system integrated with LLM, known as ChatFive, which replaced Big-Five personality inventories with conversational prompts mapped to numerical trait ratings, achieving strong agreement with standard questionnaires [17]. These findings suggest that, if LLMs can reliably convert essays and short answers to Likert scale scores, they might also be able to convert workload descriptions to TLX ratings.

In this paper, we use NASA-TLX, the most widely used instrument for measuring six dimensions (mental, physical, temporal, performance, effort and frustration), each scored on a 1-7 scale [9]. For this survey questionnaire, various digital implementations have been tried, but they all rely on the same method: using sliders or numeric entry. We introduce a two-stage pipeline to explore this

concept. First, a validator filters out off-topic or too-short answers. Then an LLM rater, processes each validated response, with their questions, such that the LLM has more clarity while assigning a number between 1 and 7 for one dimension at a time. We fix model settings and prompts to increase repeatability and constrain outputs to grading. This LLM assisted UI, lets participants express their feelings freely without particularly picking a number and gives the researchers' enough nuances to understand participants real feedback along with TLX-compatible numeric data for analysis.

We will compare these text-derived scores to both traditional slider responses and independent human coders, applying Intraclass Correlation Coefficients ICC(2,k) [8], weighted kappa, and TOST equivalence testing [26] to assess agreement and practical equivalence. Our contributions are as follows:

- An intelligent system assisted by an LLM for text-to-NASA-TLX conversion, preserving scale fidelity without sliders.
- A rigorous validation framework, including human and slider baselines, multi-model comparison, and equivalence tests.
- An analysis of the *nuances* that text reveals—cross-dimensional interactions, contextual factors, reasoning patterns, and confidence indicators—that sliders alone cannot capture.

2 Related Works

2.1 Survey Response Methods

Surveys are fundamental for data collection in various fields, including social sciences [29], psychology [24], and market research [16]. Originating in the late 19th century, foundational work such as Galton's studies on statistical inquiries [7] shaped survey methodology. These initial methods, which integrated new statistical techniques, influenced the development of contemporary survey designs and analytical approaches [22], enabling more robust data-driven inferences.

With the widespread adoption of the internet, digital surveys, particularly web-based ones, have become increasingly prevalent. This transition to digital data collection led researchers to examine how the survey medium affects responses. Miller et al [19] identified web-based surveys to offer advantages over paper-based surveys in terms of reduced administration time and cost. Carini et al [4] conducted a comparative study of web and paper surveys, finding small but significant mode effects on college students' responses, especially in technology-related contexts. Hayslett et al [10] further supported the use of web-based surveys. Prior research also explored the efficiency of web-based surveys in terms of cost and time across various group sizes, indicating web-based surveys more efficient than paper-based surveys [31]. In HCI, the shift toward web-based surveys after the 2000s led researchers to adopt similar survey administration techniques, focusing on design guidelines from question formatting to proper administering methods [23]. Since then, many researchers have adopted web-based surveys for their user studies [15]. Whether using web-based survey methods with a Likert rating scale, the issue of differentiation on rating scales persists, as concluded by Heerwegh et al [11], who found that people using web surveys were less likely to give varied responses compared to those in face-to-face surveys. Previously, researchers also tried to implement conversational chatbots to replace web surveys to compare response quality. One such work by Kim et al [14] concluded that they got higher response quality data, but they also reported it as less satisfying. Similar work by Xiao et al [33] using a pre-existing chatbot, which compared the web-based survey method with their method, which yielded clearer free-text responses compared to the traditional method. Although conventional chatbots can be beneficial, they have drawbacks such as requiring typed answers while going back and forth, resulting in shorter answers and unreliability, as concluded by Zarouali et al [36].

2.2 Automated Analysis of Open-Ended Survey Responses

Analyzing open-ended responses can be challenging. Traditional methods can fall short. More recently, advances in natural language processing (NLP) have enabled automated coding of textual survey responses. Mellon et al [18] used GPT-3 to automatically categorize open-ended "most important issue" survey responses into predefined thematic categories. They achieved 93.9 percent accuracy in agreement with a human coder categorizing political survey responses. More importantly, their approach focused on categorical categorizations rather than numerical rating conversion, which represents a key distinction from quantitative assessment tasks. Building on similar principles, researchers at IFPRI used LLMs to code occupation entries across multiple languages in developing countries [28]. Their system achieved approximately 85 percent accuracy, showcasing the capabilities of current LLMs for processing survey data, including cross-linguistic data. Following that recent advancement of LLMs opens new possibilities for assessment across different domains. LLMs were used more recently in automatic grading across different educational contexts. Farzi et al [6] used LLMs and developed a question bank retrieval system to evaluate and grade students' exams based on relevant information. Fagbohun et al [5] work highlights the use case of large language models in grading, offering consistent, scalable, and personalized LLMs. LLMs for grading short-answer questions in the field of physics by Yan et al [35] showed that LLMs provide consistent and accurate grading for simpler questions, which shows that LLMs can focus on a specific task at hand without deviating from the topic. The introduction of large language models into academic settings has proven to be useful due to their exceptional reasoning capabilities and understanding of context based on prior knowledge, as demonstrated by Schneider et al [25]. Lee et al [17] system showed that LLMs is a very good alternative to traditional Big-Five personality Likert scales with a conversational chatbot interface. In their study, participants engaged in natural dialogue about their personality traits, and the system mapped these conversational exchanges to numerical personality scores. They found that ChatFive performance was on par with established personality inventories, while participants reported higher engagement and satisfaction compared to traditional questionnaire formats. Although their approach required multi-turn dialogue rather than single-response text analysis, this demonstrated that an LLM can be reliable enough for psychological assessments. Xie et al [34] multi-agentic grading

system significantly improves the accuracy, consistency, and fairness of automated short answer grading by copying human grading processes. This was done by dividing the task into stages—rubric generation, grading, and review—the system demonstrated notable performance enhancements on multiple datasets.

Some of the other notable work which has involved LLMs is by Kaur et. al [13], where they discuss the pros and cons of using LLMs as synthetic personas to substitute human survey respondents. Similarly, Kaiser et al. [12] show that LLM-based digital twins achieve above-chance agreement with human survey data. Their results shows that LLMs can capture meaningful data similar to humans.

