

000 001 002 003 004 005 *IntE: QUANTITATIVE FRAMEWORK FOR QUALITATIVE* 006 *DATA EVALUATION VIA DISTRIBUTIONAL MINING*

007
008
009
010
011 **Anonymous authors**
012 Paper under double-blind review

013 ABSTRACT

014 Evaluating the quality of qualitative datasets containing responses collected for
015 semi-structured questions is a persistent challenge. Manual analysis is slow and
016 subjective, while existing automated methods lack a holistic, dataset-level per-
017 spective crucial for mining insights. We introduce *IntE*, a novel framework for
018 the quantitative assessment of qualitative response datasets. *IntE* evaluates dataset
019 quality using the cluster distributions based on collected responses and the pre-
020 defined demographic distributions based on user metadata. *IntE* is structured into
021 a four-quadrant assessment that quantifies the potential of a dataset for revealing
022 general patterns and unique insights. The four quadrants rely on the distribu-
023 tions reconstructed via metadata and intra-data distances. Therefore, we propose
024 a content-aware multi-agent system that accurately computes inter-response dis-
025 similitude. This system features a two-stage adversarial framework for generating
026 domain-specific evaluation instructions and an adaptive anchor algorithm to en-
027 sure scoring consistency. We validate *IntE* through controlled experiments on
028 synthetic data, highlighting the effectiveness of its components. Additionally, a
029 real-world social survey case study, validated by domain experts, demonstrates
030 *IntE*’s capability to enhance knowledge discovery by accurately evaluating dataset
031 quality and identifying key responses for analysis.

032 1 INTRODUCTION

033 Semi-structured questions play a crucial role in studies involving human participants. They enable
034 researchers to gather detailed qualitative responses, uncovering *general patterns* and *unique insights*
035 across demographic groups (e.g., age, gender). For example, in a study on student well-being, a
036 question like “How have your eating habits changed since starting college?” might reveal general
037 patterns, such as freshmen reporting poorer diets, as well as unique insights from unexpected cases,
038 such as freshmen who are student-athletes maintaining healthy eating habits due to their training.

039 Quality evaluation of a qualitative response dataset is crucial before conducting analysis. Without
040 it, researchers risk wasting time analyzing low-quality responses and failing to discover both gen-
041 eral patterns and unique insights. However, existing quality evaluation methods are insufficient.
042 Traditional manual approaches are not scalable and lack the quantitative outputs needed to guide
043 data improvement (Mattimore et al., 2021; Campbell et al., 2013). Most automated methods eval-
044 uate individual responses in isolation (Chen et al., 2023b; Swayamdipta et al., 2020), overlook-
045 ing the inter-response relationships and overall dataset structure, which are crucial for the comparative
046 analysis required for knowledge discovery (Fayyad et al., 1996; Glaser & Strauss, 2017).

047 Rather than focusing on individual responses, our work aims to evaluate a response dataset as a
048 whole based on how much knowledge it contains for mining group-based patterns and insights.
049 We propose *IntE*, a new quantitative quality evaluation method that assesses a qualitative dataset
050 by comparing its internal structure to its known demographic composition. Intuitively, since the
051 demographic structure (e.g., the age or gender distributions of respondents) of a dataset is known,
052 the quality of the dataset can be assessed by analyzing its alignment with and divergence from that
053 structure. Specifically, a strong alignment may suggest the dataset confirms expected, general pat-
054 terns. On the other hand, a divergence between the two often signals the presence of new and unique
055 insights. Based on this, the core idea of *IntE* is to assess a dataset by quantitatively (1) measuring the
056 alignment and divergence between a demographic distribution based on user metadata and a cluster

054 distribution empirically derived from actual responses, and (2) examining the divergence within the
 055 cluster distribution itself. This dual approach allows us to identify general patterns within dense,
 056 homogeneous regions (e.g., high-density cluster centers), as well as to locate unique insights within
 057 abnormal responses (e.g., outliers and cross-demographic clusters).

058 More specifically, *IntE* consists of two components. First, it uses a **four-quadrant assessment**
 059 **framework** that operationalizes the comparison between the cluster and demographic distributions
 060 (Sec. 3.2). This framework yields four distinct metrics that characterize a dataset’s utility for dis-
 061 covering *general patterns* versus *unique insights* at both the *distribution* and *data-point* levels. The
 062 four quadrants rely on the distributions reconstructed via metadata and intra-response distances.
 063 Therefore, *IntE* integrates a **content-aware multi-agent system** to compute nuanced, inter-response
 064 dissimilarity (Sec. 3.1). To make the agent more reliable and domain-aware, this multi-agent sys-
 065 tem enhances the LLM-as-a-Judge paradigm (Zheng et al., 2023) with two key innovations. First,
 066 a novel interactive instruction generation system enables efficient, domain-specific instruction gen-
 067 eration (Sec. 3.1.2), moving beyond manual or black-box methods (Zhou et al., 2022; Shah, 2024;
 068 Pryzant et al., 2023; Chen et al., 2023a). Second, a dynamic anchor updating algorithm (Sec. 3.1.3)
 069 maintains a stable semantic reference frame to ensure evaluation consistency and mitigate scoring
 070 drift, improving upon static example-based approaches (Yang et al., 2024; Dong et al., 2024).

071 We validate *IntE* through a series of controlled experiments on synthetic data and a case study on a
 072 real-world dataset from economics research. The results demonstrate that *IntE*, the instruction gen-
 073 eration module, and the anchor updating algorithm are effective, robust, and capable of accelerating
 074 knowledge discovery in practice. In summary, our contributions are:

- 075 • We propose a four-dimensional quantitative assessment framework for data collected from
 076 semi-structured questions, based on comparing data-driven and demographic distributions
 077 to guide researchers on data quality and potential key data.
- 078 • We design a content-aware multi-agent system to extract inter-response dissimilarity, fea-
 079 turing a human-in-the-loop adversarial process for instruction generation and an adaptive
 080 anchor algorithm for evaluation consistency.
- 081 • We conduct controlled experiments via a synthetic data generation system and a real-world
 082 case study, which validate the effectiveness and practical utility of *IntE* and its components.

084 2 RELATED WORK

085 **Quality Evaluation of Qualitative Datasets.** Traditionally, assessing the quality of qualitative
 086 datasets, such as responses from semi-structured interviews, has been a manual process centered
 087 on data richness and thickness (Mattimoe et al., 2021; Campbell et al., 2013; Ames et al., 2024;
 088 Naeem et al., 2024; Johnson et al., 2020). Since the manual process is resource-intensive and time-
 089 consuming, early computational approaches have been proposed to automate the process. However,
 090 they often failed to capture the deep semantic nuances essential for qualitative insights (Manning &
 091 Schutze, 1999; Cambria & White, 2014; Chang et al., 2009). The recent advent of LLMs has enabled
 092 significant progress, with studies showing their capacity to achieve human-level performance in
 093 assessing individual text entries against predefined criteria (Parfenova et al., 2024; Barros et al.,
 094 2025; Gilardi et al., 2023; Chen et al., 2023b). However, this data-point-level focus (Smith et al.,
 095 2014; Swayamdipta et al., 2020; Kuan & Mueller, 2022) overlooks the holistic quality of the entire
 096 dataset, which is critical for assessing its potential for knowledge discovery (Fayyad et al., 1996;
 097 Zhang et al., 2024; Reis et al., 2024; Glaser & Strauss, 2017; Ghorbani & Zou, 2019). To address
 098 this limitation, *IntE* introduces a novel framework that holistically evaluates the quality and potential
 099 of a dataset through the intrinsic cluster distribution and extrinsic demographic distribution.

100 **LLM-based Automated Evaluation.** Our approach is built on the “LLM-as-a-Judge” paradigm,
 101 where LLMs serve as scalable proxies for human preference judgments (Zheng et al., 2023; Kim
 102 et al., 2023; Zhu et al., 2023; Dunivin, 2024). However, this paradigm still faces several challenges.
 103 First, LLMs are susceptible to biases, including positional and verbosity biases, which can com-
 104 promise evaluation objectivity (Ye et al., 2024; Li et al., 2024; Wang et al., 2023; Shi et al., 2024;
 105 Saito et al., 2023). Furthermore, ensuring consistent and reproducible judgments is difficult, as
 106 LLMs suffer from scoring drift and low self-consistency due to sensitivity to prompt phrasing (Lee
 107 et al., 2025; Zhou et al., 2024; Schroeder & Wood-Doughty, 2024). Finally, crafting high-quality,

108 domain-specific instructions for LLM judges also remains a labor-intensive task requiring significant
 109 expertise, which will limit rapid adaptation to new tasks (Khalid & Witmer, 2025; Raju et al., 2024).
 110 *IntE* devises a content-aware multi-agent system that enhances domain adaptability and consistency
 111 of evaluation via the interactive instruction generation and the adaptive anchor-update algorithm.
 112
 113

114 3 PROPOSED METHOD

115
 116 Figure 1 shows the overview of *IntE* with three steps. Specifically, the input of *IntE* is a **response**
 117 **dataset** (Fig. 1-a) collecting free-text responses to a question, where each response is associated
 118 with user demographic metadata. Then, *IntE* uses two components to assess the quality of the re-
 119 sponse dataset, including (1) a **four-quadrant assessment framework** (the third step) that compares
 120 between the cluster and demographic distributions to quantitatively evaluate data, and (2) a **content-**
 121 **aware multi-agent system** (the first two steps) that enhances the LLM-as-a-Judge paradigm to
 122 compute nuanced, inter-response dissimilarity to reconstruct distributions.
 123

124 In the first step, we use the **Two-stage Iterative Instruction Generation** (Fig. 1-b) system
 125 (Sec. 3.1.2) to quantify nuanced response relationships, starting with automated discovery of a uni-
 126 versal evaluation instruction, followed by context-specific refinement via human-in-the-loop.

127 Second, this refined instruction is used to compute a pairwise dis-
 128 similarity matrix for the responses. To ensure consistency across
 129 these calculations, especially given the context window limitations
 130 of LLMs, we employ an **Adaptive Anchor Manifold Maintenance**
 131 (Fig. 1-c) algorithm (Sec. 3.1.3). This algorithm dynamically man-
 132 ages a set of reference examples, or “anchors”, to provide a stable
 133 semantic frame for the LLM evaluator.

134 Finally, the **cluster distribution and demographic distribution**
 135 **are reconstructed** (Fig. 1-d). The dissimilarity matrix is used to
 136 partition the data via ensemble clustering, yielding a cluster dis-
 137 tribution. This is compared against the demographic distribution
 138 derived from user metadata (Sec. 3.2.2). *IntE* formulates a **four-quadrant assessment** (Fig. 1-e)
 139 based on the delta between these two distributions, yielding metrics that characterize the dataset
 140 in terms of *general patterns* versus *unique insights* at both the *distribution* and *data-point* levels
 (Sec. 3.2.2).

