Reliability-Aware Survey Platform with Behavioral Logs and LLM-based Evaluation

RE:MIND

Department of Data Science, Hanyang University / 2022055369 Yun Juchan & 2022004048 Kim Donghyeun

Abstract

Project Summary

•**Objective:** This project aims to resolve the "Reward Hunter" issue in incentive-based online surveys by developing a web platform that automatically assesses response reliability. The goal is to shift from simple respondent matching to a quality-assured data collection system.

•**Methodology:** We proposed a hybrid evaluation framework:

1.Quantitative Model (Q_score): An **XGBoost** classifier trained on the Mouse-Chase dataset to detect behavioral anomalies (e.g., mouse dynamics, straightlining).

2.Qualitative Model (L_score): An **LLM-based** module to evaluate the relevance and logic of open-ended text responses.

• **Achievement :** We successfully implemented a full-stack prototype integrating React, Spring Boot, and AI modules. The system calculates a unified reliability score (R) to automate reward distribution, effectively filtering careless responses and ensuring high-quality data.

Introduction & Background

1. The Problem: The Rise of "Reward Hunters" and C/IER

While financial incentives in online surveys effectively recruit participants, they introduce a critical side effect: the emergence of **"Reward Hunters"** who prioritize monetary gain over genuine engagement.

Careless Insufficient Effort Responding (C/IER): These participants exhibit specific low-quality behaviors, known as C/IER, to maximize their earnings per hour·

- **Observable Patterns:** Common manifestations include **Straightlining** (selecting the same option repeatedly to save effort) and **Speeding** (answering faster than cognitive processing allows).
- **Platform Gap:** Current commercial platforms focus primarily on **quantity matching** (connecting researchers with respondents) rather than **quality assurance**, leaving researchers vulnerable to noisy and unreliable data.

2. Limitations of Traditional Quality Control

Conventional methods used to filter low-quality responses suffer from significant structural weaknesses:

- **Vulnerability of Static Filters:** Simple mechanisms like **Attention Checks** (e.g., "Select 'Strongly Agree' to prove you are reading") are easily recognized and bypassed by experienced reward hunters who have learned to spot these "trap" questions.
- **Inflexibility of Single Metrics:** Rule-based metrics such as **Longstring** (checking for repeated answers) or response time thresholds are too rigid. They fail to detect **complex low-quality patterns**, such as users who randomly vary answers to evade straightlining detection or provide gibberish in open-ended text.

3. Proposed Solution: AI-Driven Multimodal Assessment

To overcome these limitations, we propose a **platform-level intervention** that automatically calculates a reliability score (R) for every individual response by integrating two distinct data modalities:

- **Quantitative Analysis (Paradata):** We utilize **Behavioral Logs**—including mouse trajectories, click dynamics, and page interaction times—as unconscious signals of respondent motivation, which are harder to fake than static answers.
- **Qualitative Analysis (Semantic):** We employ **LLMs** to evaluate the semantic relevance and logical consistency of open-ended responses, detecting "nonsensical" or "minimum-effort" text that traditional statistical filters miss.

Methodology

Methodology I: Quantitative Model (Q_score)

•**Dataset:** PsychArchives **Mouse-Chase Reading Dataset**.

•**Proxy Labeling:** Used experimental stimuli conditions as motivation labels (Standard = 0, Appeal/Warning = 1).

•**Input Features (Behavioral Logs):**

•**Careless Indicators:** IRV (Variability), Longstring, Mahalanobis Distance.

•**Mouse Dynamics:** Velocity (logvX), Trajectory (MRPM), Flips, Acceleration.

•**Time:** Page interaction time (log_page_time).

•**Model: XGBoost Classifier**

•Captures non-linear interactions between behavioral features (e.g., short time + straightlining).

•**Performance:** Recall 0.77 (Test set), effectively identifying high-motivation respondents.

Methodology II: Qualitative Model (L_score)

•**Model: GPT-4o (or o3) API** optimized for text evaluation.