To our knowledge, no prior study has developed and validated an end-to-end pipeline that automatically validates on-task free-text survey responses and then maps single-turn text reflections directly into the numeric 1–7 ratings of a workload instrument (NASA-TLX). Our work is distinguishable from other work, specifically personality dialogue mapping [17], which performs conversational assessment to report Big Five scores, whereas we validate an end-to-end pipeline that filters on-task free-text workload reflections in real-time and then maps it directly to NASA-TLX 1-7 ratings. We further compare LLM-derived ratings against both slider baselines and independent human graders using ICC(2,k) and weighted kappa, providing the first evidence that text-based workload assessment can substitute for conventional numerical scales.

3 LLM Setup

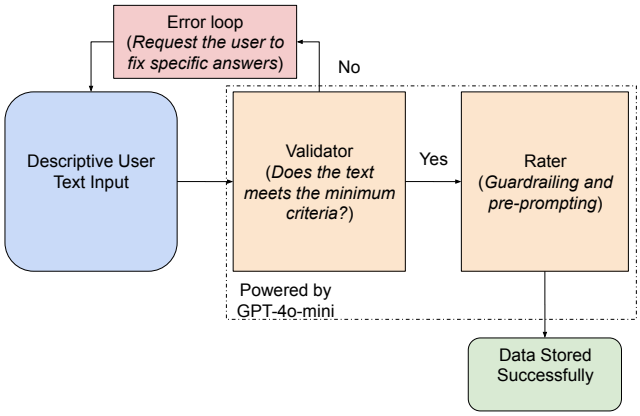


Figure 1: End-to-end text-to-LLM pipeline with error-reprompt loop. Input is sent to the GPT-4o-mini validator for validation, and then to the rater after passing the validation for final data storage. Failures from the validator are sent back to the participant to revise their particular input.

When participants finish providing free-text responses, each NASA-TLX dimension undergoes a two-stage automated processing pipeline (see Fig. 1). It first goes through a validator which judges, whether a response is informative and related to the task. A rater then maps a validated response to numerical scales on the conventional 1-7 range.

Validation and rating steps both use OpenAI's gpt-4o-mini model [20], which is conservatively decoded to ensure repeatability with a temperature of zero and a top_p level of 1.0. The response format is in JSON. Every API call has a 15-second timeout with two automatic retries in the event of a transient error. The model has no memory of participant data beyond the single prompt; however, we maintain a log of model ID and timestamp per processed submission for future reference. To safeguard participants' privacy, the pipeline sends only three inputs: the dimension name, the canonical version of the NASA-TLX question text, and the written response entered by a participant. We did not include any identifying information, such as difficulty levels, round numbers, or participant IDs.

The validator acts as gate to free-text responses that applies an explicit rubric. A response passes when it ensures that the participant's response is adequate, regardless of the difficulty level and is natural language that clearly refers to the participant's experience with sliding puzzle task only. Responses must contain a minimum of 8 words and avoid apparent gibberish, random content, or off-topic material. The validator categorizes accepted responses into quality tiers: high-quality responses contain clear, multi-sentence reflections of about 15 words or more; medium-quality responses are specific but brief, at 10-15 words; and low-quality responses meet minimum acceptability standards, at 8-10 words. The validator returns a JSON object containing pass/fail status, brief reasoning, and quality assessment. Failed responses trigger client-side prompts for revision and resubmission.

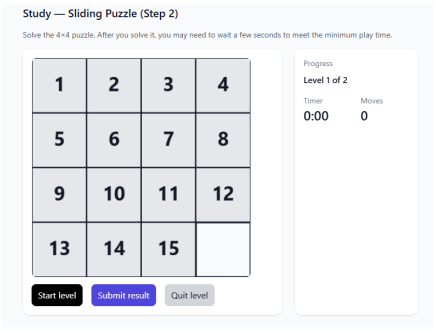
For validated responses, the rater maps text to numerical scores using explicit anchors where 1 represents "very low" and 7 represents "very high" across all dimensions. However, to prevent polarity confusion, the Performance dimension uses inverted anchoring where 1 indicates "very high success" and 7 indicates "very low success." The rater bases scores exclusively on participant responses while paying careful attention to negations and linguistic nuances. Each rating call returns a JSON object containing the numerical score and a brief explanation retained for auditability.

To make sure that the model does not deviate from the main task, we use fixed prompts and decoding parameters (temperature= 0) and enforce JSON-constrained output. When an API failure occurs due to a network issue, the front-end displays relevant error messages. Although an offline heuristic fallback is present in the codebase, it is not used when collecting data. To perform analyses that require harmonized polarity across rating scales, we employ a pre-specified recoding step that adjusts both slider and text-based scales and then calculates agreement metrics or calibration statistics. All the participants data for both the methods where saved in CSV (Comma Separated Value) format with dedicated folders for each participant.

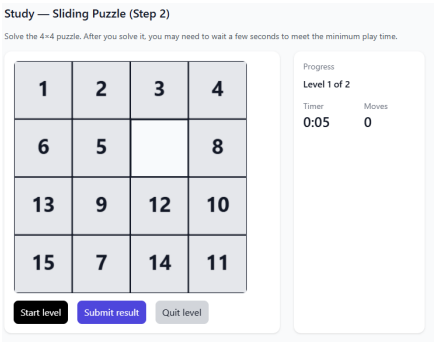
4 Experiment

4.1 Hypotheses

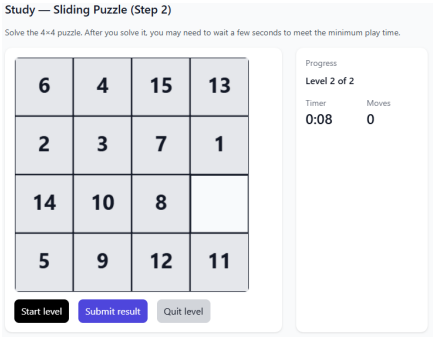
We conducted a comprehensive study involving 37 participants to understand the viability and effectiveness of our LLM-based text survey input method in its ability to convert textual input into Likert scale data. The study incorporated two different survey methodologies: (1) *Likert scale-based slider scale* and (2) *an LLM-based text input*. This approach enabled us to gather diverse perspectives and



(a) UI with buttons and solved puzzle example.



(b) Level 1 puzzle layout.



(c) Level 2 puzzle layout.