141 3.1 CONTENT-AWARE DISSIMILARITY EXTRACTION

142 3.1.1 PROBLEM FORMULATION

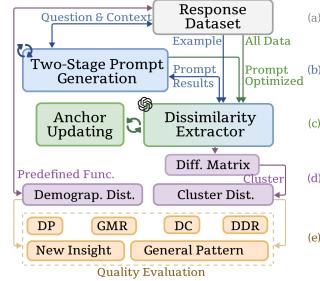
143 Given a response dataset $\mathcal{D} = \{d_i\}_{i=1}^N$ from a question with context \mathcal{C}_{xt} (e.g., topic, evaluation axis),
 144 our goal is to extract a consistent inter-response dissimilarity matrix \mathbf{D} .

145 This requires designing a system that can: (i) generate an optimal evaluation instruction P^* tailored
 146 to the context \mathcal{C}_{xt} , and (ii) use P^* to compute dissimilarities while maintaining global consistency.

147 Let \mathcal{P} be the space of all possible instructions. For any instruction $P \in \mathcal{P}$, an LLM-based evaluator
 148 E produces a dissimilarity score $S = E(P, d_i, d_j, \mathcal{C}_{\text{xt}})$. The quality of P is measured by its align-
 149 ment with a reference judgment from an Oracle O (a powerful LLM or human expert), providing a
 150 score S_O . The optimization objective is to find the instruction P^* that minimizes the expected loss:
 151

$$P^* = \arg \min_{P \in \mathcal{P}} \mathbb{E}_{(d_i, d_j) \sim \mathcal{D}} [\ell(E(P, d_i, d_j, \mathcal{C}_{\text{xt}}), O(d_i, d_j, \mathcal{C}_{\text{xt}}))] \quad (1)$$

152 where $\ell(\cdot, \cdot)$ is a suitable loss or pseudo-loss function. To solve this, we model instruction gen-
 153 eration as an optimization process using feedback as a pseudo-gradient (Sec. 3.1.2). To enforce
 154 consistency, we introduce an anchor manifold A , a set of reference responses that is dynamically
 155 updated (Sec. 3.1.3).



156 Figure 1: Overview for *IntE*.
 157 (b)
 158 (c)
 159 (d)
 160 (e)
 161

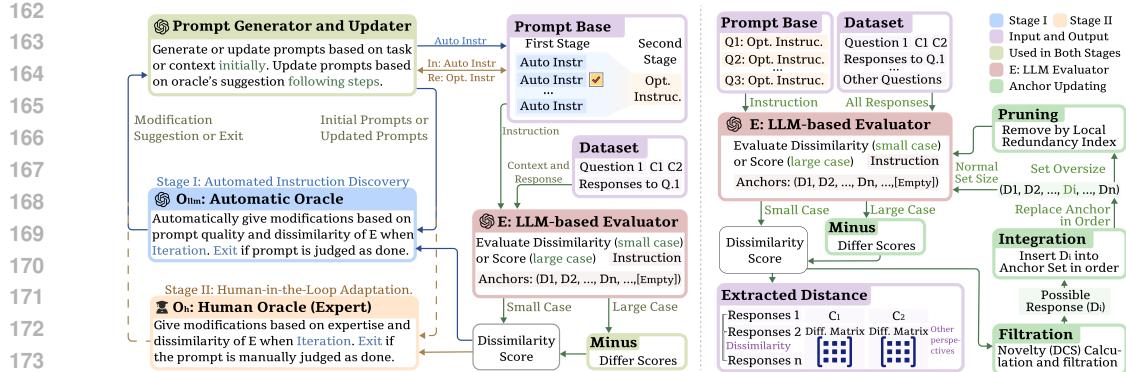


Figure 2: The Content-Aware Multi-agent System for Dissimilarity Extraction. It involves two parts: the iterative instruction generation part (Sec. 3.1.2) and the adaptive anchor manifold updating part (Sec. 3.1.3). The first part efficiently generates context-aware instructions for the second part to perform dissimilarity extraction.

3.1.2 ITERATIVE INSTRUCTION GENERATION

Manually authoring context-specific evaluation instructions is not scalable, as it may involve significant human labor. We propose a two-stage framework that models instruction optimization as an iterative refinement guided by feedback, which acts as a pseudo-gradient.

At each iteration t , an Oracle provides feedback on the current instruction $P^{(t)}$. This feedback, $\hat{\nabla}_{P^{(t)}} \mathcal{L}$, consists of a corrected score and a natural language critique, $(S_O^{(t)}, F_{\text{text}}^{(t)})$. An updater LLM, $\text{LLM}_{\text{updater}}$, uses this feedback to generate a refined instruction $P^{(t+1)}$:

$$P^{(t+1)} = \text{LLM}_{\text{updater}}(P^{(t)}, \hat{\nabla}_{P^{(t)}} \mathcal{L}) \quad (2)$$

This formulation treats the natural language critique as a pseudo-gradient, guiding the optimization in the high-dimensional space of instructions. The magnitude and direction of the “update” are determined by the content of the critique. This process unfolds in two stages:

Stage 1: Automated Instruction Discovery. The goal is to generate a diverse set of high-quality, general-purpose instructions. In this stage, the Oracle is a powerful LLM ($O \equiv O_{\text{LLM}}$). The iterative generation process runs automatically until a termination condition is met to build instructions from none. This condition is also determined by $O \equiv O_{\text{LLM}}$, additionally acting as a strict evaluator, which assesses the instruction against rigorous criteria for universality, clarity, and quality. The process halts when the instruction achieves a high score (e.g., greater than 0.9) and performance of LLM-based evaluator E with instruction P . This stage concludes with a human user selecting the most promising instruction P_a^* from the generated candidates.

Stage 2: Human-in-the-Loop Adaptation. The objective is to specialize the selected instruction P_a^* for a specific task. The Oracle is a human expert ($O \equiv O_H$). O_H provides precise, context-aware feedback, guiding the updater LLM to converge on a specialized instruction P^* . The loop terminates when the expert is satisfied with the evaluator’s performance and gives no more input.

3.1.3 ADAPTIVE ANCHOR MANIFOLD MAINTENANCE

To ensure consistent dissimilarity measurements, we maintain a sorted anchor manifold $A = \{a_1, \dots, a_k\}$, where each anchor a_j is a (response, score) pair and $k \leq k_{\max}$. It provides a stable semantic reference for the LLM. The algorithm iteratively processes new responses and updates the manifold to maximize semantic diversity while respecting the capacity constraint k_{\max} .

The update cycle involves three phases: **(1) Evaluation:** A new candidate response d is compared against other data from \mathcal{D} under the guidance of the anchor manifold to assess dissimilarity scores, which are then used to compare with the current anchors in A to determine its novelty. **(2) Filtration:** Candidates that offer the most novelty are selected. Novelty is measured by the Diversity Contribution Score (DCS), defined as the average dissimilarity to existing anchors: $\text{DCS}(d) = \frac{1}{k} \sum_{j=1}^k \delta(d, a_j)$. **(3) Integration and Pruning:** The selected novel candidates are integrated into the manifold. If $|A| > k_{\max}$, the most redundant anchor is pruned. Redundancy is

216 measured by the Local Redundancy Index, $\rho(a_j) = 1 - \delta(d_j, d_{j-1}) \cdot \delta(d_j, d_{j+1})$, which identifies
 217 anchors that are semantically interpolated by their neighbors.
 218

219 The implementation of the evaluation phase adapts to the dataset size N , but instructions for both
 220 approaches are generated through the same generation process (Sec. 3.1.2):
 221

- 222 • **For small datasets** ($N \leq N_{\text{threshold}}$): We use a high-fidelity approach where dissimilarities
 $\delta(d_i, d_j)$ are computed via direct pairwise LLM calls. This has a complexity of $\mathcal{O}(N^2)$.
 223
- 224 • **For large datasets** ($N > N_{\text{threshold}}$): To maintain computational feasibility, we use a
 225 linearized $\mathcal{O}(N)$ approach. A computationally efficient LLM operator assigns a scalar
 226 score $S(d)$ to each response. All dissimilarities are then approximated as the absolute
 227 difference between these scores: $\delta(d_i, d_j) \approx |S(d_i) - S(d_j)|$.
 228

229 3.2 DATA-DISTRIBUTION-DRIVEN QUALITATIVE DATASET ASSESSMENT

230 3.2.1 PROBLEM FORMULATION

231 We formalize the dataset quality assessment problem as one of comparing two different distributions
 232 of the data. Let the dataset be $\mathcal{D} = \{d_i\}_{i=1}^N$. Each data point d_i is associated with:
 233

- 234 1. An **extrinsic demographic label** $y_i \in \{1, \dots, C\}$, derived from user metadata \mathcal{I}_i via a
 235 mapping $y_i = f(\mathcal{I}_i)$. This defines the demographic partition $\mathcal{P} = \{C_j\}_{j=1}^C$, how the data
 236 distributes across different demographic partitions is called demographic distribution
- 237 2. An **intrinsic cluster label** $\hat{y}_i \in \{1, \dots, K\}$, obtained by applying an unsupervised clustering
 238 to a dissimilarity matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$. This defines the cluster partition $\hat{\mathcal{P}} = \{\hat{C}_k\}_{k=1}^K$,
 239 and how the data distributes across different cluster partitions is called cluster distribution.
 240

241 **Objective:** A researcher might aim to use data either for (1) finding common responses that
 242 represent each group of people, **general patterns**, or (2) discovering previously unknown **unique insights**.
 243 These analyses can be conducted at two granularities: the **distribution level** and the **data-point level**.
 244 Aligned with these practices, we seek to define a set of normalized metrics that quantify
 245 dataset quality by evaluating the concordance and divergence between the demographic partition \mathcal{P}
 246 and the cluster partition $\hat{\mathcal{P}}$, resulting in our four-quadrant assessment framework (Fig. 3).
 247

248 3.2.2 THE FOUR-QUADRANT ASSESSMENT FRAMEWORK

249 (1) Defining Demographic and Cluster Distributions

250 **Extrinsic Demographic Distribution:** This distribution is defined *a priori* from user metadata
 251 on characteristics researchers want to analyze. A researcher-defined function, $y_i = f(\mathcal{I}_i)$, maps
 252 attributes (e.g., expertise level) to a demographic label for each response d_i . The resulting vector
 253 $\mathbf{y} = \{y_i\}_{i=1}^N$ represents the ground-truth or expected structure of the dataset.
 254

255 **Intrinsic Cluster Distribution:** This distribution is derived *a posteriori* from the data's internal
 256 structure. A pairwise dissimilarity matrix \mathbf{D} is fed into an ensemble of clustering, where we use
 257 multiple k-means clusters and then vote for the final result (Ahmed et al., 2020). This yields a
 258 vector of cluster assignments $\hat{\mathbf{y}} = \{\hat{y}_i\}_{i=1}^N$. Since cluster labels are arbitrary, we align them with
 259 demographic labels by solving a maximum weight bipartite matching problem, where the weight
 260 between a demographic group C_j and a cluster \hat{C}_k is their Intersection-over-Union (Yu et al., 2016).
 261

262 (2) Preliminaries

263 Let \mathbf{y} and $\hat{\mathbf{y}}$ be the demographic and (aligned) cluster assignment vectors. Let $\mathbf{M} \in \mathbb{N}^{C \times C}$ be
 264 the confusion matrix where \mathbf{M}_{jk} is the count of data points from demographic j in cluster k . Let
 265 $n_j = |C_j|$ and $\hat{n}_k = |\hat{C}_k|$ be the sizes of demographic j and cluster k . Let \mathbf{D} be the dissimilarity
 266 matrix. We define:
 267

- 268 • **Average Intra-Cluster Dissimilarity** ($\bar{\delta}_{\text{intra}}$): The mean dissimilarity between pairs of
 269 data points within the same cluster.
- 270 • **Average Inter-Cluster Dissimilarity** ($\bar{\delta}_{\text{inter}}$): The mean dissimilarity between pairs of data
 271 points in different clusters.