•**Evaluation Criteria:**

•**Relevance:** Is the answer related to the question?

•**Informativeness:** Does it provide specific details?

•**Coherence:** Is the logic consistent?

•**Factuality (RAG):** Checks against external knowledge for fact-based surveys.

•**Process:** Prompts generate a normalized score (0–1) representing the "sincerity" of the text.

Methodology III: Score Fusion & Policy

•Hybrid Scoring Formula:

R = (0.5 x Q_score) + (0.5 X L_score)

•Combines behavioral (Q) and content (L) signals equally.

•**Decision Policy (Thresholds):**

•**Auto-Reward (R >= 0.70):** High reliability, immediate payment.

•**Manual Review (0.40 =< 0.70):** Ambiguous, requires admin check.

•**Reject (R < 0.40):** Low quality, payment withheld

Experimental Analysis

1. Quantitative Model Performance

- **High Sensitivity:** Achieved a Recall of 0.7749 (Test set), confirming the model is optimized to minimize false negatives and correctly identify diligent respondents.
- **Behavioral Proxies:** Feature importance analysis confirmed that Mouse Dynamics (logaY, MRPM) are strong predictors of motivation, outperforming linear baselines by capturing complex non-linear interactions.

2. Decision Policy Validation

- **Data-Driven Thresholds:** Established Auto-Reward (R >= 0.7) and Reject (R < 0.40) zones based on validation set confidence to ensure payout accuracy.
- **Hybrid Efficacy:** The fusion logic successfully caught ambiguous cases (e.g.High Behavior score but Low Content score), flagging them for manual review rather than erroneously rewarding them.

System Implementation & Service

System Process Overview

The platform operates through a streamlined four-stage pipeline:
Collection ➡ Parallel Evaluation ➡ Aggregation ➡ Deployment.

1. Data Collection & Setup

Researcher: Registers the survey and defines evaluation criteria (quantitative/qualitative). If criteria are omitted, the system automatically recommends optimal settings.

Participant: Accesses the survey via the **Web Interface**. Upon submission, response data is immediately routed to the backend pipeline.

2. Parallel Evaluation Pipeline

Incoming data is split into two independent processing engines:

Quantitative Logic (Python Engine):

Executes script-based and rule-based assessments using the internal **Knowledge Base (KB)**.

Performs vector embedding tasks to analyze behavioral logic and quantitative metrics.

Qualitative Logic (LLM Engine):

Applies **Context Engineering** to prepare text data.

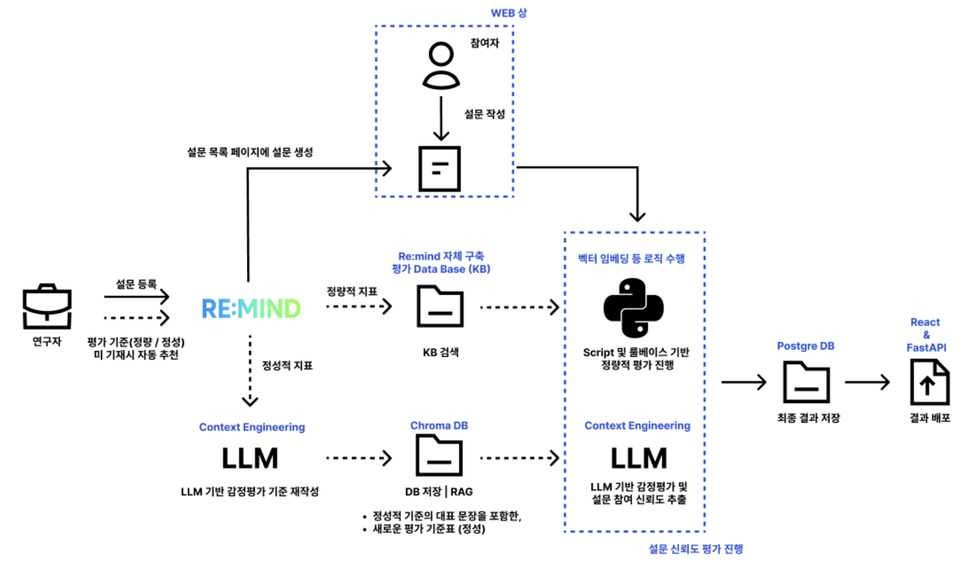
Utilizes **RAG (Retrieval-Augmented Generation)** with **Chroma DB** to retrieve relevant context and assess the semantic reliability of open-ended responses.