Figure 2: Sliding Puzzle interface and tasks. The UI (a) shows the main controls and solved state where the *Start* button begins the puzzle, the *Submit* button checks completion, and the *Quit level* button allows quitting after minimum time, while (b) and (c) illustrate the unsolved Level 1 and Level 2 puzzles.

insights, ensuring a robust analysis of the research hypotheses. We propose the following hypotheses:

- **H1:** When compared to the slider scale rating, the LLM will yield a similar Likert rating.
- **H2:** When it comes to describing workload, participants will prefer the text input.
- **H3:** The participants will prefer the text when it comes to capturing contextual nuances.

4.2 Participants

We recruited 37 participants from our university (22 male, 15 female). The ages of our participants ranged from 21 to 30 (M_{age} = 21.92, SD_{age} = 2.68). We also asked the participants about their puzzle experience on a 3-point scale (*1 = none, 2 = some, 3 = a lot*), and the mean and standard deviation were ($M_{\text{experience}}$ = 1.97, $SD_{\text{experience}}$ = 0.55). The distribution of puzzle experience was skewed toward some experience. We targeted a statistical power of 0.80 to detect medium effect sizes (Cohen's d = 0.5).

4.3 Apparatus

Our study application was delivered as a desktop, browser-based puzzle system. The experiment ran on a laptop equipped with an Intel Core i9 (12th generation) CPU, Nvidia GeForce RTX 3090 (8 GB), and 32 GB RAM, providing ample compute for local rendering and data collection. All large-language-model components were accessed via hosted APIs; no on-device LLM inference was performed. Only the study client is executed locally, with interaction logging handled by the application and model requests sent over the network to the respective API services.

4.4 Task Design

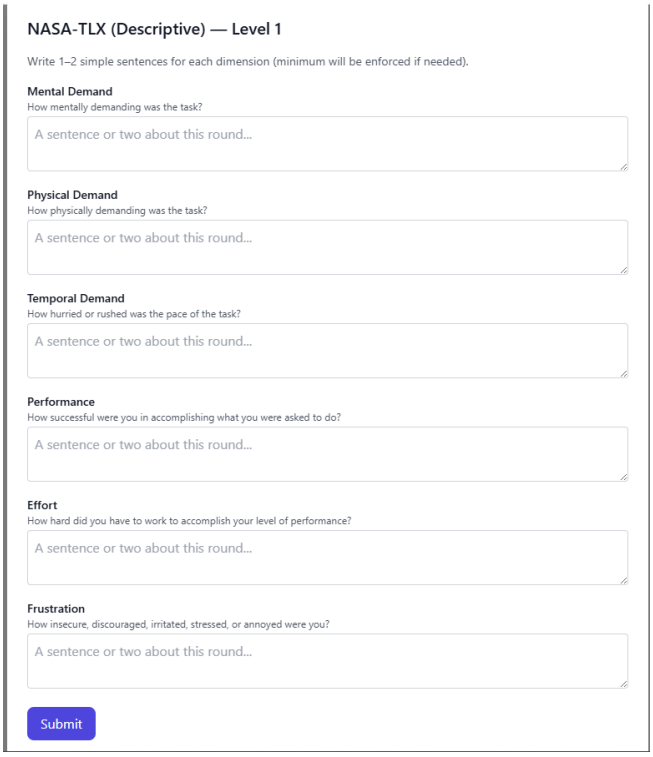
The task was carried out in a normal desktop environment. The participants completed randomized rounds of a 4x4 sliding-tile puzzle under two difficulty levels (easy, hard) (see Fig. 2). To ensure sufficient engagement, a minimum play time was put into condition, which is twenty seconds for the easy level and thirty seconds for the hard level. To distinguish between easy and hard levels, the difficulty was calibrated by Manhattan-distance targets. The study was web-hosted by a Java application served via Uvicorn backend displaying the tile grid, a timer, and a move counter (see Fig. 2a). Participants used three buttons:

- **Start Level:** Begins timer and move counter and randomizes the tiles.
- **Submit Result:** Records completion time, move count, and logs "solved" status.
- **Quit Level:** Enabled only after the minimum time; Allows early exit under frustration; logs partial progress.

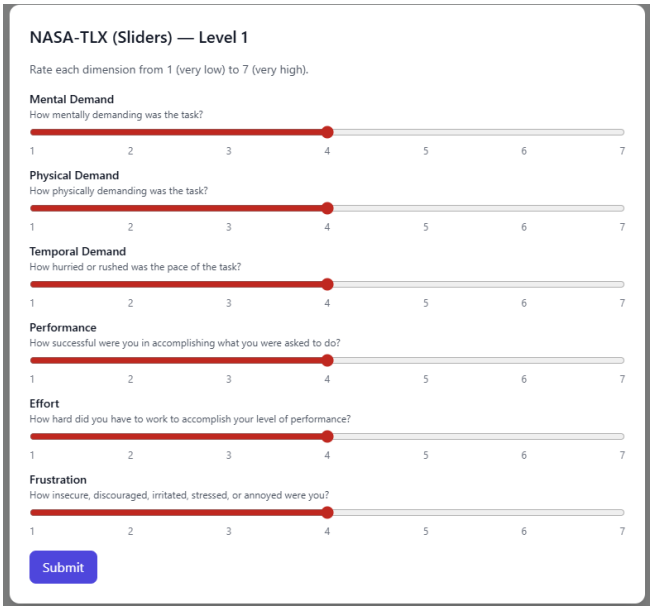
For each trial, participants clicked the start level, manipulated the tiles until they believed it was solved and checked by clicking the submit button. If it is solved, it will move forward to the survey mode; else it will pop up with a gentle message stating it is not finished. The survey input methods, that is, the slider and text (see Fig. 3), were counterbalanced across two sequences to control for order effects. The backend captured completion time, number of moves and timestamps along with the text and slider ratings. The text responses were forwarded via API to GPT-4o-mini for automated NASA-TLX conversion. All logs were saved in csv format for subsequent analysis.

4.5 Study Design

To understand how the text-based as a survey input stands when it comes to in comparison with traditional slider ratings, we employed a within-subjects study design. It was counterbalanced to compare automated text-to-NASA-TLX conversion against traditional slider



(a) NASA-TLX (text-based). Participants provide 1–2 sentences for each demand dimension.



(b) NASA-TLX (Sliders). Participants rate each demand dimension on a 7-point scale.