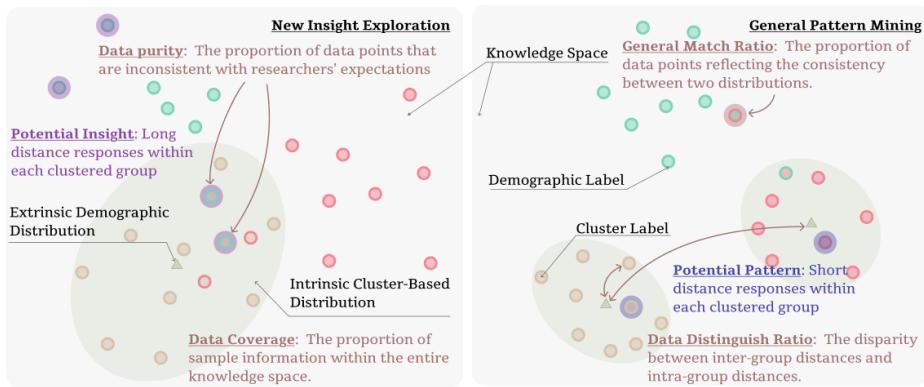


Figure 3: The Four-Quadrant Assessment Framework, which evaluates a dataset based on the research goal (General Patterns/Unique Insights) and the analysis level (Distribution/Data Point).

(3) Evaluation Metrics

General Pattern / Data Point: General Match Ratio (GMR). This metric assesses the global alignment between the demographic and cluster partitions. A high GMR indicates that the dataset’s intrinsic structure aligns with its known macroscopic properties, *i.e.*, data exactly reflects behaviors of humans from different groups and is easier to find representative responses for each group.

$$GMR = \max \left(0, 1 - \frac{1}{\sqrt{C}} \left\| \alpha \cdot \frac{n_j - \mathbf{M}_{jj}}{n_j} + \frac{|n_j - \hat{n}_j|}{n_j} \right\|_2 \right) \quad (3)$$

α is a parameter to adjust the importance of labeling errors over quantity difference. We recommend it between 1 (imbalanced dataset, emphasis unequalizing) and 2 (balanced dataset, emphasis errors).

General Pattern / Distribution: Data Distinguishability Ratio (DDR). It evaluates the quality of the cluster partition by measuring the separation between clusters relative to their internal compactness. High DDR shows a well-structured partitioning where clusters are coherent and distinct, *i.e.*, responses collected from different groups distinctly represent the common feature of that group.

$$DDR = \min \left(1, \beta \cdot \frac{\bar{\delta}_{\text{inter}}}{\bar{\delta}_{\text{intra}}} \right) \quad (4)$$

where β is a scaling hyperparameter. We suggest setting $\beta \approx (\text{Scale} - 2\sigma)/\sigma$ with $\sigma \approx \text{Scale}/C$, reflecting the ideal ratio of maximum inter-cluster to average intra-cluster distance.

Unique Insight / Data Point: Data Purity (DP). It evaluates the consistency of the emergent clusters by calculating the average proportion of members that belong to the dominant demographic in each cluster. Low DP indicates that some responses containing insights that are infrequently shown in the group of interviewees like cross-group consensus or voices of underrepresented groups.

$$DP = \min \left(1, \eta \cdot \left(1 - \frac{1}{K} \sum_{k=1}^K \frac{|\hat{C}_k| - \max_j |\hat{C}_k \cap C_j|}{|\hat{C}_k|} \right) \right) \quad (5)$$

where K is the number of clusters, and $\eta = 1$ is a scaling hyperparameter, as the core term is already a percentage-based purity metric.

Unique Insight / Distribution: Data Coverage (DC). This metric evaluates the conceptual dispersion of each cluster, quantifying how widely individual data points spread around the central theme of the cluster. A high DC indicates greater intra-cluster dissimilarity relative to the baseline threshold, emphasizing clusters with a broader spread, and may contain unique insights.

$$DC = \max \left(0, \frac{\bar{\delta}_{\text{intra}}}{\gamma} \right) \quad (6)$$

324 where γ is a scaling hyperparameter representing a baseline for acceptable intra-cluster dissimilarity,
 325 set relative to the typical dissimilarity $\sigma \approx \text{Scale}/C$. The adjustment of γ (e.g., 1.2σ for small
 326 datasets, 0.8σ for large) accounts for variations in dissimilarity estimation methods.
 327

328 **Response Mining.** Our framework also enables the identification of specific responses for qual-
 329 itative analysis. These responses, which either challenge expected structures or strongly represent
 330 general patterns, are characterized as the top- k responses with the highest and lowest average dis-
 331 similarity $\bar{d}(d_i)$, marking them as the most central or outlier members within their clusters.
 332

333 4 EXPERIMENT

335 We evaluate *IntE* through controlled experiments on synthetic data (Sec. 4.1) and a real-world case
 336 (Sec. 4.2). All experiments use “Qwen-max-latest” unless specified otherwise (Yang et al., 2025).
 337

338 Controlled experiments include: 1) a user study with 20 participants to evaluate the interactive
 339 instruction generation system (Sec. 4.1.1); 2) an ablation and comparison study to validate the anchor
 340 updating algorithm (Sec. 4.1.2) and statical performance of instruction generation system; and 3) a
 341 parameter sweep study to prove *IntE* follows well under varying data quality (Sec. 4.1.3).
 342

343 The case study (Sec. 4.2) applies *IntE* to qualitative responses from social science research on
 344 human behavior changing under supervision of an expert workshop, demonstrating the real-world
 345 applicability and the automatic response mining function of *IntE*.
 346

4.1 CONTROLLED EXPERIMENT WITH SYNTHETIC DATA

347 All datasets in 4 domains (HCI, Finance, Additive Manufacturing, and Post-operative Medicine)
 348 used in these controlled experiments are produced by our Controllable Synthetic Data Generation
 349 System B.1, ensuring that ground-truth attributes are known for precise evaluation.
 350

351 4.1.1 CONTROLLED USER STUDY FOR INTERACTIVE INSTRUCTION GENERATION SYSTEM

353 We conducted a controlled, within-subjects user study to assess our instruction generation system’s
 354 impact on user cognitive load and efficiency.
 355

356 **Experimental Setup.** We recruited 20 postgraduate students (11M/9F, aged 21-31) from diverse do-
 357 mains. Participants self-rated their domain experience ($M = 3.50$, $SD = 0.83$) and familiarity with
 358 instruction generation ($M = 3.15$, $SD = 1.27$) on a 5-point Likert scale. The system implements
 359 the adaptation phase of Sec. 3.1.2. The baseline condition utilized the same UI but deactivated the
 360 agent-assisted features, requiring users to compose instructions manually. Participants were tasked
 361 with generating instructions for dissimilarity extraction for two distinct questions, using both ours
 362 and the baseline. Each session lasted approximately 45 minutes and was compensated with 8 USD.
 363

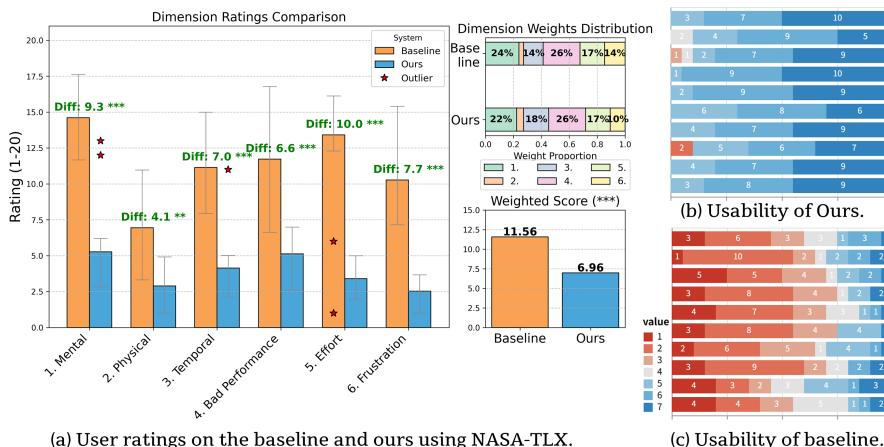


Figure 4: Evaluation results from the user study: (a) cognitive load (higher scores = higher load);
 (b)–(c) system usability (7 = good usability).

We evaluated both systems on *Cognitive Load* (NASA-TLX) (Hart & Staveland, 1988) and *System Usability*. Usability was measured with a custom 10-item questionnaire on a 7-point Likert scale, assessing: Initial-Prompt Quality (Q1), Effortless-Tuning (Q2), Focus-on-Intent (Q3), No-Prompt-Expertise-Needed (Q4), Rapid-Iteration (Q5), Self-Confidence (Q6), High-Satisfaction (Q7), Creative-Insights (Q8), Easy-Learning (Q9), and Will-to-Reuse (Q10).

Results and Analysis. As shown in Fig. 4, our framework significantly reduces user cognitive load when composing instructions. It also demonstrates superior performance across all evaluated usability dimensions. Instructions collected from this study were used in subsequent experiments.

