# Improving Human Answers Quality by Machine Questions Number and Context Factors

**Haonan Zhao (haonan.zhao@unitn.it)**
University of Trento, Trento, Italy

**Xiaoyue Li*** **(xiaoyue.li@unitn.it)**
University of Trento, Trento, Italy

## Abstract

Mobile phones provide an opportunity for a symbiotic interaction between humans and machines, which allows phones to collect human-centric data at anytime and anywhere. However, low-quality answers, which refer to the wrong answers, may be provided by users when they are asked excessive questions or in unsuitable contexts (e.g., driving). To solve this problem, we aim to design a methodology to collect more correct answers. We propose to use answer reaction time to annotate answer quality, to find a suitable number of daily questions, and the context factors that need to be considered according to their history records. We validated our methodology via the public dataset, which was collected by an extensive four-week in-the-wild study at the University of Trento, Italy. The results reveal that the context information and the number of daily questions are factors that can impact user answer behavior. These factors, therefore, influence the answer quality.

**Keywords:** Human-Centric AI; Social Computing; Context; Data Quality.

## Introduction

Smartphones, now an indispensable facet of our daily lives, facilitate the receipt of notifications and provision of responses at any location and any time (Runyan et al., 2013). Current research in the Human-Computer Interaction domain has leveraged the use of smartphones to ask questions for human answers and sensor data to understand human behavior, see, e.g., social interactions (Eagle & Pentland, 2009), points of interest (Alessandretti, Sapiezynski, Sekara, Lehmann, & Baronchelli, 2018), and social connections (Miritello, Lara, Cebrian, & Moro, 2013). Also, the capacity to process context (Schilit & Theimer, 1994; Giunchiglia, 1993) , which allows humans to adjust their behavior according to their environment, is a critical aspect of various fields, including Human-Centered AI, Psychology, and Data Mining (Giunchiglia, Li, Busso, & Rodas-Britez, 2023; Hoareau & Satoh, 2009; Li, Rodas-Britez, Busso, & Giunchiglia, 2022). But the inherent complexity of context often poses challenges in directly acquiring high-quality context data. Currently, smartphones offer an opportunity to collect subjective context data (e.g., via time-use diaries) directly and promptly from users, namely, ask context questions and get answers from humans directly.

While it is possible to use the Experience Sampling Method (Larson & Csikszentmihalyi, 2014) or time diaries (Bowers, 1939) to gather data about the human context via smartphone, the collection of such data presents two significant challenges: the first relates to the quality of human input data. The low-quality data, namely the given wrong answers or without given answers, can not help machines to know humans well because this kind of data does not meet the standards required by AI (Mehrotra, Vermeulen, Pejovic, & Musolesi, 2015). The second challenge is the considerable number of questions that humans are expected to answer due to the fact that many fields (e.g., deep learning and AI) need to get a huge amount of data from them. The frequent questions from smartphones can be intrusive, especially when they interrupt users during periods of activity (e.g., driving a car or taking a meeting). This can result in annoyance and a reduction in work and study efficiency (Renaud, Ramsay, & Hair, 2006). This can invariably lead to survey fatigue, where participants become weary or lose interest in the survey, resulting in low answer quality (Jeong et al., 2023).

In light of the above challenges, it is urgent to devise a methodology that sends less questions via the suitable context to require human answers while preserving the quality of the answers received. Consequently, our research objective is *to evaluate the feasibility of amassing high-quality answers by making sure of the number of daily questions and the suitable context information, thereby minimizing disruptions to the participant while accruing high-quality data from each participant.* Hereby, the specific kind of answer quality means the quality of participant answers to machine-generated context questions sent by smartphone. To quantify the quality of the answers, we follow the studies from (Bison & Zhao, 2023; van Berkel et al., 2019), which propose that a reduced reaction time contributes to enhanced answer quality; the reaction time means the elapsed time from notification to answer.

The data analysis was conducted on an ESM interval-based experiment, compilation from 268 students over a span of four weeks, approximately 400,000 responses with corresponding sensor collections that we have executed at the University of Trento. This experiment was fundamentally predicated on the utilization of an ESM application christened i-Log (Zeni, Zaihrayeu, & Giunchiglia, 2014), which facilitates the collection of sensor data, predominantly but not exclusively from the smartphone, and the administration of queries about their situational context, in the form of *time diaries*, i.e., clusters of queries about the prevailing context, posed

*Corresponding author



In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

multiple times, at various programmable junctures of the day (Bowers, 1939).