Figure 3: Two modes of NASA-TLX input. The text-based format (a) captures qualitative reflections, which are then processed by the LLM, while the slider format (b) captures quantitative ratings on a 7-point scale.

ratings. Thirty-six participants each completed two puzzle trials on a 4×4 grid, one "easy" trial (Manhattan distance 8–16) and one "hard" trial (Manhattan distance 56–76). After each trial, participants provided survey responses via two methods, that is slider ratings and text. Free text responses were forwarded via API to GPT-4o-mini for conversion into NASA-TLX scores. Participants were assigned to two sequences in alternating order to control order effects. Sequence A (easy followed by slider and then text; hard followed by text and then slider) or Sequence B (easy followed by text and then slider; hard followed by slider and then text). The primary analysis we did is two one-sided tests (TOST) to assess equivalence between LLM-converted and slider-based within predefined bounds. Also, as secondary analyses, we did linear and rank-order associations through Pearson correlations and Spearman's ρ, along with evaluating reliability using Intraclass Correlation Coefficients (ICC) and quadratic-weighted Cohen's κ.

4.6 Procedure

Upon arrival, recruited participants were asked to review an informed consent document and then followed by a demographics section digitally. Once the initial paperwork and survey were completed, the researcher explained how to use the system and what to expect. The researcher also elaborated on what kind of survey input methods they will be using, which are a traditional slider method and a text box method, where they can write down what they experienced in that round. Afterwards, once the participant understood everything and was comfortable starting the study, the researcher asked them to take a seat in front of the setup and begin their study. The main task was divided into two levels, while the survey input methods were administered using a Latin Square design to avoid bias. The participant went through the easy level and then the hard level. After each level, they were asked either to use the slider first or the text box first, depending on the sequence A or B. When the participants completed both easy and hard levels along with their surveys, a post-study survey was also administered. The time required to complete the study was approximately 30 minutes. Participants were each compensated with 10 dollars in Amazon gift cards.

5 Results

In this section we discuss the outcomes of the study and see how the LLM performed against the traditional Likert scale.

5.1 Dataset and exclusions

We recruited 37 participants. We excluded one participant data due to straightlining (P36) that is providing uniform ratings ("4" across all dimensions)($N = 36$). With 6 NASA–TLX dimensions and 2 difficulty conditions (Easy/Hard) per participant, this yields $36 \times 6 \times 2 = 432$ text-based items; all subsequent analyses use these $N = 432$ items.

5.2 NASA-TLX

We tested to compare whether *Hard* increased workload relative to *Easy* within subjects using one-sided Wilcoxon signed-rank tests (Hard > Easy). We ran it separately for the traditional Likert scale sliders and for the LLM generated ratings, with Holm correction

Table 1: Within-subject manipulation check (Hard > Easy) using Wilcoxon signed-rank tests with Holm correction across dimensions. Per-dimension paired $n = 36$.

	Slider		LLM	
Dim.	Med. Δ	p_{Holm}	Med. Δ	p_{Holm}
Effort	**+1**	**0.0039**	**+1**	**0.0031**

Table 2: TOST between LLM-derived Likert ratings and slider ratings ($\Delta = \pm 0.5$ Likert categories; 90% CI). Easy and Hard each use $n = 216$ paired items.

Condition	Mean diff (LLM–Slider)	90% CI	Equivalence
Easy	0.032	[−0.027, 0.092]	Equivalent ($p < .001$)
Hard	0.042	[−0.031, 0.115]	Equivalent ($p < .001$)

across the six dimensions. Only *Effort* was more under *Hard* in both measures; all other dimensions were non-significant after correction (see Table 1). We also ran an equivalence test between LLM text-derived ratings and slider ratings using two one-sided tests (TOST) with equivalence bounds $\Delta = \pm 0.5$ Likert categories and 90% CIs. LLM and slider were statistically equivalent in both difficulty conditions (Table 2). According to the bar plots (see Fig. 4) LLM and slider ratings closely agree across all NASA-TLX dimensions in both easy and hard as well.

5.3 Human Inter-Rater Reliability

Two human raters independently mapped free-text notes to 1–7 NASA–TLX overall workload scores using a locked rubric given to them. The rubric consisted of 1–2 indicating very low demand and highly successful, 3–4 indicating mild to moderate demand and successful with noticeable effort, 5–6 reflects high to very high demand with marginal success and 7 denotes extreme demand and unsuccessful at solving the puzzle. Before scoring the original dataset, they completed a brief calibration on data collected during pilot study which is separate from the original study data and then proceeded independently, blinded to each others' scores, to all LLM outputs and as well as the slider values.

Inter-rater reliability was substantial on the full dataset ($N = 432$) where Cohen's κ_{qw} (H1–H2) = 0.789, 95% CI [0.737, 0.828]; ICC(2,1) = 0.789, 95% CI [0.739, 0.830] (two-way random, absolute agreement, single-measure); and the corresponding average-measure ICC(2,2) = $\frac{2\times 0.789}{1+0.789} \approx 0.882$, 95% CI [0.850, 0.907]. Given the agreement, we used the mean of the two raters ($H12_{\text{avg}}$) as the human consensus reference for further comparison with the LLM generated grades (see Table 3).

5.4 Model Agreement to Human Consensus

We compared the LLM's 1–7 predictions to the human consensus on the same $N = 432$ items (see Table 4), and we found that discrete scale alignment was high across all the models. For the model that was used for the user study *GPT-4o-mini*, we found that the Cohen's κ_{qw} ($H12_{\text{avg\_round}}$ vs LLM) = 0.759, 95% CI [0.705, 0.804]. The gap to human–human agreement is $\Delta\kappa = 0.789 - 0.759 = 0.030$; overlapping CIs indicate closeness to human grade on the discrete scale.

Table 3: Human inter-rater reliability on 1–7 TLX overall workload from free text ($N = 432$). 95% CIs via bootstrap for ICC; κ CI as reported.

Metric	Estimate	95% CI
Cohen's κ_{qw} (H1–H2)	0.789	[0.737, 0.828]
ICC(2,1)	0.789	[0.739, 0.830]
ICC(2,2)	0.882	[0.850, 0.907]

Table 4: LLM agreement with human consensus on the 1–7 scale ($N = 432$). 95% CIs via nonparametric bootstrap except κ.