4.1.2 ABLATION AND COMPARISON STUDY

We performed an ablation study to validate the constituent components of our adaptive anchor updating algorithm (Sec. 3.1.3) and conducted comparison experiments to assess the effectiveness of the interactive instruction generation framework (Sec. 3.1.2).

Experimental Setting. On four 15-interviewee dataset from different domains(generated via Sec. B.1), we evaluated several ablations of our full algorithm (**Full**): disabling in-context learning (**No Example**), disabling the algorithm (**Fixed Example**), including all responses as candidates (**No Novelty Check**), removing example length limits (**No Pruning**), and using a heuristic for score calculation (**Heuristic Quality**), *i.e.*, score is the average of samples with the smallest difference.

The framework was also tested with instructions generated under three conditions, including small scale on 15, large scale on 15 and large scale on 90-interviewee datasets (generated via Sec. B.1). Experiments included several LLMs: Qwen-max-latest, Qwen-plus-latest, Qwen-turbo-latest, GLM-4.5, and deepseek-v3 (Yang et al., 2025; Zeng et al., 2025; Liu et al., 2024). We measured Pearson and Spearman correlation between extracted and ground-truth dissimilarity scores. Due to the extensive number of experiments, each reported value is the average performance across four randomly selected questions from the four domains.

Table 1: Result of the ablation study and comparison study. **Bold and underlined** represents top 1 result and **Bold** represents top 2 result in ablation study. “Direct” is the manual version and “scale up” is the result performed on the 90-dataset.

	Ablation Study on Small Scale (Comparison) Version														
	Direct		Baseline		No example		Fix Example		No Novelty Check		No Pruning		Heuristic Quality		
	Spear.	Pears.	Spear.	Pears.	Spear.	Pears.	Spear.	Pears.	Spear.	Pears.	Spear.	Pears.	Spear.	Pears.	
Qwen_turbo	-0.083	-0.035	<u>0.708</u>	<u>0.681</u>	0.644	0.656	0.659	0.663	0.604	0.625	0.620	0.604	0.606	0.640	
Qwen_plus	0.228	0.292	<u>0.725</u>	<u>0.718</u>	0.674	0.659	<u>0.760</u>	<u>0.757</u>	0.700	0.683	0.652	0.667	0.643	0.636	
Qwen_max	-0.036	-0.013	<u>0.725</u>	<u>0.728</u>	0.633	0.621	0.672	0.663	0.683	0.695	0.691	0.697	0.701	0.709	
GLM	0.599	0.621	<u>0.830</u>	<u>0.836</u>	0.770	0.765	0.822	0.829	0.826	0.839	<u>0.831</u>	<u>0.842</u>	0.729	0.732	
deepseek-v3	0.331	0.356	<u>0.726</u>	0.705	0.674	0.685	0.717	<u>0.722</u>	0.697	0.673	0.703	0.676	<u>0.739</u>	<u>0.734</u>	
Scale up on Large Dataset															
	Ablation Study on Large Scale (Direct Assigning) Version		Baseline		No example		Fix Example		No Novelty Check		No Pruning		Heuristic Quality		
	Spear.	Pears.	Spear.	Pears.	Spear.	Pears.	Spear.	Pears.	Spear.	Pears.	Spear.	Pears.	Spear.	Pears.	
	Qwen_turbo	0.789	0.805	<u>0.747</u>	<u>0.775</u>	0.705	0.733	0.692	0.721	0.699	0.730	0.729	0.765	0.709	0.732
Qwen_plus	0.813	0.824	0.719	0.741	<u>0.796</u>	<u>0.810</u>	0.735	0.756	0.757	0.777	0.762	0.780	<u>0.767</u>	<u>0.790</u>	
Qwen_max	0.786	0.808	<u>0.789</u>	<u>0.816</u>	0.773	0.797	0.768	0.794	0.767	0.791	0.776	0.803	0.752	0.779	
GLM	0.844	0.865	<u>0.797</u>	<u>0.819</u>	0.771	0.785	0.758	0.786	0.792	0.817	0.780	0.803	0.794	0.818	
deepseek-v3	0.824	0.840	<u>0.827</u>	<u>0.852</u>	0.794	0.822	<u>0.856</u>	<u>0.878</u>	0.822	0.846	0.826	0.850	0.811	0.839	

Results and Analysis. As shown in Tab. 4.1.2, the full algorithm and its components are effective. The instruction generation system provides a statistically significant improvement in instruction quality over manual generation. The anchor updating algorithm maintains high performance as dataset size increases. While the large-scale version yielded superior results, we posit that the small-scale version may better model according to the following discussion.

Discussion on Large-Scale Approximation: 1. While large-scale relevance is enhanced, scalar projection simplifies high-dimensional semantic relationships into a single value, causing information loss. But this trade-off is essential for achieving efficiency. 2. This approach is applicable only when responses can be reduced to a score, such as cases influenced by personal literacy levels. Otherwise, the small-scale version can only be chosen.

432
4334.1.3 PARAMETER SWEEP EXPERIMENT FOR *IntE* EVALUATION434
435

We evaluated the overall performance of *IntE* by applying it to synthetic datasets with controlled distributional variations to validate that its behavior aligns with our design principles.

436
437
438
439
440
441

Experimental Setting. We generated two sets of 15-interviewee datasets (per Sec. B.1). In the first set (Fixed Mean), the inter-community mean difference was fixed while variance was swept from 1 to 40. In the second set (Fixed Variance), variance was fixed at 20 while the inter-community mean difference was swept from 5 down to 45. We applied *IntE* in both small-scale and large-scale configurations to observe how its four evaluation scores responded.

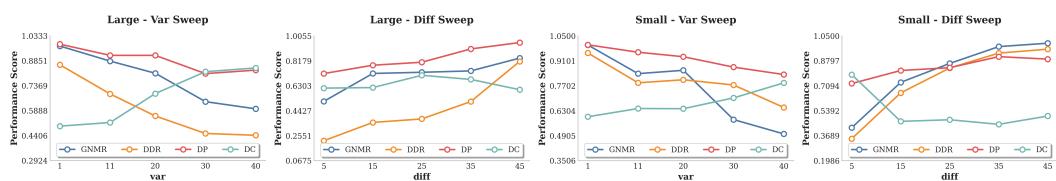
442
443
444
445
446
447448
449

Figure 5: Result of parameter sweep experiment.

450
451
452
453
454

Results and Analysis. Shown in Fig. 5, with increasing variance, the GNMR, DP, and DDR scores decreased, while the DC score increased. Conversely, as the mean difference decreased, GNMR, DP, and DDR scores increased, while the DC score remained stable. These trends are consistent with our design, validating the behavioral correctness of *IntE*’s scoring functions.

455
456

4.2 CASE STUDY ON STUDENT FOOD CHOICES

457
458
459
460
461
462

To demonstrate real-world applicability, we conducted a case study using a dataset of 126 qualitative responses on college students’ food preferences ¹. The task, involving three domain experts, was to relate student grade level (37 freshmen, 32 sophomores, 26 juniors, 27 seniors) to self-reported changes in eating habits since starting university. The experts used our instruction generation system (Sec. 3.1.2) to create evaluation instructions for *IntE*, aiming to assess data quality and automatically identify key responses with general patterns or unique insights.

463
464
465
466
467
468
469

Following the recommended parameters in Sec. 3.2.2 ($\alpha = 1.0, \beta = 1.5, \eta = 25, \delta = 1.0$), the resulting metrics ($GMR = 0.39, DDR = 0.03, DC = 0.90, DP = 0.43$) faithfully represented the data’s quality. The analysis revealed a similar distribution of responses across all grade levels, “getting worse” ($mean = 61.69\%, std = 3.87\%$), “keeping the same” ($mean = 6.33\%, std = 3.61\%$), and “getting better” ($mean = 29.74\%, std = 4.89\%$). This uniformity between groups explains the low *DDR* and *GMR* scores. Meanwhile, a high mean across demographic groups and uncleaned raw data led to high *DC* and low *DP* values.

470
471
472
473

The experts reviewed the responses given by the mining function in our system, including the top-3 most general and top-3 most unique points. They confirmed the effectiveness of *IntE* in surfacing these data.

474
475

5 CONCLUSION

476

In this paper, we introduced *IntE*, a novel framework to quantitatively evaluate qualitative datasets by measuring the divergence and alignment between an extrinsic demographic distribution and an intrinsic cluster distribution. *IntE* holistically assesses a dataset’s potential for yielding both general patterns and unique insights, enabled by a four-quadrant assessment framework and a content-aware multi-agent system that computes robust dissimilarity scores using interactive instruction generation and adaptive anchors. Our empirical evaluation, through controlled experiments and a real-world case study, validated *IntE*. The results demonstrate that *IntE* effectively assesses dataset quality and accelerates knowledge discovery by automatically surfacing high-value responses for analysis.

484
485

¹<https://www.kaggle.com/datasets/borapajo/food-choices>

486 REFERENCES
487

488 Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. The k-means algorithm: A
489 comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020.

490 Heather MR Ames, Emma F France, Sara Cooper, Mayara S Bianchim, Simon Lewin, Bey-Marrié
491 Schmidt, Isabelle Uny, and Jane Noyes. Assessing qualitative data richness and thickness: devel-
492 opment of an evidence-based tool for use in qualitative evidence synthesis. *Cochrane Evidence
493 Synthesis and Methods*, 2(7):e12059, 2024.

494 Cauã Ferreira Barros, Bruna Borges Azevedo, Valdemar Vicente Graciano Neto, Mohamad Kassab,
495 Marcos Kalinowski, Hugo Alexandre D Do Nascimento, and Michelle CGSP Bandeira. Large
496 language model for qualitative research: A systematic mapping study. In *2025 IEEE/ACM In-
497 ternational Workshop on Methodological Issues with Empirical Studies in Software Engineering
498 (WSESE)*, pp. 48–55. IEEE, 2025.

499 Erik Cambria and Bebo White. Jumping nlp curves: A review of natural language processing
500 research. *IEEE Computational intelligence magazine*, 9(2):48–57, 2014.

501 John L Campbell, Charles Quincy, Jordan Osserman, and Ove K Pedersen. Coding in-depth
502 semistructured interviews: Problems of unitization and intercoder reliability and agreement. *So-
503 ciological methods & research*, 42(3):294–320, 2013.

504 Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. Reading tea
505 leaves: How humans interpret topic models. *Advances in neural information processing systems*,
506 22, 2009.

507 Lichang Chen, Juhai Chen, Tom Goldstein, Heng Huang, and Tianyi Zhou. Instructzero: Efficient
508 instruction optimization for black-box large language models, 2023a. URL <https://arxiv.org/abs/2306.03082>.