Subsequently, we propose an innovative approach that, through meticulous selection and optimization of the timing of queries posed, can alleviate user fatigue and enhance the reliability of the amassed data. We present a methodology that employs Machine Learning methodologies, particularly Random Forest, to predict the time periods when a participant is most likely to provide high-quality responses. Here we selected 170 participants who have responded to all queries in the experiment, in the period, we collected approximately 100,000 responses, including the time diaries about the context surrounding the participant, by context, we refer to the person's situational context, including their internal state as embedded in a reference context or environment that is shared with other people. In this manner, we not only reduce the frequency of query delivery to participants but also identify the optimal times for them to provide high-quality responses, thereby enhancing the quality of substantial data in the human-centered AI system.

## Related Work

The collection of human context data is achieved through two methods: the use of smartphone sensors and the acquisition of information directly from participants. The former operates continuously on the user's phone, while the latter leverages Time Use Diaries (TUDs) (Bowers, 1939). Mobile self-reports (Boase & Ling, 2013) have garnered substantial momentum in recent years; this methodology primarily collects data through a self-completed time diary report that logs a participant's activities at fixed time intervals, which offers a comprehensive view of how individuals allocate their time across various activities, yielding valuable insights into patterns of behavior and lifestyle (Bauman, Bittman, & Gershuny, 2019; Sarker, Hoque, Uddin, & Alsanoosy, 2021). However, assessing the accuracy of self-report responses has received little attention in the previous, despite the considerable implications for study results. Researchers tend to presuppose the accuracy of all ESM responses, without further validating this assumption (van Berkel et al., 2019).

Comprehending the user's context is pivotal for ascertaining an opportune moment to dispatch messages. In psychology, the ESM (Larson & Csikszentmihalyi, 2014) empowers researchers to devise intensive, repeated measure questionnaires that furnish comprehensive context information about the interviewee. Moreover, the incessant progression of smartphones has culminated in the emergence of a novel technology known as smartphone-based ESM (Mehrotra et al., 2015). This technology enables the documentation of human behavior and experiences by querying users and scrutinizing people's time-use. Users can be queried about human behavior, encompassing daily events and moods, via their smartphones (Clark & Watson, 1988). A challenge in architecting ESM systems is identifying a moment to pose questions that enhances participant engagement and ameliorates the quality of data collection.

As underscored by Mehrotra et al. (Mehrotra et al., 2015), obtaining high-quality data with ESM is challenging because ESM consistently sends questions at the same time interval, which can sometimes annoy participants. Numerous studies aim to improve the response rate of questions, i.e., the number of answered questions. For instance, from the perspective of smartphone usage, Boukhechba et al. (Boukhechba et al., 2018) observed a higher response rate when ESM questions were sent after a phone call compared to using social media for text communication. Similarly, Berkel et al. (van Berkel et al., 2020) also found that active phone use before the ESM question could increase the response rate. From the situational context perspective, researchers have explored how changes in location (e.g., at home and outdoors) (Sun, Rhemtulla, & Vazire, 2021), different times of the day (Pielot et al., 2017), and social interactions (Boukhechba et al., 2018) impact the response rate of ESM questions. Although these investigations have not extensively probed into the quality of responses or devised a method for its enhancement, Bison et al. (Bison & Zhao, 2023) established that elevated reaction time for questions could degrade the quality of answers. Consequently, we also select reaction time as a pivotal determinant of the quality of human answers.

## Methodology

In the sociological survey research, the Total Survey Error (TSE) paradigm (Amaya, Biemer, & Kinyon, 2020; Biemer, 2010; Sen, Floeck, Weller, Weiss, & Wagner, 2019) is extensively discussed to maximize data quality within constraints, which is beneficial for optimizing surveys. We propose a methodology to improve data quality by recognizing specific constraint factors. These factors are considered and used in the questioning mechanism of machines and are not influenced by various human subjective responses.

### Answer quality and reaction time

In the area of human-centric data collection, researchers aspire to obtain immediate answers to questions from participants (van Berkel et al., 2019; Bison & Zhao, 2023; Bison, Zhao, & Giunchiglia, 2024). This aspiration stems from the understanding that the longer of delay between the request and the answer, the greater the risk of introducing errors into answers due to, e.g., memory forget effects. As the study (Janssens, Bos, Rosmalen, Wichers, & Riese, 2018) observed, some researchers stated that they elected to permit participants with a relatively short delay to answer questions to facilitate real-time assessment and mitigate recall bias. However, to enhance participants' compliance, some degree of delay must be allowed, as participants are engaged in their regular lives and may not always be in a position to respond instantaneously. Motivated by these studies, we measure the quality of answers by *reaction time*. We define *reaction time* as the time interval between a participant receiving questions and starting to give answers. We formalize *reaction time*, de-

noted as $t_r$, as:

$$t_r = t_a - t_n \tag{1}$$

where $t_a$ is *answer time*, which is the time when a participant starts to answer the questions. The $t_n$ is *notification time*, which is the time when a participant receives the questions that are sent by machines. If the participant did not answer a question (e.g., $t_a$ = '*null*'), we set $t_r = \infty$. In real-world datasets, the notification time and answer time are always recorded by *timestamp* datatype, which stores the year, month, day, hour, minute and second values, e.g., '2024-07-01 10:11:20' in MySQL Database. To facilitate the representation and application of reaction time, we invariably set the unit of reaction time to minutes, e.g., '2.8' minutes.