Model	n	κ_{qw}	Spearman ρ	MAE (1–7)
gpt-4o-mini[20]	432	0.759 [0.705, 0.804]	0.837 [0.791, 0.872]	0.825 [0.746, 0.909]
gpt-4.1-mini[21]	422	0.854 [0.821, 0.881]	0.891 [0.862, 0.913]	0.622 [0.560, 0.698]
claude-sonnet-4-20250514[2]	432	0.847 [0.817, 0.872]	0.893 [0.869, 0.911]	0.638 [0.576, 0.701]
ministral-8b-latest[1]	427	0.848 [0.814, 0.876]	0.893 [0.867, 0.914]	0.639 [0.571, 0.707]
grok-3-mini[32]	406	0.849 [0.816, 0.878]	0.895 [0.872, 0.915]	0.658 [0.589, 0.732]

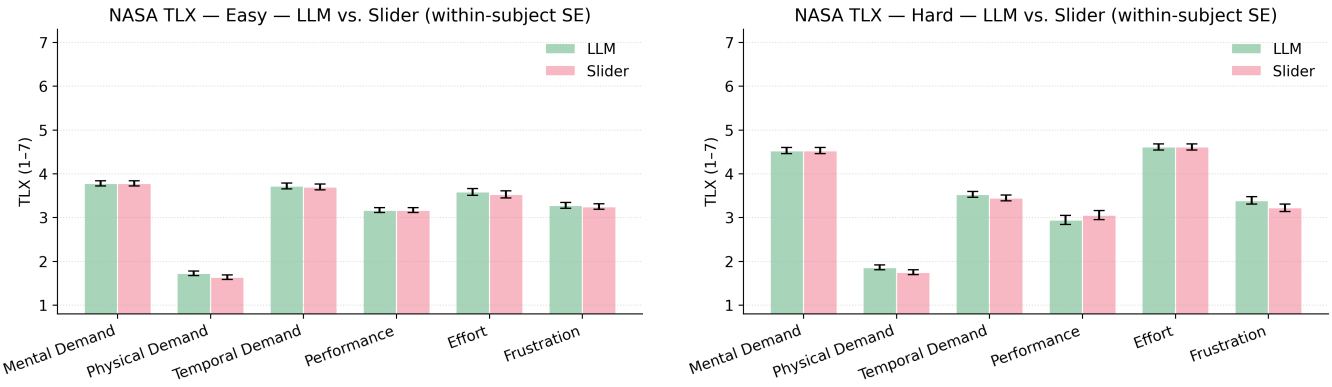
To test the rank-order consistency, we ran Spearman correlations between the LLMs and the human ratings which ranged from $\rho = 0.837$ to $\rho = 0.895$, indicating that the model largely preserved the relative ordering of items. The MAE also suggests that the deviations were modest ranging between 0.62 and 0.83, which corresponds to average disagreements of less than one rating. Together, this indicates that our method is generalizable across other LLMs which can produce ratings that are close to human consensus in absolute terms.

5.5 Benchmarking model

To see how other models performed with the same data collected through the user study, we conducted a benchmarking of different models (see Table 5). This benchmarking was performed after the original study using *gpt-4o-mini* to test the viability and to determine if our method is generalizable across other models.

We used the same prompt structure, temperature = 0 and JSON-constrained decoding as we used for GPT-4o-mini. *Coverage* denotes the fraction of benchmark items that had a parsable, schema–valid JSON output. *Pass rate* denotes the proportion of those covered instances that satisfied all validation checks, and latencies are end-to-end for each item median 50th percentile and 95th percentile in seconds.

Out of the six models tested, only one model *claude-sonnet-4-20250514*—achieved 100% coverage along with our *gpt-4o-mini* which was the model initially used for the study. *ministral-8b-latest* was the fastest overall (p50 ≈ 0.40 s, p95 ≈ 0.60 s) while maintaining near–perfect coverage (99.1%). *gpt-4.1-mini* was also quick (p50 ≈



(a) Easy **(b) Hard**

Figure 4: NASA–TLX (1–7) means by dimension, LLM vs. slider. Error bars: within-subject SE (Cousineau–Morey).

0.77 s, p95 ≈ 1.19 s) with 97.7% coverage. *grok-3-mini* lagged substantially on latency (p50 ≈ 9.42 s, p95 ≈ 19.17 s) and reached 94.0% coverage. Although we ran the benchmarking for all the other models the initial model used in this case to collect the data which is *gpt-4o-mini* only cost 0.074 dollars for the whole user study to be conducted. We exclude *gemini-2.5-flash* and a 3B–class model due to incomplete / unreliable logging, to avoid skewing cross–model comparisons.

Including the model agreement to human consensus (see Section 5.4) along with model benchmarking in terms of overall operational metrics we can see that our method is not only close to human ratings but also generalizes reliably across a diverse set of LLMs.

5.6 Post Study

We also ran a small post-study survey asking participants about their thoughts and preferences over the two methods. Firstly, to understand the participants' naturalism towards giving feedback or filling, we asked the participant *Which method felt more natural to describe workload?* in which they had an option to choose whether it was the *slider* or the *text-based*. According to the results analyzed from 36 participants, we found that approximately 53% prefer the text over the slider scale. Secondly, they were asked which method of survey input helped in capturing the context/feelings better *Which captured nuances or context better?*. Also, for this, we had two options similar to the previous questions. To support our claims after analyzing the results, we got 97.2% preference for text-based over the slider scale.

The third *If descriptive answers are summarized into 1–7 later, how fair/accurate would that feel?* and the fourth *Which method did you prefer overall and why?* questions were descriptive input in nature. In these two questions, the participants shared what they felt about the respective questions. These narratives were extracted for further discussions on our hypotheses in the discussion section 6.

5.7 Thematic Analysis

We also analyzed the text given by each participants and constructed a reflexive thematic analysis from both the levels. This thematic coding was done to analyze and delve deeper into the nuances to understand why participants felt a certain way and to showcase that typically during only Likert scale surveying we tend to omit the valuable feedback. Wilcoxon signed-rank tests with holm correction (see section 5.2 only the *effort* dimension showed significant differences while the other TLX-dimensions are non-significant. This analysis indicates that level difficulty was not the one responsible for it but other factors like timer pressure, rushing made easy level seem harder while strategically approaching the hard level made it look easier than before. The coding was based on semantic and iterative, with themes developed on analytic patterns rather than count-based.