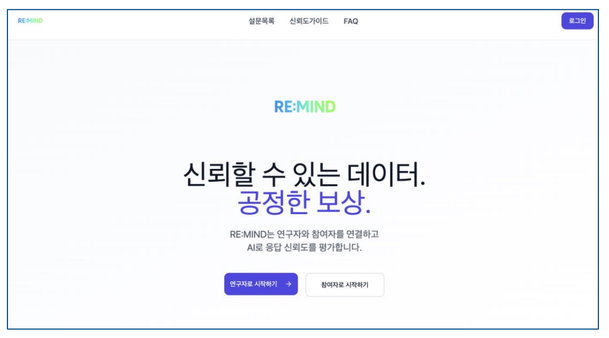
3. Data Aggregation & Storage

SQL DB: The outputs from both the quantitative script engine and the LLM engine are merged and securely stored as the final result dataset.

4. Result Deployment

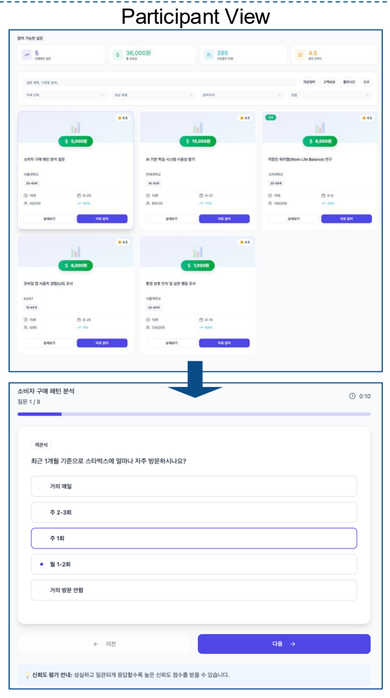
Visualization: The aggregated reliability reports are deployed via **FastAPI** to the **React Frontend**, providing researchers with a visualized dashboard of data quality.





Researcher View

Participant View



1. Participant View

•**Action:** Completes surveys via a dynamic web interface.

•**Process: Real-time, invisible logging** of behavioral paradata (mouse, time).

•**Result: Automated rewards** processed immediately based on the Reliability Score (R)

2. Researcher View

•**Setup:** Configures survey targets and evaluation criteria .

•**Monitor:** Visualizes data quality (Q, L, R) via **"Traffic Light" dashboard indicators**.

•**Control: One-click filtering** of low-quality responses (R < 0.40) before export .

Conclusion & Future Work

Conclusion

•**Paradigm Shift:** Developed a platform focused on **Quality Assurance**, shifting away from simple respondent matching.

•**Hybrid AI Efficacy:** Demonstrated that combining **Behavioral Logs (Q)** and **LLM Analysis (L)** outperforms traditional rule-based filters in detecting "Reward Hunters".

•**Automated Fairness:** Enabled a transparent, data-driven reward distribution system based on the unified reliability score (R).

Future Work

•**Real-World Validation:** Conduct pilot surveys to validate system usability and scoring accuracy with actual participants.

•**Model Optimization:** Retrain the quantitative model using large-scale commercial datasets to enhance robustness against noise.

•**Advanced Scoring:** Incorporate uncertainty metrics (confidence intervals) into the fusion logic for safer automated decisions.

Contact Information: Jun-Chan Yun (juchan563@gmail.com)
Academic Advisor: Prof. Hyoung-Sook Kim (khsook12@hanyang.ac.kr)