The reaction time to a question can influence the quality of the answer. The shorter reaction time leads to better answer quality. To quantify answer quality, we binary classify answer quality as high-quality or low-quality responses.

$$Answer \quad Quality = \begin{cases} High, & \text{if } t_r < RT \\ Low, & \text{if } t_r \geq RT \end{cases} \tag{2}$$

where $RT$ is a cut-off point as the time threshold that distinguishes answer quality. If the reaction time of an answer $t_r$ is less than the reaction time cut-off point $RT$, the quality of the answer is considered 'High'. On the other hand, if the reaction time $t_r$ is greater or equal to $RT$, the quality of the answer is considered 'Low'. Of course, if the participant did not answer a question, we have $t_r = \infty$. Hence, this answer is considered a low-quality answer.

There is a relationship between reaction time and answer quality. For different datasets, the cut-off point of reaction time is hard to determine, depending on the proportion of good-quality answers in the databases and the expected quality from answer collectors, etc. Findings from pencil-paper diary studies suggest that the majority of participants respond within a 10 to 20-minute window (Csikszentmihalyi & Larson, 1987). The ESM answers delayed by more than 15 minutes exhibit diminished reliability (Wichers et al., 2007; Delespaul, 1995). Hence, we proposed the cut-off point of reaction time, $RT$, can be set as setting **15 minutes** for differentiating between high-quality and low-quality answers.

## Answer quality factors

We present two types of factors, namely *number of questions* and *context information*, which can have an impact on the answer quality.

**Number of questions:** The number of questions is always designed by the experiment manager before the beginning of an experiment for better questioning, e.g., considering how many answers are needed from the participants. Experiment managers are willing to get plenty of answers from participants, so they would like to ask many questions per day (Eisele et al., 2022). Jeong et al. (Jeong et al., 2023) indicate that an excessive quantity of questions can invariably precipitate survey fatigue, a state wherein participants experience weariness or diminished interest in the survey, thereby compromising the accuracy of answers. Consequently, the number of questions emerges as a factor influencing answer quality. Our aim is to send fewer questions to participants but get a higher percentage of high-quality answers. And we propose that the number of daily questions should be the number of high-quality answers that could be provided by most participants.

**Context information:** Context is widely used in the HCI and Knowledge Representation (KR) communities, which is defined as 'any information that can be used to characterize the situation of an entity' (Dey, 2001). Context is the situational setting of an individual that encompasses their internal condition within a common reference environment with others (Giunchiglia, Bignotti, & Zeni, 2017; Intille, Rondoni, Kukla, Ancona, & Bao, 2003; Runyan et al., 2013). Current research on context primarily finds context could impact notifications (Mishra, Lowens, Lord, Caine, & Kotz, 2017) and enhancement of question response rates (Sun et al., 2021), thereby impacting the accuracy of answers. Therefore, we choose context information as the factor influencing answer quality. The contexts we consider in this paper are as follows:

**Time context:** We follow the suggestions from Chen and Kotz (Chen & Kotz, 2000), as time is a fundamental and natural context in which human, physical, and computing contexts are recorded over time to compile a context history. Time also impacts whether a participant will answer a question (Sun et al., 2021). We segmented the time context for a survey into three categories: *hour*, *weekday* and *day*.

**Situational context:** Situational context is a multidimensional concept where environmental, social, and technical conditions interact. Current research (Mishra et al., 2017) said context can impact if a participant gives feedback to a notification; thereby, context could influence the answer quality. In this paper, we consider situational context as three dimensions that are the *event context*, the *spatial context*, and the *social context*, which are defined in (Giunchiglia et al., 2017).

The *number of question* factor has been proven to impact answer quality in our experiment in the previous section. Our experiments have demonstrated that the factors *hour*, *weekday* and *day* (from time context) plus *activity* (from event context), *location* (from spatial context), and *interacting people* (from social context) impact quality of answers.

## Workflow

We elucidate a step-to-step general workflow, which aims to find the number of questions and context information that can impact the answer quality. The workflow includes three main steps.