- **Pacing Paradox on 'Level 1'.** Level 1 was set to be on easy difficulty which led participants to believe that they would be able to finish it quickly with speed. This led to rushing increasing in numerous unwanted moves and near-misses making the level 1 seemingly hard than it is (Table 6, Theme A).
- **Strategy Stabilizes 'Level 2'.** Participants also stated that after Level 2 they went for more patterned based approaches which ultimately let them perceive Level 2 which was the harder level of the two seem easier (Table 6, Theme B).
- **Error–Repair Loop.** A lot of participants also shared that most of the error went to fixing or undo moves which they saw was making it more complicated than solving the issue (Table 6, Theme C).
- **Comparative Framing.** Although level 1 was easier than level 2 in terms of complexity, participants mentioned that level 2 was easier than level 1 (Table 6, Theme D).

6 Discussion

Our main goal was to understand can an LLM grade as closely as a human grade, along with participants' natural preferences for describing their experience during a user study.

6.1 Score Alignment between Human and LLM

Our first research hypotheses is that if the descriptive text survey method backed by an LLM will have similar results when compared

Table 5: Model performance on the 432 items (only models with complete logs). Latencies in seconds.

Model	*n*	Cov. (%)	Pass (%)	JSON fail (%)	p50 s	p95 s	Tokens in	Tokens out
gpt-4o-mini[20]	432	100.0	100.0	0.0	0.81	1.41	160 954	24 555
ministral-8b-latest[1]	428	99.1	99.1	0.0	0.40	0.60	164 094	22 588
gpt-4.1-mini[21]	422	97.7	97.7	0.0	0.77	1.19	162 411	23 000
claude-sonnet-4-20250514[2]	432	100.0	100.0	0.0	2.02	3.13	182 697	28 974
grok-3-mini[32]	406	94.0	94.0	0.0	9.42	19.17	157 565	22 329

Table 6: Reflexive thematic analysis: themes, central concepts, and illustrative extracts (numbers-only task).

Theme	**Central Organizing Concept (COC)**	**Illustrative Extracts (max 3)**
Theme A — The Pacing Paradox on *Level 1*	Expectations of finishing Level 1 quicker led to timer monitoring and rushing.	"I didn't notice the timers, but after a while I felt a more than slight rush to get it done." [Temporal, Level 1] "The task was not too mentally demanding, but the timer rushed me a bit making it a bit more mentally demanding." [Mental, Level 1] "The task felt very hurried and rushed. I believe this is because of the timer and move counter." [Temporal, Level 1]
Theme B — Strategy Stabilizes *Level 2*	Familiarity and strategizing based on previous attempts made Level 2 easier than intended compared to Level 1.	"Less than the previous puzzle, since by now I was more familiar with the puzzle game." [Mental, Level 2] "I completed the task faster than the previous puzzle which I believe was supposed to be easier." [Performance, Level 2] "This task was less demanding than the last one, since I am more familiar with the mechanics of this puzzle." [Mental, Level 2]
Theme C — The Error–Repair Loop	Stuck with unwanted moves and attempts to fix them led to a poorer experience.	"Frustrated at first since I made a few mistakes I couldn't undo, but less at the end." [Frustration, Level 2] "I got a little bit frustrated when I realized my mistake as I messed up my progress before trying to fix a smaller part of a bigger problem near the end." [Frustration, Level 1] "It felt easy at first, but then I made a big mistake which caused me to take a long time and a lot of mental effort to fix." [Mental, Level 1]
Theme D — Comparative Framing & Expectation–Outcome Mismatch	Prior attempts shaped what participants defined as hard.	"I feel like the second time was easier because I had already attempted before, so this was not that mentally demanding." [Mental, Level 2] "I felt like it was easier this time around. I felt like I worked the same, but more efficiently." [Effort, Level 2] "I completed the task faster than the previous puzzle which I believe was supposed to be easier." [Performance, Level 2]

to slider scales. The findings in section 5.2 and section 5.4 offer substantial support for H1. The results show that TOST had equivalence when compared to traditional likert scale rating. Also, when compared to human graders, yielding strong results in favour of H1. This leads us to believe that even an LLM which is not fine-tuned to the information it's receiving with proper prompt engineering and guard-railing, it is able to understand what needs to be done. In the post survey questions, we also asked the participants in Q3 *If descriptive answers are summarized into 1–7 later, how fair/accurate would that feel?*. A lot of participants said in favor of fairness, which is evident from the remarks made by P1 *"I would say that it would be fair as most descriptive answers were close to how I felt on the sliders."*, P26 said *"More accurate because when writing out my answers I understand what I would rate my feelings better than if I were to just rate it right away."* as well as P2 said *"Yes I think that my answers be converted into a 1-7 is a lot more convenient for me because it takes a lot more effort to think how I felt and how it would correlate to a number."*. Given all the evidences it shows that an LLM can also convert narratives to numbers while keeping the context rich texts for further understanding for the researchers conducting the user study, allowing them to make necessary changes towards better experience for the participants.

6.2 Natural Feedback Mechanism

Our second research hypotheses is that if participants will find the text survey input to be more natural when compared to slider scales. Analysis of our post survey question Q1 *Which method felt more natural to describe workload?* (see section 5.6) yielded in support of H2. It is also evident from participants' responses to Q4 *Which method did you prefer overall and why?* where P1 quoted *"I preferred descriptive, as I would think about my experience with the puzzle more."*, P22 said *"I preferred being able to speak my mind thoroughly through text since it conveys my experience from my own perspective."* as well as P24 said *"I preferred the descriptive text because I could explain and give more nuances to subjective questions that may not be able to be described with just a number."*. While some participants also preferred the slider scale because of it quickness and less thinking,

which is evident from the remarks made by P31 said *"I preferred the slider because it was quicker than typing out my answer."* and P11 said *"I prefer the slider because it was more quicker and less time consuming as I don't have to type out everything."*. Also, participant 15 mentioned *"Although I would need clarification on some of the sliders, I felt it was somewhat easier. I think it's also due to wanting to answer the questions more quickly."*. These indicate that participants favored the slider scale because of its quickness, which can lead to a decrease in the quality of survey input, resulting in losing its validity overall. Also, this can affect a researcher who needs quality over quantity. So, where quickness is necessary, using a slider scale will be a better option, but when it comes to proper understanding of the study itself and to aid the researcher to be able to extract and provide more insights, along with incorporating participant feedback in future work, descriptive text should be preferred over sliders.