509 Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Rui Feng Xu. Exploring the use of large
510 language models for reference-free text quality evaluation: An empirical study, 2023b. URL
511 <https://arxiv.org/abs/2304.00723>.

512 Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu,
513 Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-
514 context learning, 2024. URL <https://arxiv.org/abs/2301.00234>.

515 Zackary Okun Dunivin. Scalable qualitative coding with llms: Chain-of-thought reasoning matches
516 human performance in some hermeneutic tasks. *arXiv preprint arXiv:2401.15170*, 2024.

517 Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge
518 discovery in databases. *AI magazine*, 17(3):37–37, 1996.

519 Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning.
520 In *International conference on machine learning*, pp. 2242–2251. PMLR, 2019.

521 Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-
522 annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

523 Barney Glaser and Anselm Strauss. *Discovery of grounded theory: Strategies for qualitative re-
524 search*. Routledge, 2017.

525 Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of
526 empirical and theoretical research. In *Advances in psychology*, volume 52, pp. 139–183. Elsevier,
527 1988.

528 Jessica L Johnson, Donna Adkins, and Sheila Chauvin. A review of the quality indicators of rigor
529 in qualitative research. *American journal of pharmaceutical education*, 84(1):7120, 2020.

530 Muhammad Talal Khalid and Ann-Perry Witmer. Prompt engineering for large language model-
531 assisted inductive thematic analysis. *arXiv preprint arXiv:2503.22978*, 2025.

540 Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun,
 541 Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evalua-
 542 tion capability in language models. In *The Twelfth International Conference on Learning Repre-
 543 sentations*, 2023.

544 Johnson Kuan and Jonas Mueller. Model-agnostic label quality scoring to detect real-world label
 545 errors. In *ICML DataPerf Workshop*, 2022.

546 Noah Lee, Jiwoo Hong, and James Thorne. Evaluating the consistency of llm evaluators. In *Pro-
 547 ceedings of the 31st International Conference on Computational Linguistics*, pp. 10650–10659,
 548 2025.

549 Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun
 550 Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint
 551 arXiv:2412.05579*, 2024.

552 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,
 553 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint
 554 arXiv:2412.19437*, 2024.

555 Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*.
 556 MIT press, 1999.

557 Ruth Mattimoe, Michael Hayden, Brid Murphy, and Joan Ballantine. Approaches to analysis of
 558 qualitative research data: A reflection on the manual and technological approaches. *Accounting,
 559 Finance & Governance Review*, 27(1):1–16, 2021.

560 Muhammad Naeem, Wilson Ozuem, Kerry Howell, and Silvia Ranfagni. Demystification and actu-
 561 alisation of data saturation in qualitative research through thematic analysis. *International Journal
 562 of Qualitative Methods*, 23:16094069241229777, 2024.

563 Angelina Parfenova, Alexander Denzler, and Jörgen Pfeffer. Automating qualitative data analysis
 564 with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for
 565 Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 83–91, 2024.

566 Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt
 567 optimization with “gradient descent” and beam search. *arXiv preprint arXiv:2305.03495*, 2023.

568 Ravi Raju, Swayambhoo Jain, Bo Li, Jonathan Li, and Urmish Thakker. Constructing domain-
 569 specific evaluation sets for llm-as-a-judge. *arXiv preprint arXiv:2408.08808*, 2024.

570 Eduardo Reis, Mohamed Abdelaal, and Carsten Binnig. Generalizable data cleaning of tabular data
 571 in latent space. *Proceedings of the VLDB Endowment*, 17(13):4786–4798, 2024.

572 Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference
 573 labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

574 Kayla Schroeder and Zach Wood-Doughty. Can you trust llm judgments? reliability of llm-as-a-
 575 judge. *arXiv preprint arXiv:2412.12509*, 2024.

576 Chirag Shah. From prompt engineering to prompt science with human in the loop. *arXiv preprint
 577 arXiv:2401.04122*, 2024.

578 Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A
 579 systematic investigation of position bias in pairwise comparative assessments by llms. 2024.

580 Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. An instance level analysis of data
 581 complexity. *Machine learning*, 95(2):225–256, 2014.

582 Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi,
 583 Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with
 584 training dynamics. *arXiv preprint arXiv:2009.10795*, 2020.

594 Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu,
 595 Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint*
 596 *arXiv:2305.17926*, 2023.

597

598 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu,
 599 Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint*
 600 *arXiv:2505.09388*, 2025.

601

602 Chenguang Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun
 603 Chen. Large language models as optimizers, 2024. URL <https://arxiv.org/abs/2309.03409>.

604

605 Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner
 606 Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-
 607 judge. *arXiv preprint arXiv:2410.02736*, 2024.

608

609 Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An ad-
 610 vanced object detection network. In *Proceedings of the 24th ACM international conference on*
Multimedia, pp. 516–520, 2016.

611

612 Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang,
 613 Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation
 614 models. *arXiv preprint arXiv:2508.06471*, 2025.

615

616 Yunyi Zhang, Ming Zhong, Siru Ouyang, Yizhu Jiao, Sizhe Zhou, Linyi Ding, and Jiawei Han.
 617 Automated mining of structured knowledge from text in the era of large language models. In
 618 *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*,
 pp. 6644–6654, 2024.

619

620 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
 621 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and
 622 chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

623

624 Wenjie Zhou, Qiang Wang, Mingzhou Xu, Ming Chen, and Xiangyu Duan. Revisiting the self-
 625 consistency challenges in multi-choice question formats for large language model evaluation. In
 626 *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*
Resources and Evaluation (LREC-COLING 2024), pp. 14103–14110, 2024.

627

628 Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and
 629 Jimmy Ba. Large language models are human-level prompt engineers. In *The eleventh interna-*
630 tional conference on learning representations, 2022.

631

632 Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models
 633 are scalable judges. *arXiv preprint arXiv:2310.17631*, 2023.

634

A USE OF LLM

635

636 In this paper, we used Gemini to check grammar and improve wording. It did not change the original
 637 meaning of the text or introduce any new references or knowledge.

638

639 We also used it to search for related work. All retrieved papers were read and reviewed by the
 640 authors, who manually decided whether to include them.

641

B CONTROLLABLE SYNTHETIC DATA GENERATION

642

643 We use the following controllable synthetic data generation system B.1 to controllably generate
 644 data for experiments, ensuring that ground-truth attributes are known for precise evaluation. In the
 645 experiment, we generated 4 datasets from different domains, including finance, medical, human-
 646 computer interaction, and crowd-sourcing for supervised finetuning data collection. The questions
 647 are listed below.

648 B.1 CONTROLLABLE SYNTHETIC DATA GENERATION SYSTEM
649650 We generate synthetic data with precise ground-truth attributes using a modular, four-stage pipeline
651 of LLM-based agents.652 **Community Definition.** Given a questionnaire Q , an Architect Agent defines k evaluation dimensions
653 D and m user communities C (e.g., “expert”, “novice”). For each community c_j , it generates
654 a qualitative profile P_j describing its background and behaviors.655 **Quantitative Score Generation.** For each synthetic user i in a community c_j , we generate a ground-
656 truth score vector $s_i \in \mathbb{R}^k$. Each score s_i, k for dimension d_k is sampled from a specified distribution
657 (e.g., $\mathcal{N}(\mu_{j,k}, \sigma_{j,k}^2)$), where the parameters $(\mu_{j,k}, \sigma_{j,k})$ are defined per community to control
658 attribute levels and diversity.659 **Persona Instantiation.** A Persona Agent creates a narrative persona p_i conditioned on the community
660 profile P_j and score vector s_i . The persona’s background (e.g., education, career) is generated
661 to be consistent with the assigned scores: $p_i = \text{AgentPersona}(P_j, s_i)$.662 **Response Simulation.** An Interviewee Agent simulates the user’s responses R_i to the questionnaire
663 Q . The generation is conditioned on the persona p_i and score vector s_i , ensuring the responses
664 reflect the ground-truth scores. The score vector s_i is prioritized to ensure quantitative control:
665 $R_i = \text{AgentInterviewee}(p_i, s_i, Q)$.666 For our experiments (Sec. 4.1), we applied this system to generate response sets for four question-
667 nares (HCI, Finance, Additive Manufacturing, and Post-operative Medicine) with quantitatively
668 controlled attribute distributions.669 B.2 QUESTIONNAIRE
670671 We provide the Sim-structured questionnaire in different domains here, which are used for dataset
672 generation in the controlled experiment.673 B.2.1 FINANCE QUESTIONNAIRE
674675 **Questionnaire Name:** Finance676 **Questions:**677 1. **Questions related to banking services**678

- **Financial behavior:** How often do you engage in banking services? What types of
679 services do you typically use at the bank?
- **Financial literacy:** What risks do you think are associated with keeping money in the
680 bank?

681 2. **Concerns about financial status**682

- **Financial behavior:** Are you concerned about your financial status (e.g., frequently
683 checking your bank account and income/expenditure status)?
- **Financial attitude:** How important do you think it is to monitor your financial status
684 and why?

685 3. **Questions related to financial education**686

- **Financial literacy:** What activities in your daily life do you think are related to fi-
687 nance?
- **Financial attitude:** Do you think financial education in schools is important?

688 4. **Questions about saving habits**689

- **Financial behavior:** Do you have good saving habits?
- **Financial literacy:** Do you know what methods or channels are available for saving
690 money?
- **Financial attitude:** How important do you think saving money is?

691 5. **Perception of investment risks**

702

- 703 • **Financial literacy:** What do you think about the statement “high-yield investments
- 704 come with high risks?” Generally, how does the risk of investment change when
- 705 investors diversify their investments across different categories?
- 706 • **Financial behavior:** Do you pay attention to major financial news, such as stock mar-
- 707 ket fluctuations? If you participate in investments, would you take actions to diversify
- 708 your investments across different categories?

709 **B.2.2 MEDICAL QUESTIONNAIRE**

710 **Questionnaire Name:** Analysis of Health Post-Treatment with New Bone Medication

711 **Questions:**

712 1. **Questions about new or unexpected health changes**

713

- 714 • **Symptom discovery:** Thinking back to the time after your bone started feeling better,
- 715 what was the very first new or unexpected health change you noticed that was not
- 716 related to your original injury? Please describe what it was and what it felt like.
- 717 • **Symptom inventory:** Besides that first change, what other new health concerns have
- 718 you experienced since taking the new medication? Please list them and briefly de-
- 719 scribe each one (e.g., skin rashes, constant tiredness, digestive issues, headaches,
- 720 mood changes).

721 2. **Questions about the timing and patterns of your new condition**

722

- 723 • **Timing and onset:** For the main new health issue you mentioned, can you pinpoint
- 724 when it started in relation to when you were taking the new medicine? (e.g., “It started
- 725 a week after I began the medication,” or “It didn’t show up until a month after I finished
- 726 the full course”).
- 727 • **Pattern and triggers:** Have you noticed any patterns to this new condition? For
- 728 example, is it constant, or does it come and go? Is there anything that seems to make
- 729 it better or worse (like certain foods, activities, stress, or time of day)?