**Step 1: Annotating answer quality.** Answers can be annotated with labels by considering their quality as 'High' or 'Low'. This can be achieved by first calculating the reaction time of these answers using Equation (1) (reaction time ($t_r$) values of unanswered records are set as $\infty$). Subsequently, we use Equation (2) to indicate the quality of the answers (where

*RT* is set as 15 minutes). Based on the answer quality, we annotated answers with labels (e.g., 'High' or 'Low'; '1' or '0', etc).

**Step 2: Determining the daily question number for acquiring more high-quality answers.** The number of daily questions is designed to be the number of high-quality answers that most participants can provide every day. First, we calculate the average number of high-quality answers that each participant provides. Secondly, participants are grouped based on their average number of high-quality answers, i.e., two participants are in one group only if they have the same average number of high-quality answers. Finally, we can find the group with the most participants. The number of daily questions is an integer that is closest to the average number of high-quality answers in this group.

**Step 3: Identifying context factors influencing answer quality.** To test whether there is a relationship between answer quality and context information, we first use the Chi-Square statistical test to evaluate whether there is a statistical difference between various context factors and the reaction time of answers; here, we assume the reaction time was independent of context factors. The result of the *P value* will show the relationship between answer quality and context information. For example, in the relationship between activity and reaction time, if the $P < 0.01$, that there is a strong significance between the context information and reaction time. Based on the Equation (2), we can know there is a strong significance between the context information and answer quality. Next, we calculate the number of high-quality answers that participants can provide based on different context factors, which enables the observation of the different contexts to influence participants' high-quality answers.

Given the above, answers from a dataset are annotated with labels by considering their quality from *Step 1*. We acquire a certain number of daily questions that we should send to the participants every day from *Step 2*. The evaluation experiments using machine learning algorithms can be designed to test whether sending a certain number of daily questions and considering the observed context information (from *Step 3*) can achieve a greater percentage of high-quality answers than the original dataset, which means considering the number of questions and context information can improve the answer quality.

# Experiment

## Dataset

Our group collected the WeNet dataset. The WeNet dataset engaged in excess of 10,000 students across eight distinct nations, with the objective of investigating human lifestyles [1]. As a European initiative, the data collection procedure was executed in compliance with the EU General Data Protection Regulation (GDPR) and procured approval from the ethics committee. The pertinent details of the data have been lucidly delineated in (Giunchiglia et al., 2021), which provided detailed experimental protocols, participant recruitment, and data cleaning methodologies.

In the data collection experiments, we employed a mobile sensing i-Log application (Zeni et al., 2014), which sends time diary questions to all the participants. These questions include three HETUS questions[2]: *Location*: 'Where are you?'; *Activity*: 'What are you doing?'; *Interacting people*: 'With whom are you?'. All the questions were asked in thirty minutes in the first two weeks and one hour in the last two weeks. The variables of time diaries we use include:

- *participantid*: The participant's identifiers, value integers from 0 to 169.
- *notification time*: The time when questions are arrived at the smartphone, e.g., '1970-01-18 10:05:09'.
- *answer time*: The time when questions are opened by the participant, e.g., '1970-01-18 10:10:26'.
- *location*: User answers of the `location` question representing their location, which provides the participant with 26 categories such as home, workplace, university, restaurant, etc.
- *activity*: The user answers the `activity` question representing their event, providing the participant with 34 answer categories such as sleeping, eating, working, etc.
- *interacting people*: The user answers the `interacting people` representing their interacted people, which provides the participant with 8 categories such as nobody, partner, friends, etc.

To implement our methodology, we used the WeNet-Italy dataset, which is a constituent of the WeNet dataset. WeNet-Italy dataset was collected by conducting this investigation at the University of Trento, where all participant students were approached via email. From the pool of over 5,000 students, a group of 350 was selected, based on their individual fields of study. These chosen students were provided with supplementary instructions on the downloading and utilization of the application. In the final analysis, the experiment was successfully completed by 170 participants. The demographic information is shown in Table 1.

## Answer quality annotation

To clarify the annotation process of answer quality. We show the part of the pre-processed and anonymized (e.g., '1970-01-08') dataset, as shown in the *Day*, *Activity Answer*, *Notification Time* and *Answer Time* columns in Table 2, where values of *Activity Answer* are participants' answers to the question of 'What are you doing?'. Based on our workflow, we aim

[1] For an exhaustive description of the project, as well as the option to download the WeNet dataset, kindly refer to the provided URL: `https://www.internetofus.eu/`.

[2] *Harmonized European Time Use Surveys (HETUS)* standard can be found in `https://ec.europa.eu/eurostat/web/time-use-surveys`.