6.3 Contextual Understanding and Textual Nuances

Our third research hypotheses is that if the LLM can understand and capture nuances. Upon further analysis of post survey question Q2 *Which captured nuances or context better?* has led to support of H3 (see section 5.6). The result showed that almost all the participants agreed upon the fact that text-based survey input is where they were able to express their feelings more: one such remark by P12 *"I would prefer the descriptive option, give me the flexibility of expression"*. Also, it made the participant think about their experience rather than putting a number to it, as expressed by P1 *"I preferred descriptive, as I would think about my experience with the puzzle more."*. P23 also mentioned *"I preferred the descriptive text because I could explain and give more nuances to subjective questions that may not be able to be described with just a number."* along with that, P24 also mentioned *"I preferred the descriptive text as I was able to point out and write down exactly what I thought about each thing."*. These remarks evidently make up for the fact that when it comes to capturing nuances and participants' ability to articulate what they felt throughout the experience, via a descriptive medium is better. These are further useful for researchers for post-study analysis and improvements in future work. While this descriptive text alone could be hectic and very manual work for the researcher to understand and work around the statistics, an LLM can make that easier with its cutting-edge state-of-the-art models' understanding of language and context, given enough guardrails and prompt engineering.

We also did a manual reflexive thematic analysis on the descriptive text to understand participants' perspectives on the study (see section 5.7). We found various reasons and a more in-depth understanding of the user study was like from the participants' point of view. Based on the textual input, we extracted four main themes that resulted in understanding why participants felt Level 2 easier than Level 1 and no significant differences between other NASA-TLX dimensions except *effort*. This led us to understand why the textual nuances are important, and hence conducted a thematic analysis. While the Likert scale ratings provided standardized measure of perceived workload, the descriptive responses revealed the underlying reasons driving these perceptions. While analyzing the Temporal demand dimension, the Likert rating just tells how rushed the task felt, but descriptive responses indicated the why behind it. For example, in regards to pacing paradox, P3 stated *"I was rushed as I was looking at the timer on the side, but the pace of the task seemed pretty balanced overall."* which does indicate that the complexity of the puzzle, being easy or hard, did not cause this effect rather because of the timer. More examples with each respective theme are given in the Table 6. The descriptive texts revealed that participants' workload perceptions were not shaped by task demands but by their strategizing, learning from previous mistakes, and temporal expectations, which would have been hard to understand and missing in purely Likert scale ratings. These findings underscore the importance of incorporating qualitative methods in workload research to understand not just what participants experience.

7 Limitations and Future Work

Our work offers valuable insights into LLM based survey input methods for descriptive text. However, it has its own limitations suggesting future research directions. Although we ran a study with the general *gpt-4o-mini* API, it can perform better and give more accurate results once it can be fine-tuned with enough data collected. This improvement could potentially provide more accuracy and closeness to human-like Likert grading. Additionally, the study was conducted in a controlled setting and primarily focused on the NASA-TLX survey instrument with well defined constructs. As, a result, the findings may not generalize to more complex constructs or other survey types. Future research should extend this methodology to other types of surveys, such as System Usability Study (SUS), Simulator Sickness Questionnaire (SSQ) and other available survey instruments with more complex study designs.

Also, the entire user study was conducted based on the given factor that the participant must understand and be able to write English. This limits us to a certain population whose first language is English. With recent developments, a lot of LLM models can understand multiple languages. Another future work can be to make sure this descriptive survey method is able to take input in more than one language and process it equally, as it has showcased for the English language. The current state of the model it is able to to give a numerical grade and a brief explanation for the grade. A future direction can also be that the model is able to automatically bin the participants' responses and create codes to further simplify understanding of the surveys for the researchers which in current state is done manually.

While looking at the post-survey data, our method was heavily preferred for capturing nuances, but the overall preference was marginal approval. This can be further improved by introducing voice as input instead of text, which is easier to convey participants feelings than writing, potentially reducing user exhaustion. In future work, we will be implementing a speech-to-text model where they can record their thoughts and feelings about the experience and simultaneously convert it into text for downstream processing.

Some applications that can be developed include understanding how participants feel while running studies in real-time. It can be a feedback monitoring system that takes live feedback based on pre-configured survey questionnaires, which will understand the

participants' feelings and collect data in intervals to truly understand how participant actually felt during the study as it can evolve over time. One such study can be a flight training simulator, which can understand the mental load, physical load and many other aspects that we can only get while they are running the study.

8 Conclusion

Our research indicates that a text-based survey input method backed by an LLM can serve as an effective tool providing a nuanced contextual understanding and accurate ratings of participants' experiences compared with traditional Likert-scale ratings. Across 36 participants engaged in the puzzle-solving task with varying difficulty levels, we found that our LLM-based survey input method performed equivalently, if not better, while improving response quality by reducing the incidence of low-quality inputs such as straight-lining, which is commonly observed in Likert scale ratings.

The ability of LLMs to deliver equivalent Likert scale ratings, as indicated by the TOST analysis, highlights LLMs' ability to understand natural language and convert them into grades based on the input. Its ability to capture deeper into participant experience using free text makes it a more promising tool for more interactive, context-sensitive surveys. For reproducibility and transparency, we provide a public repository containing the materials required to run the study at *https://github.com/Akashdeep1000/NarrativetoNumbers*.

As this field advances, LLM-based survey input methods offer a promising avenue for improving efficiency, accuracy and depth of survey methodologies. By paving the way for more dynamic data collection methods in human-computer interaction research, LLMs may help bridge the gap between traditional quantitative measures and the rich, qualitative insights necessary for understanding complex human experiences.

9 GenAI Usage Disclosure

We didn't use the GenAI tool at any stage beyond improving grammar of the manuscript.

Acknowledgments

This work is supported in part by NSF Award DRL-2506427 and Army Research Lab Award W912CG2320004. The authors would also like to thank the anonymous reviewers for their valuable feedback.