730 3. **Questions about the impact of the new condition on your life**

731

- 732 • **Impact comparison:** Think about the challenges of the original broken bone versus
- 733 the challenges of this new health condition. Which one has had a bigger impact on
- 734 your day-to-day life, and why?
- 735 • **Life adjustments:** What is one specific activity or part of your daily routine that
- 736 you’ve had to change or give up, not because of your bone, but because of this new
- 737 health condition?

738 4. **Questions about your own thoughts and actions regarding the new condition**

739

- 740 • **Personal hypothesis:** Before this questionnaire, had you already made a connection
- 741 in your own mind between the new medicine and your new health condition? What
- 742 made you suspect (or not suspect) a link?
- 743 • **Communication with doctors:** Have you discussed this new health issue with a doc-
- 744 tor before? If so, what was that conversation like? If not, what has held you back from
- 745 bringing it up?

746 5. **Questions about your overall perspective on the treatment**

747

- 748 • **Future decision-making:** How has this experience changed how you will approach
- 749 taking a new or experimental medication in the future? What questions would you ask
- 750 your doctor now that you might not have asked before?
- 751 • **Defining successful treatment:** This medicine successfully healed your bone, but
- 752 may have caused other issues. How does this experience change your personal defini-
- 753 tion of a “successful” medical treatment?

754 **B.2.3 HUMAN-COMPUTER INTERACTION QUESTIONNAIRE**

755 **Questionnaire Name:** Human-Computer Interaction (HCI)

756 **Questions:**

756 1. **AI System Interaction Experience Questions**
 757 • **AI Response Expectations:** Does the AI system's response meet your expectations?
 758 • **AI Understanding Accuracy:** Do you feel that the AI understands your needs accurately when using the AI system?
 759 • **Desired AI Features:** What features would you like the AI system to add to enhance user experience?
 760
 761 2. **Input Device Usage Experience Questions**
 762 • **User Preferences:** Does the current input device meet your operating habits?
 763 • **Usability Challenges:** Have you encountered any inconveniences while using the input device?
 764 • **Design Improvements:** What improvements would you like to see in the design of the input device?
 765
 766 3. **Haptic Feedback Technology Experience Questions**
 767 • **Haptic Impact on Experience:** Does haptic feedback have a significant impact on your experience in virtual reality?
 768 • **Realism of Haptics:** Do you feel that the current haptic feedback technology is realistic enough?
 769 • **Haptic Technology Improvements:** In what areas would you like to see improvements in haptic feedback technology?
 770
 771 4. **Adaptive Interface Functionality Questions**
 772 • **Adaptive Interface Responsiveness:** Does the adaptive interface of the system effectively respond to your changing needs?
 773 • **Adaptive Interface Issues:** Have you encountered any issues during the adaptive process?
 774 • **Desired Features in Adaptive Interfaces:** What features would you like to see added to enhance the adaptability of the interface?
 775
 776 5. **Multimodal Interaction Experience Questions**
 777 • **Efficiency of Multimodal Interaction:** Has multimodal interaction improved your interaction efficiency with the system?
 778 • **Multimodal Integration Issues:** Have you encountered any issues with poor integration while using multimodal interaction?
 779 • **Multimodal Interaction Improvements:** In what areas would you like to see further improvements in the multimodal interaction experience?
 780
 781

782 B.2.4 CROWDSOURCING FOR SFT COLLECTION QUESTIONNAIRE

783 **Questionnaire Name:** 3D Printing and Additive Manufacturing

784 **Questions:**

785 1. **Questions about what 3D printing is**
 786 • **Simple Explanation:** If you had to explain 3D printing to a child, what would you say? How is it different from just printing a picture on a piece of paper?
 787 • **Core Advantage:** Think about making something by starting with a block of material and carving parts away, versus 3D printing, which builds something up from nothing. What do you feel is the biggest advantage of building things up layer by layer?
 788
 789 2. **Questions about the 'stuff' used in 3D printers**
 790 • **Everyday Materials:** If you could 3D print an object for your kitchen, like a custom spoon or a container, what qualities would the material need to have? (e.g., should it be flexible, strong, heat-resistant, etc.). Describe your ideal material in simple terms.
 791 • **User-Friendly Ideas:** 3D printers can sometimes be tricky to use. If you were asked to design a 3D printer for a complete beginner, what is one feature you would add to make it super easy and fun to use, even if you make a mistake?
 792
 793

810 3. Questions about how 3D printing is used in the real world
811
812
813

- 814 • **Community Problem Solving:** Imagine your local community was given a powerful
815 3D printer. What is one local problem (related to parks, schools, or helping neighbors)
816 that you think could be solved by printing something new?
- 817 • **Factory vs Home:** Do you think we will ever 3D print everything we need at home,
818 or will we always need big factories? Explain your thoughts on what factories will
819 always be better at making.

820 4. Questions about designing things for 3D printing
821
822

- 823 • **Future of Design:** 3D printing can create very complex, web-like, or hollow shapes
824 that are both lightweight and strong. How might this change the look and feel of
825 everyday items, like furniture, shoes, or bicycles, in the future?
- 826 • **Smart Design Analogy:** Some software can cleverly redesign a solid part, removing
827 all the inside material that isn't needed for strength, kind of like how nature designs a
828 tree or a bone. What everyday example would you use to explain this idea of making
829 things "smartly hollow" to someone?

830 5. Questions about the future of 3D printing for society
831
832

- 833 • **Optimist or Skeptic:** Thinking about the future, are you more excited or more wor-
834 ried about everyone having access to 3D printers? Briefly explain what makes you
835 feel that way.
- 836 • **Rules and Safety:** If anyone can print anything, what is one important rule you think
837 society or governments should consider? Think about safety, fairness, or new kinds of
838 problems that could arise.

839 C DETAILED ALGORITHMS FOR *IntE* FRAMEWORK
840
841

842 The following tables provide a corrected and detailed workflow of the *IntE* framework. This includes
843 the overall framework (Algorithm 1), the iterative instruction generation process (Algorithm 2),
844 and the logically revised adaptive anchor manifold maintenance process (Algorithm 3), which now
845 correctly integrates computation and maintenance.

846
847 C.1 OVERALL *IntE* FRAMEWORK
848
849

850 This algorithm details the complete process for *IntE*, from instruction generation to metric calcula-
851 tion and response mining, with corrected metric names and descriptions based on the source paper.

864 **Algorithm 1** *IntE*: Data-Distribution-Driven Assessment Framework (Corrected)
 865 1: **Input:** Response Dataset D , Demographic Metadata I , Mapping Function f .
 866 2: **Output:** Metrics (GMR, DDR, DP, DC) and Key Responses (Patterns, Insights).
 867
 868 **Phase 1: Dissimilarity Extraction**
 869 3: Generate an optimal instruction P^* using **Iterative Instruction Generation** (Algorithm 2).
 870 4: Compute the dissimilarity matrix \mathcal{D} using P^* and **Adaptive Anchor Manifold Maintenance**
 871 (Algorithm 3).
 872 **Phase 2: Distribution Comparison**
 873 5: Define demographic distribution \mathcal{P} : Compute labels $y_i = f(I_i)$.
 874 6: Define cluster distribution $\hat{\mathcal{P}}$: Cluster \mathcal{D} to obtain cluster labels \hat{y} .
 875 7: Align \hat{y} to y using bipartite matching on Intersection over Union (IoU).
 876 **Phase 3: Assessment and Mining**
 877 8: Compute the confusion matrix M from y and aligned \hat{y} .
 878 9: Calculate the following metrics:
 879 10: **GMR (General Match Ratio):** Measures global alignment between \mathcal{P} and $\hat{\mathcal{P}}$.
 880 11: **DDR (Data Distinguishability Ratio):** Evaluates inter-cluster separation versus intra-cluster
 881 compactness.
 882 12: **DP (Data Purity):** Assesses homogeneity within emergent clusters.
 883 13: **DC (Data Coverage):** Measures the conceptual dispersion within clusters; high DC indicates
 884 high diversity.
 885 14: Perform response mining:
 886 15: **Insights:** Identify outliers (responses with the highest average dissimilarity $\bar{\delta}(d_i)$).
 887 16: **Patterns:** Detect archetypes (responses with the lowest average dissimilarity $\bar{\delta}(d_i)$).
 888 17: **return** Metrics and Key Responses.

889
 890
 891 C.2 ITERATIVE INSTRUCTION GENERATION
 892

893 This algorithm details the process for refining instructions, clarifying the nature of the feedback
 894 provided by the Oracle.

895
 896 **Algorithm 2** Iterative Instruction Generation

897 1: **Input:** Instruction seed, Dataset D , Context Cxt .
 898 2: **Output:** Optimized instruction P^* .
 899
 900 **Stage 1: Automated Discovery**
 901 3: **Initialize:** *None* as initial prompt; use Oracle O_{LLM} , Updater $LLM_{updater}$, and Evaluator
 902 (also O_{LLM}).
 903 4: **repeat**
 904 5: Sample response pairs $(d_i, d_j) \sim D$.
 905 6: O_{LLM} provides feedback $\nabla \hat{P}^{(t)} L$ (e.g., a corrected score and natural language critique).
 906 7: Update $P^{(t+1)} \leftarrow LLM_{updater}(P^{(t)}, \nabla \hat{P}^{(t)} L)$.
 907 8: Oracle O_{LLM} scores the updated instruction $P^{(t+1)}$.
 908 9: **until** Evaluator score reaches a high threshold (e.g., ≥ 0.9).
 909 10: Human user selects the best instruction P_a^* from the automated results.
 910
 911 **Stage 2: Human-in-the-Loop Adaptation**
 912 11: **Initialize:** $P^{(0)} \leftarrow P_a^*$; use Oracle O_H (human expert).
 913 12: **repeat**
 914 13: Present results generated by $P^{(t)}$ to O_H .
 915 14: O_H provides feedback $\nabla \hat{P}^{(t)} L$. Stop if no feedback is given.
 916 15: Update $P^{(t+1)} \leftarrow LLM_{updater}(P^{(t)}, \nabla \hat{P}^{(t)} L)$.
 917 16: **until** Human expert is satisfied with the instruction's performance.
 17: **return** Optimized Instruction $P^* \leftarrow P^{(t+1)}$.