Table 1: The demographic information.

| Feature | Sex | | Degree | | Department | | | |
|---|---|---|---|---|---|---|---|---|
| | Female | Male | BSc | MA+PhD | Information Science | Industrial | Business | Sociology |
| Number | 86 | 84 | 138 | 32 | 53 | 23 | 44 | 50 |
| Percentage | 50.58% | 49.42% | 81.17% | 18.83% | 31.18% | 13.52% | 25.88% | 29.41% |

to calculate the reaction time and then annotate answer quality for each record in our WeNet-Italy dataset, as examples of *Reaction Time* and *Answer Quality* columns in Table 2.

Each record includes the value of *Notification Time*, namely the time when the participant's smartphone receives this question, and the value of *Answer Time*, namely when the participant starts to answer this question. Based on the Equation 1, we can calculate the reaction time for each record that has been answered. If a participant gives no answer to the question, we set *Reaction Time* values $\infty$. Secondly, after knowing the reaction time for all of the records in the WeNet-Italy dataset, based on Equation 2 and cut-off point $RT = 15$ *minutes*, we can annotate the values of the *Answer Quality* column as '1' or '0', where '1' represents the high-quality answer, and '0' represents the low-quality answer.

Table 2: Example for a part of the annotated dataset.

| Day | Activity Answer | Notification Time | Answer Time | **Reaction Time** | **Answer Quality** |
|---|---|---|---|---|---|
| 1970-01-18 | {"Eating"} | 1970-01-18 10:05:09 | 1970-01-18 10:10:09 | **5** | **1** |
| 1970-01-18 | null | 1970-01-18 12:06:01 | null | $\infty$ | **0** |
| 1970-01-18 | {"Sport"} | 1970-01-18 14:03:08 | 1970-01-18 14:12:22 | **9.14** | **1** |
| 1970-01-19 | null | 1970-01-19 12:06:01 | null | $\infty$ | **0** |
| 1970-01-19 | {"Studying"} | 1970-01-19 13:06:06 | 1970-01-19 13:43:16 | **37.1** | **0** |

**Result:** After labeling answer quality into the WeNet-Italy dataset, we can conduct the statistical analysis on the quality of answers in the original WeNet-Italy dataset (which is the baseline of our evaluation, as shown in Table 3). The 34.26% of questions were answered by high quality. In addition, on Monday and Saturday, participants give more high-quality answers (occupying 38.56% and 39.17%), while on Friday, participants give less high-quality answers (occupying 31.24%). It means we find that, at the beginning of weekdays/weekends, people always provide more high-quality answers.

## Daily questions number

For each participant in the WeNet-Italy dataset, we calculate the average number of high-quality answers that she/he provided per day. These participants are grouped according to the average number. The results show that the biggest group includes 20 participants. They provided an average of 9.6 high-quality answers every day. The second biggest group includes 14 participants, who gave an average of 17.8 answers with high quality per day. However, for the smallest group, 4 participants averagely answered 2.4 questions with high-quality answers. As discussed in *Step 2*, the daily question number should be the integer around the average number of high-quality answers of the biggest group,i.e., 10.

**Result:** Upon analyzing the number of high-quality answers each participant can contribute daily, we have elected to present ten questions per day. This decision aims to minimize any potential disruption to the participants' daily routines and maximize high-quality answers.

## Context information

To find the context information that influence the answer quality. We first employed the Chi-Square statistical tests on context factors (namely weekday, day, hour, activity, location and interacting people from WeNet-Italy dataset) with the reaction time of answers. To avoid the Type I error, we first have employed the Benjamini-Hochberg Procedure (Beyer, 1978) to process data. Then when assuming the reaction time was independent of context factors, the Chi-Square results showed that there is a significant discrepancy between our observed counts and the expected counts, which proving the proposed context factors have a strong significance with the reaction time. As we have discussed that the reaction time of answers involves the standard for evaluating answer quality, therefore we can know the proposed context factors influence answer quality. Due to the limitation of the space of this paper, we do not show the complete Chi-Square tests results in detail. We give some result examples: For the activity factor, the median reaction time of responses fluctuated from a minimum of 12 minutes when the participant engages in social media/phone/chat activities to a maximum of 44 minutes during free time activities ($\chi^2 = 1107.75$, $p<0.01$). For the interacting people factor, the median reaction time varies from 12 minutes when the participant is alone to 68 minutes in a social setting ($\chi^2 = 746.38$, $p<0.01$). For the location factor, the median reaction time spans from 9 minutes when the participant is at a university to 44 minutes when outdoors ($\chi^2 = 674.69$, $p<0.01$).