References

[1] Mistral AI. 2025. Ministral 8B. https://mistral.ai/. Large language model. Model identifier: ministral-8b-latest.
[2] Anthropic. 2025. Claude Sonnet 4. https://www.anthropic.com/. Large language model. Model identifier: claude-sonnet-4-20250514; version date: 2025-05-14.
[3] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[4] Robert M Carini, John C Hayek, George D Kuh, John M Kennedy, and Judith A Ouimet. 2003. College student responses to web and paper surveys: Does mode matter? *Research in Higher Education* 44 (2003), 1–19.
[5] O Fagbohun, NP Iduwe, M Abdullahi, A Ifaturoti, and OM Nwanna. 2024. Beyond traditional assessment: Exploring the impact of large language models on grading practices. *Journal of Artifical Intelligence and Machine Learning & Data Science* 2, 1 (2024), 1–8.
[6] Naghmeh Farzi and Laura Dietz. 2024. Exam++: Llm-based answerability metrics for ir evaluation. In *Proceedings of LLM4Eval: The First Workshop on Large Language Models for Evaluation in Information Retrieval.*
[7] Francis Galton. 1883. *Inquiries into human faculty and its development.* Macmillan.
[8] Kilem L. Gwet. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters* (4th ed.). Advanced Analytics, LLC, Gaithersburg, MD.
[9] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology.* Vol. 52. Elsevier, 139–183.
[10] Michele M. Hayslett and Barbara M. Wildemuth. 2004. Pixels or pencils? The relative effectiveness of Web-based versus paper surveys. *Library & Information Science Research* 26, 1 (2004), 73–93. doi:10.1016/j.lisr.2003.11.005
[11] Dirk Heerwegh and Geert Loosveldt. 2008. Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public opinion quarterly* 72, 5 (2008), 836–846.
[12] Carolin Kaiser, Jakob Kaiser, Vladimir Manewitsch, Lea Rau, and Rene Schallner. 2025. Simulating Human Opinions with Large Language Models: Opportunities and Challenges for Personalized Survey Data Modeling. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization.* 82–86.
[13] Arshnoor Kaur, Amanda Aird, Harris Borman, Andrea Nicastro, Anna Leontjeva, Luiz Pizzato, and Dan Jermyn. 2025. Synthetic Voices: Evaluating the Fidelity of LLM-Generated Personas in Representing People's Financial Wellbeing. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization.* 185–193.
[14] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI conference on human factors in computing systems.* 1–12.
[15] Andrey Krekhov, Sebastian Cmentowski, Katharina Emmerich, Maic Masuch, and Jens Krüger. 2018. GulliVR: A walking-oriented technique for navigation in virtual reality games based on virtual body resizing. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play.* 243–256.
[16] Anna Krizanova, George Lăzăroiu, Lubica Gajanova, Jana Kliestikova, Margareta Nadanyiova, and Dominika Moravcikova. 2019. The effectiveness of marketing communication and importance of its evaluation in an online environment. *Sustainability* 11, 24 (2019), 7016.
[17] Jungjae Lee, Yubin Choi, Minhyuk Song, and Sanghyun Park. 2024. ChatFive: Enhancing User Experience in Likert Scale Personality Test through Interactive Conversation with LLM Agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces.* 1–8.
[18] Jonathan Mellon, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman. 2024. Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics* 11, 1 (2024), 20531680241231468.
[19] James Miller, John Daly, Murray Wood, Andrew Brooks, and Marc Roper. 1996. Electronic bulletin board distributed questionnaires for exploratory research. *Journal of Information Science* 22, 2 (1996), 107–115.
[20] OpenAI. 2024. GPT-4o mini. https://platform.openai.com/ Large language model accessed via API.
[21] OpenAI. 2025. GPT-4.1 mini. https://platform.openai.com/. Large language model. Model identifier: gpt-4.1-mini.
[22] Michael D Ornstein. 2013. *A companion to survey research.* SAGE Publications Ltd.
[23] A Ant Ozok. 2007. Survey design and implementation in HCI. In *The human-computer interaction handbook.* CRC Press, 1177–1196.
[24] Cleo Protogerou and Martin S Hagger. 2020. A checklist to assess the quality of survey studies in psychology. *Methods in Psychology* 3 (2020), 100031.
[25] Johannes Schneider, Bernd Schenk, Christina Niklaus, and Michaelis Vlachos. 2023. Towards llm-based autograding for short textual answers. *arXiv preprint arXiv:2309.11508* (2023).
[26] Donald J Schuirmann. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics* 15 (1987), 657–680.
[27] Eleanor Singer and Mick P Couper. 2017. Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)* 11, 2 (2017), 115–134.
[28] Tushar Singh and Himangshu Kumar. 2025. AI in qualitative research: Using large language models to code survey responses in native languages. IFPRI Blog. https://www.ifpri.org/blog/ai-in-qualitative-research-using-large-language-models-to-code-survey-responses-in-native-languages/
[29] Tom William Smith, Jibum Kim, Achim Koch, and Alison Park. 2006. *Social-science research and the general social surveys.* NORC/University of Chicago.
[30] Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. 2000. *The psychology of survey response.* Cambridge University Press.
[31] Constantin E Uhlig, Berthold Seitz, Nicole Eter, Julia Promesberger, and Holger Busse. 2014. Efficiencies of Internet-based digital and paper-based scientific surveys and the estimated costs and time for different-sized cohorts. *PloS one* 9, 10 (2014), e108441.

[32] xAI. 2025. Grok-3 Mini. https://x.ai/. Large language model. Model identifier: grok-3-mini.

[33] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 3 (2020), 1–37.

[34] Wenjing Xie, Juxin Niu, Chun Jason Xue, and Nan Guan. 2024. Grade Like a Human: Rethinking Automated Assessment with Large Language Models. http://arxiv.org/abs/2405.19694 arXiv:2405.19694 [cs].

[35] ZhenTing Yan, Rui Zhang, and Fei Jia. 2024. Exploring the Potential of Large Language Models as a Grading Tool for Conceptual Short-Answer Questions in Introductory Physics. In *Proceedings of the 2024 9th International Conference on Distance Education and Learning*. 308–314.

[36] Brahim Zarouali, Theo Araujo, Jakob Ohme, and Claes de Vreese. 2024. Comparing chatbots and online surveys for (longitudinal) data collection: an investigation of response characteristics, data quality, and user evaluation. *Communication Methods and Measures* 18, 1 (2024), 72–91.