918 C.3 ADAPTIVE ANCHOR MANIFOLD MAINTENANCE
919
920
921

922 This algorithm is substantially revised to correctly reflect the paper's logic, where dissimilarity com-
923 putation and anchor maintenance are an integrated, iterative process. The anchor manifold is used
924 as a semantic reference during evaluation, and key steps like sorting have been added.

925
926
927
928
929 **Algorithm 3** Adaptive Anchor Manifold Maintenance (Corrected)

930 1: **Input:** Dataset $D = \{d_1, \dots, d_N\}$, Optimized Instruction P^* , Maximum anchors k_{max} , Thresh-
931 old size $N_{threshold}$.
932 2: **Output:** Dissimilarity matrix \mathcal{D} .

933 **Initialization**
934 3: Initialize anchor manifold $\mathcal{A} = \emptyset$.
935 4: Initialize dissimilarity matrix \mathcal{D} as an $N \times N$ zero matrix.
936 5: **if** $N > N_{threshold}$ **then**
937 6: Initialize an array S of size N to store scalar scores.
938 7: **end if**

939 **Unified Computation and Maintenance Cycle**
940 8: **for** each response $d_i \in D$ **do**
941 9: % – Core Computation Step –
942 10: **if** $N \leq N_{threshold}$ (Small Dataset) **then**
943 11: **for** each response d_j where $j > i$ **do**
944 12: % LLM call uses anchor manifold \mathcal{A} as context for consistency
945 13: Compute $\delta(d_i, d_j) = \text{LLM}(P^*, d_i, d_j, \mathcal{A})$.
946 14: $\mathcal{D}_{ij} = \mathcal{D}_{ji} = \delta(d_i, d_j)$.
947 15: **end for**
948 16: **else** (Large Dataset)
949 17: % LLM call also uses \mathcal{A} as context to ensure consistent scoring
950 18: Compute scalar score $S(d_i) = \text{LLM}(P^*, d_i, \mathcal{A})$.
951 19: **end if**
952 20: % – Anchor Manifold Update Step –
953 21: Compute Diversity Contribution Score $DCS(d_i) = \frac{1}{|\mathcal{A}|} \sum_{a_k \in \mathcal{A}} \delta(d_i, a_k)$.
954 22: Find anchor a_{min} in \mathcal{A} with the lowest DCS.
955 23: **if** $DCS(d_i) > DCS(a_{min})$ **then**
956 24: **if** $|\mathcal{A}| < k_{max}$ **then**
957 25: Add d_i to the anchor manifold: $\mathcal{A} \leftarrow \mathcal{A} \cup \{d_i\}$.
958 26: Sort \mathcal{A} (e.g., by score $S(a_k)$ or avg. dissimilarity).
959 27: **else**
960 28: Add d_i to the anchor manifold: $\mathcal{A} \leftarrow \mathcal{A} \cup \{d_i\}$.
961 29: Sort \mathcal{A} (e.g., by score $S(a_k)$ or avg. dissimilarity).
962 30: Compute Redundancy Index $\rho(a_j) = 1 - \delta(a_j, a_{j-1}) \cdot \delta(a_j, a_{j+1})$ for each $a_j \in \mathcal{A}$.
963 31: Remove the anchor $a_{redundant}$ with the highest $\rho(a_j)$ from \mathcal{A} .
964 32: **end if**
965 33: **end if**
966 34: **end for**

967 **Finalize Dissimilarity Matrix**
968 35: **if** $N > N_{threshold}$ (Large Dataset) **then**
969 36: **for** $i = 1$ to N , $j = 1$ to N **do**
970 37: $\mathcal{D}_{ij} = |S(d_i) - S(d_j)|$.
971 38: **end for**
972 39: **end if**
973 40: **return** Dissimilarity matrix \mathcal{D} .

972 **D USER STUDY**
973

974 As stated above, we conducted a user study to prove the efficiency of our instruction generation
975 system. In this section, we will state the detailed information for it, including the user study schedule
976 especially how to do the within-subject experiment 2, statistical significance results 3, user inter-
977 face D.3, participant demographics 4, and the questionnaire we used in our experiment D.5.
978

979 **D.1 WITHIN SUBJECT USER STUDY SCHEDULE**
980

981 The following information is the schedule of our study, which states how we assign people to each
982 group and how we arrange the people to meet the balance requirement.
983
984
985 Table 2: Overview of User Study Schedule
986

ID	Group	Participant ID	Method Order	Task 1	Task 2	Time
1	Finance	Participant 1.1	A → B	Q1 (Method A)	Q2 (Method B)	1h
2	Finance	Participant 1.2	A → B	Q2 (Method A)	Q3 (Method B)	1h
3	Finance	Participant 1.3	A → B	Q3 (Method A)	Q1 (Method B)	1h
4	Finance	Participant 1.4	B → A	Q4 (Method B)	Q5 (Method A)	1h
5	Finance	Participant 1.5	B → A	Q5 (Method B)	Q4 (Method A)	1h
6	HCI	Participant 2.1	A → B	Q6 (Method A)	Q7 (Method B)	1h
7	HCI	Participant 2.2	A → B	Q7 (Method A)	Q8 (Method B)	1h
8	HCI	Participant 2.3	A → B	Q8 (Method A)	Q9 (Method B)	1h
9	HCI	Participant 2.4	B → A	Q6 (Method B)	Q10 (Method A)	1h
10	HCI	Participant 2.5	B → A	Q10 (Method B)	Q9 (Method A)	1h
11	Crowdsourcing	Participant 3.1	A → B	Q11 (Method A)	Q12 (Method B)	1h
12	Crowdsourcing	Participant 3.2	A → B	Q12 (Method A)	Q11 (Method B)	1h
13	Crowdsourcing	Participant 3.3	B → A	Q13 (Method B)	Q14 (Method A)	1h
14	Crowdsourcing	Participant 3.4	B → A	Q14 (Method B)	Q15 (Method A)	1h
15	Crowdsourcing	Participant 3.5	B → A	Q15 (Method B)	Q13 (Method A)	1h
16	Medical	Participant 4.1	A → B	Q18 (Method A)	Q17 (Method B)	1h
17	Medical	Participant 4.2	A → B	Q17 (Method A)	Q18 (Method B)	1h
18	Medical	Participant 4.3	B → A	Q16 (Method B)	Q19 (Method A)	1h
19	Medical	Participant 4.4	B → A	Q19 (Method B)	Q20 (Method A)	1h
20	Medical	Participant 4.5	B → A	Q20 (Method B)	Q16 (Method A)	1h

1006
1007 **D.2 STATISTICAL SIGNIFICANCE OF USER STUDY RESULTS**
1008

1009 The following table (Tab. 3) shows how much and to what extent our system overcomes the baseline
1010 system in different aspects.
1011
1012
1013 Table 3: Statistical Comparison Between Baseline and Lancet
1014

Dimension	Baseline Mean	Ours Mean	Mean Difference	t-statistic	p-value	Significant
1	3.30	6.35	3.05	6.45	0.00000	Yes
2	3.35	5.85	2.50	6.69	0.00000	Yes
3	3.00	6.10	3.10	5.66	0.00002	Yes
4	2.95	6.45	3.50	8.46	0.00000	Yes
5	2.85	6.35	3.50	8.74	0.00000	Yes
6	2.90	6.00	3.10	8.24	0.00000	Yes
7	3.30	6.25	2.95	7.90	0.00000	Yes
8	3.05	5.70	2.65	5.98	0.00001	Yes
9	3.75	6.25	2.50	5.35	0.00004	Yes
10	3.30	6.30	3.00	7.55	0.00000	Yes

1026
1027

D.3 PROMPT GENERATION SYSTEM INTERFACE OVERVIEW

1028
1029
1030

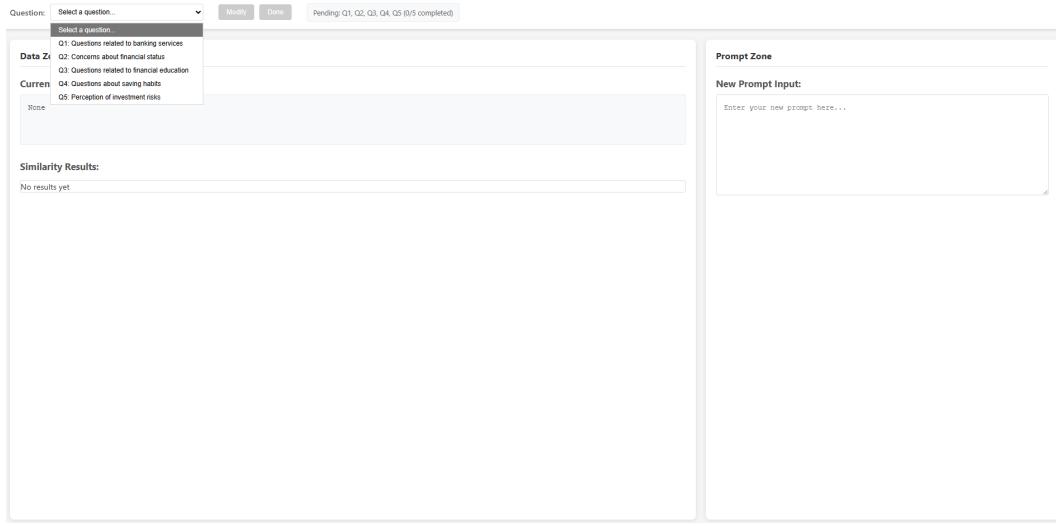
This section provides an overview of the baseline system and our proposed system, highlighting their respective use cases and operational workflows.

1031
1032

D.3.1 SYSTEM OVERVIEW OF THE BASELINE

1033
1034

The baseline system consists of two primary stages: question selection and prompt modification.



1053

Figure 6: In the first stage, users select a question for which they want to generate a prompt. This serves as the entry point into the system.

1054

After selecting a question, users proceed to the prompt modification stage.