Furthermore, we observe the weekday, day, hour, activity, location and interacting people factors that influence the number of high-quality answers. Here, we only show two examples as shown in Figure 1 and Figure 2 due to the limitation of space. Figure 1 shows the weekday factor (one of time context) influences the number of high-quality answers. The trend line indicates decreasing numbers of high-quality responses from *Monday* to *Friday*, with median values from 9.25 to 7.5. Interestingly, there is a significant increase in the median number of high-quality answers from *Friday* to *Saturday*, from 7.5 to 9.5.

Figure 2 shows the activity factor influences the number of high-quality answers from participants. The *Study* is the activity with the highest median number of high-quality an-

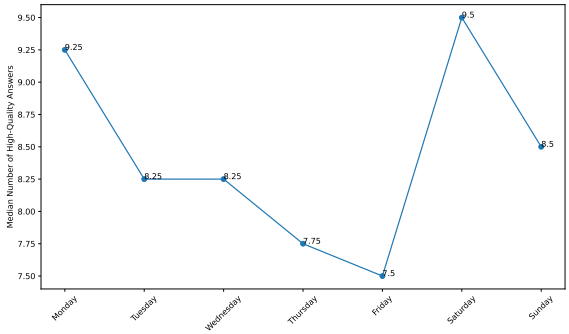


Figure 1: The median number of high-quality answers per day from participants on weekdays.

swers per day at 9.88, which demonstrates that participants provide more high-quality answers when they are studying. Due to our participants being students; they always spend the most time on study. Both the *Eating* and *Watching TV* activities have a median of 5.5 high-quality answers every day, which is the second highest value. The activities with the lowest medians are *Listening music*, *Break coffee*, *Phone calling*, and *Other* with median values of 1.12, 1, 1, and 0.88 respectively. This indicates that participants are least productive in terms of providing high-quality answers during these activities.

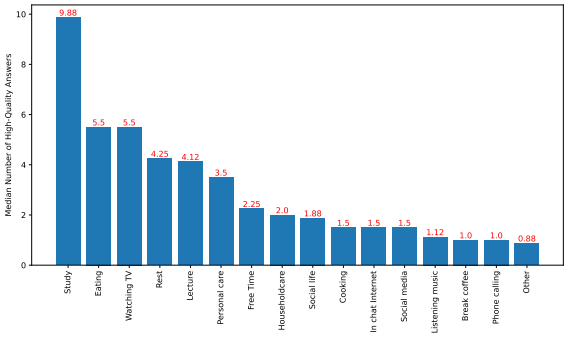


Figure 2: The median number of high-quality answers per day when participants conduct different activities.

**Result:** Based on the analysis in this section, the context information in WeNet-Italy dataset, including weekday, day, hour, activity, location and interacting people, impact acquiring high-quality answers from participants. These factors needed to be learned by machines to improve the quality of human answers.

## Evaluation

We conducted evaluation machine learning experiments by using the most common classifiers, i.e., Random Forests, Decision Trees, Artificial Neural Networks, Logistic Regression, and Gaussian Naive Bayes. We proposed we can learn from participants real context information and answer quality in the first two weeks, then predict in the following two weeks when to ask questions in a way that gets their answers with high quality. For the second two weeks, we asked 10 questions every day to participants during the machine learning suggested time slots.

From the results of Machine learning experiments, Table 3 presents the percentage of high-quality answers. The baseline is calculated by the percentage of high-quality answers out of all answers in the original WeNet-Italy dataset. We also show the percentages of high-quality answers based on the machine learning results in the second two weeks.

The baseline overall percentage is 34.26%. The overall percentage of high-quality answers predicted by the Random Forest classifier is the highest at 62.28The other rows break down these percentages by each day of the week. For each day, the percentages of high-quality answers predicted by the classifiers are higher than the baseline percentage, indicating an improvement across all days. The highest percentage of high-quality answers is predicted by the Random Forest classifier on Monday (64.03%), while the lowest is predicted by the Logistic Regression classifier on Thursday (56.02%). All of the algorithms we test can get a huge improvement from the original baseline, which can support our methodology; that is, sending a suitable number of questions (fewer than before) based on the participant context information can improve the quality of the answers. We infer this because fewer questions and under specific contexts make the participants feel less annoyed. Hence, they prefer to provide more high-quality answers.

Table 3: Comparation of percentages of high-quality answers.

| Weekday | BaseLine | Predictions Via Our Methodology | | | | |
|---|---|---|---|---|---|---|
| | | Random Forest | Decision Tree | Gaussian Naive Bayes | Artificial Neural Networks | Logistic Regression |
| Monday | 38.56% | **64.03%** | 63.78% | 60.90% | 63.30% | 58.39% |
| Tuesday | 36.80% | 62.27% | **62.57%** | 58.82% | 61.91% | 57.47% |
| Wednesday | 36.74% | **63.08%** | 60.08% | 58.65% | 61.18% | 56.47% |
| Thursday | 33.53% | 60.17% | **61.46%** | 57.35% | 58.67% | 56.02% |
| Friday | 31.24% | **60.82%** | 60.65% | 57.51% | 59.60% | 56.86% |
| Saturday | 39.17% | 63.35% | **63.40%** | 60.42% | 61.56% | 58.30% |
| Sunday | 37.64% | **61.21%** | 60.71% | 56.30% | 58.89% | 57.45% |
| Overall | 34.26% | **62.28%** | 61.87% | 58.56% | 60.73% | 57.28% |

## Conclusion

This study has demonstrated the significance of the number of daily questions and context factors on the quality of human answers in mobile surveys. By analyzing a part of the WeNet dataset, we have identified that reducing the number of daily questions and considering contextual information can enhance the quality of data collected. Our methodology offers a promising approach to optimizing mobile surveys for human-centered AI systems. The findings suggest that a thoughtful balance between the quantity of questions and the timing of their delivery can lead to more accurate and reliable data. Future research may explore the integration of additional contextual factors and the refinement of predictive models to further improve the efficacy of data collection in AI applications.

## References


Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S., & Baronchelli, A. (2018). Evidence for a conserved quantity in human mobility. *Nature human behaviour*, *2*(7), 485–491.

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total error in a big data world: adapting the tse framework to big data. *Journal of Survey Statistics and Methodology*, *8*(1), 89–119.

Bauman, A., Bittman, M., & Gershuny, J. (2019). A short history of time use research; implications for public health. *BMC public health*, *19*, 1–7.

Beyer, W. H. (1978). Crc standard mathematical tables. *West Palm Beach*.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public opinion quarterly*, *74*(5), 817–848.

Bison, I., & Zhao, H. (2023). Factors impacting the quality of user answers on smartphones [Conference paper]. In *Proceedings of the second international conference on hybrid human-machine intelligence (hhai 23)* (Vol. 3456, p. 208 – 213).

Bison, I., Zhao, H., & Giunchiglia, F. (2024). What impacts the quality of the user answers when asked about the current context? *arXiv preprint arXiv:2405.04054*.

Boase, J., & Ling, R. (2013). Measuring mobile phone use: Self-report versus log data. *Journal of Computer-Mediated Communication*, *18*(4), 508–519.

Boukhechba, M., Cai, L., Chow, P. I., Fua, K., Gerber, M. S., Teachman, B. A., & Barnes, L. E. (2018). Contextual analysis to understand compliance with smartphone-based ecological momentary assessment. In *Proceedings of the 12th eai international conference on pervasive computing technologies for healthcare* (pp. 232–238).

Bowers, R. V. (1939). *Time budgets of human behavior.* JSTOR.

Chen, G., & Kotz, D. (2000). A survey of context-aware mobile computing research.

Clark, L. A., & Watson, D. (1988). Mood and the mundane: relations between daily life events and self-reported mood. *Journal of personality and social psychology*, *54*(2), 296.

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *The Journal of nervous and mental disease*, *175*(9), 526–536.

Delespaul, P. A. (1995). Assessing schizophrenia in daily life: The experience sampling method.

Dey, A. K. (2001). Understanding and using context. *Personal and ubiquitous computing*, *5*(1), 4–7.

Eagle, N., & Pentland, A. S. (2009). Eigenbehaviors: Identifying structure in routine. *Behavioral ecology and sociobiology*, *63*(7), 1057–1066.

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151.

Giunchiglia, F. (1993). Contextual reasoning. *Epistemologia, special issue: I Linguaggi e le Macchine*, *16*, 345-364.

Giunchiglia, F., Bignotti, E., & Zeni, M. (2017). Personal context modelling and annotation. In *2017 ieee international conference on pervasive computing and communications workshops (percom workshops)* (p. 117-122).

Giunchiglia, F., Bison, I., Busso, M., Chenu-Abente, R., Rodas, M., Zeni, M., ... others (2021). A worldwide diversity pilot on daily routines and social practices (2020).

Giunchiglia, F., Li, X., Busso, M., & Rodas-Britez, M. (2023). A context model for personal data streams. In *Web and big data: 6th international joint conference, apweb-waim 2022, nanjing, china, november 25–27, 2022, proceedings, part i* (pp. 37–44).

Hoareau, C., & Satoh, I. (2009). Modeling and processing information for context-aware computing: A survey. *New Generation Computing*, *27*(3).

Intille, S. S., Rondoni, J., Kukla, C., Ancona, I., & Bao, L. (2003). A context-aware experience sampling tool. In *Chi'03 extended abstracts on human factors in computing systems* (pp. 972–973).

Janssens, K. A., Bos, E. H., Rosmalen, J. G., Wichers, M. C., & Riese, H. (2018). A qualitative approach to guide choices for designing a diary study. *BMC medical research methodology*, *18*, 1–12.

Jeong, D., Aggarwal, S., Robinson, J., Kumar, N., Spearot, A., & Park, D. S. (2023). Exhaustive or exhausting? evidence on respondent fatigue in long surveys. *Journal of Development Economics*, *161*, 102992.

Larson, R., & Csikszentmihalyi, M. (2014). The experience sampling method. In *Flow and the foundations of positive psychology* (pp. 21–34). Springer.

Li, X., Rodas-Britez, M., Busso, M., & Giunchiglia, F. (2022). Representing habits as streams of situational contexts. In *Advanced information systems engineering workshops: Caise 2022 international workshops, leuven, belgium, june 6–10, 2022, proceedings* (pp. 86–92).

Mehrotra, A., Vermeulen, J., Pejovic, V., & Musolesi, M. (2015). Ask, but don't interrupt: the case for interruptibility-aware mobile experience sampling. In *Adjunct proceedings of the 2015 acm international joint conference on pervasive and ubiquitous computing and proceedings of the 2015 acm international symposium on wearable computers* (pp. 723–732).

Miritello, G., Lara, R., Cebrian, M., & Moro, E. (2013). Limited communication capacity unveils strategies for human interaction. *Scientific reports*, *3*(1), 1–7.

Mishra, V., Lowens, B., Lord, S., Caine, K., & Kotz, D. (2017). Investigating contextual cues as indicators for ema delivery. In *Proceedings of the 2017 acm international joint conference on pervasive and ubiquitous computing and proceedings of the 2017 acm international symposium on wearable computers* (pp. 935–940).

Pielot, M., Cardoso, B., Katevas, K., Serrà, J., Matic, A., & Oliver, N. (2017). Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1–25.

Renaud, K., Ramsay, J., & Hair, M. (2006). " you've got e-mail!"... shall i deal with it now? electronic mail from the recipient's perspective. *International Journal of Human-Computer Interaction*, *21*(3), 313–332.

Runyan, J. D., Steenbergh, T. A., Bainbridge, C., Daugherty, D. A., Oke, L., & Fry, B. N. (2013). A smartphone ecological momentary assessment/intervention "app" for collecting real-time data and promoting self-awareness. *PloS one*, *8*(8), e71325.

Sarker, I. H., Hoque, M. M., Uddin, M. K., & Alsanoosy, T. (2021). Mobile data science and intelligent apps: concepts, ai-based modeling and research directions. *Mobile Networks and Applications*, *26*, 285–303.

Schilit, B. N., & Theimer, M. M. (1994). Disseminating active map information to mobile hosts. *IEEE network*, *8*(5), 22–32.

Sen, I., Floeck, F., Weller, K., Weiss, B., & Wagner, C. (2019). A total error framework for digital traces of humans. *arXiv preprint arXiv:1907.08228*.

Sun, J., Rhemtulla, M., & Vazire, S. (2021). Eavesdropping on missing data: What are university students doing when they miss experience sampling reports? *Personality and Social Psychology Bulletin*, *47*(11), 1535–1549.

van Berkel, N., Goncalves, J., Hosio, S., Sarsenbayeva, Z., Velloso, E., & Kostakos, V. (2020). Overcoming compliance bias in self-report studies: A cross-study analysis. *International Journal of Human-Computer Studies*, *134*, 1–12.

van Berkel, N., Goncalves, J., Koval, P., Hosio, S., Dingler, T., Ferreira, D., & Kostakos, V. (2019). Context-informed scheduling and analysis: improving accuracy of mobile self-reports. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–12).

Wichers, M. C., Myin-Germeys, I., Jacobs, N., Peeters, F., Kenis, G., Derom, C., ... Van Os, J. (2007). Evidence that moment-to-moment variation in positive emotions buffer genetic risk for depression: a momentary assessment twin study. *Acta Psychiatrica Scandinavica*, *115*(6), 451–457.

Zeni, M., Zaihrayeu, I., & Giunchiglia, F. (2014). Multi-device activity logging. In *Proceedings of the 2014 acm international joint conference on pervasive and ubiquitous computing: Adjunct publication* (pp. 299–302).