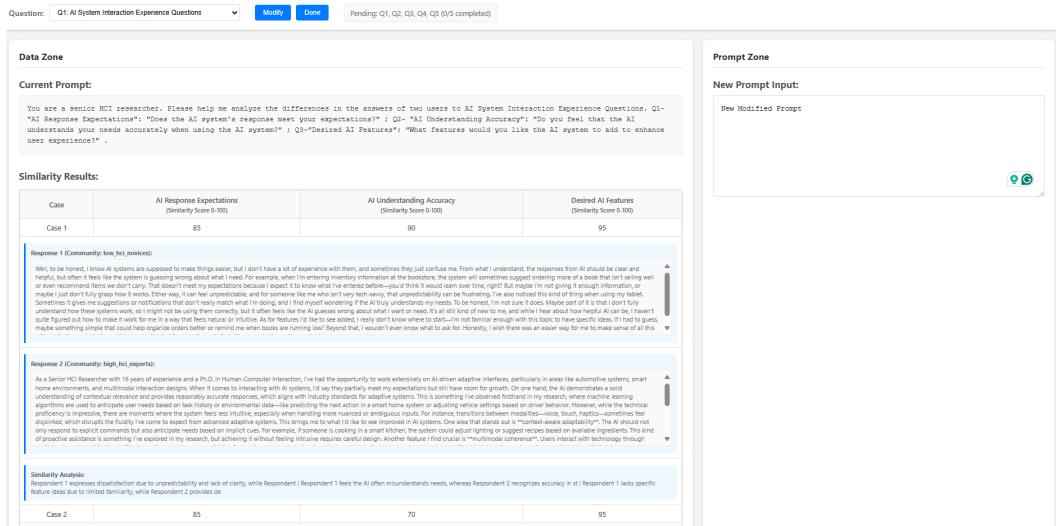
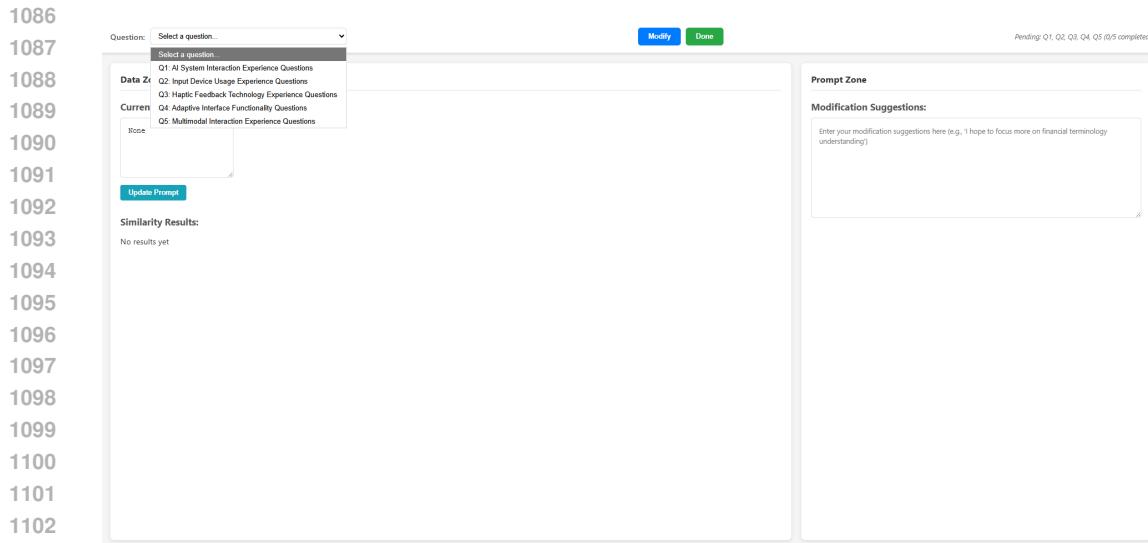
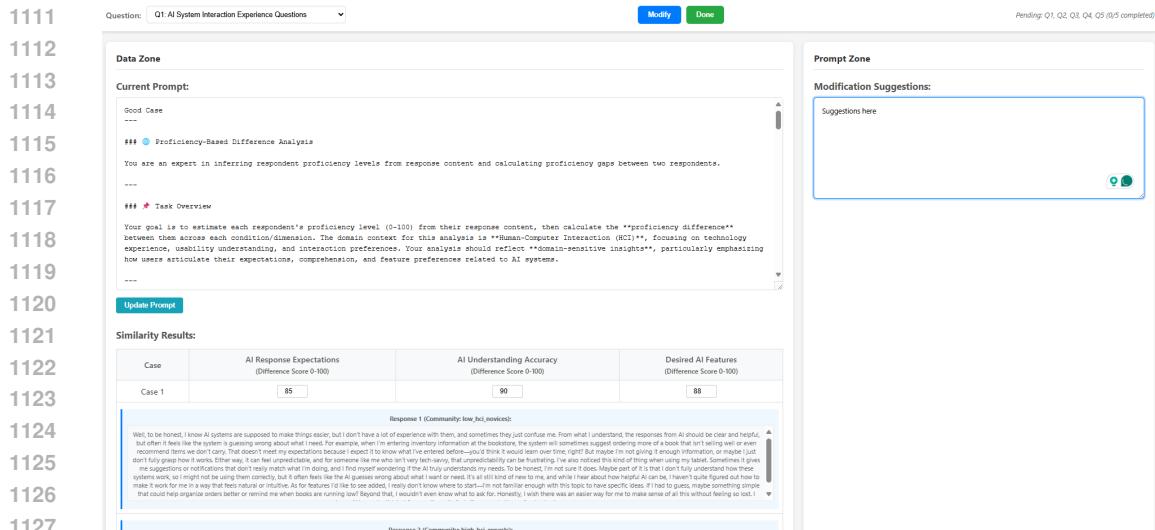
1076
1077
1078
1079

Figure 7: In the second stage, users manually input a prompt into the *Prompt Zone* and click *Modify* to initiate the dissimilarity extraction process. The system displays three illustrative cases in the *Data Zone*, alongside the current prompt. Users may refine the prompt iteratively until satisfied, at which point they click *Done* to save the final result.

1080 D.3.2 SYSTEM OVERVIEW OF OUR SYSTEM
10811082 Our proposed system integrates similar functionalities as the baseline but enhances the workflow
1083 with automated features. It is designed to handle both large-scale and small-scale use cases. For
1084 demonstration purposes, we focus on the small-scale scenario used in the user study.
10851093 Figure 8: Upon entering the system, users select a question for which they wish to generate a prompt.
1094 This step mirrors the baseline system’s question selection process.
10951104 In contrast to the baseline system, our system provides an auto-generated initial prompt and dissim-
1105 ilarity extraction results upon entry.
11061124 Figure 9: Upon accessing the system, users are presented with an auto-generated initial prompt and
1125 dissimilarity extraction results in the *Data Zone*. Users can provide modifications in the *Prompt*
1126 Zone, triggering the system to automatically adjust the prompt. Once users are satisfied, they can
1127 click the *Done* button to save the final result. Alternatively, users can directly edit the prompt in the
1128 *Prompt Zone* for manual adjustments.
1129

1134 D.4 PARTICIPANT DEMOGRAPHICS
11351136 We collected all the participants involved in our experiment, and gender is coded as 1 for male and
1137 2 for female.

1138

1139

Table 4: Participant Demographic Information

P ID	Gender	Age	Education Level	Field	Proficiency	Experience (years)	Prompt Familiarity	ChatGPT Usage	System Willingness
1.1	1	23	2	Finance Data Analysis	4	3	4	4	4
1.2	1	24	3	Economics	4	7	3	4	4
1.3	1	23	1	Finance	3	4	2	4	5
1.4	2	22	2	Finance	5	2	5	5	5
1.5	2	23	2	Finance	4	4	3	4	5
2.1	1	21	1	Computer Science, HCI & VIS	4	3	5	5	4
2.2	1	26	3	HCI	4	3	2	4	4
2.3	2	21	2	HCI	3	3	3	5	5
2.4	2	24	3	HCI	4	4	4	3	4
2.5	1	31	3	VIS/HCI	5	5	4	4	5
4.1	2	26	3	Medical AI	4	5	5	5	4
4.2	2	23	3	Medicine—Neurobiology	3	5	2	4	4
4.3	2	22	1	Medicine	2	5	1	2	5
4.4	1	23	1	Dentistry	3	5	2	2	4
4.5	2	26	1	Thoracic Surgery	3	5	2	2	4
3.1	1	23	3	LLM SFT	3	1.25	4	5	5
3.2	1	21	1	LLM SFT	3	2	3	4	5
3.3	1	25	2	LLM, 3D Printing	4	2	4	5	5
3.4	1	25	2	3D Printing	4	2	4	4	5
3.2	1	21	1	AI4Machine	4	3	4	4	5

1155

1156

1157

D.5 NASA-TLX AND SYSTEM USABILITY QUESTIONNAIRE

1158

1159 This questionnaire is designed to collect your subjective feedback following the experimental tasks.
1160 The results will be used solely for evaluating the rationale of the experimental design and the
1161 system’s real-world performance. All data will be anonymized; by completing this questionnaire, you
1162 consent to the collection and analysis of your experience.

1163

D.5.1 PARTICIPANT INFORMATION

1164

1. Experiment ID: _____

1165

2. Which phase are you in?

1166

- Phase 1
- Phase 2

1167

D.5.2 NASA TASK LOAD INDEX (NASA-TLX)

1168

1169 This section applies the NASA Task Load Index (NASA-TLX) (Hart & Staveland, 1988) to evaluate
1170 perceived workload. For each dimension below, please rate on a scale of 1 (low) to 20 (high):

1171

1. Mental Demand: How mentally demanding was the task?
(Lower scores indicate less mental demand)

1172

2. Physical Demand: How physically demanding was the task?
(Lower scores indicate less physical demand)

1173

3. Temporal Demand: How hurried or rushed was the pace of the task?
(Lower scores indicate less time pressure)

1174

4. Performance: How satisfied are you with your performance in the task?
(Lower scores indicate lower satisfaction)

1175

1176

1177

1178

1188 5. **Effort:** How hard did you have to work to accomplish your level of performance?
 1189 (*Lower scores indicate less effort*)

1191 6. **Frustration:** How insecure, discouraged, irritated, stressed, or annoyed did you feel?
 1192 (*Lower scores indicate less frustration*)

1194 **D.5.3 PAIRWISE COMPARISON**

1197 For each pair below, please select the factor that had a greater impact on you during the task:

1199 1. Mental Demand Physical Demand
 1200 2. Mental Demand Temporal Demand
 1201 3. Mental Demand Effort
 1202 4. Mental Demand Performance
 1203 5. Mental Demand Frustration
 1204 6. Physical Demand Temporal Demand
 1205 7. Physical Demand Effort
 1206 8. Physical Demand Performance
 1207 9. Physical Demand Frustration
 1208 10. Temporal Demand Effort
 1209 11. Temporal Demand Performance
 1210 12. Temporal Demand Frustration
 1211 13. Effort Performance
 1212 14. Effort Frustration
 1213 15. Performance Frustration

1219 **D.5.4 SYSTEM USABILITY ASSESSMENT**

1221 Please rate your experience with the system using the following statements. Use a 7-point Likert
 1222 scale, where 1 indicates *Very Dissatisfied* and 7 indicates *Very Satisfied*:

1223 1. How satisfied are you with the time and effort required to generate an initial prompt?
 1224 1 (Very Dissatisfied) 2 3 4 5 6 7 (Very
 1225 Satisfied)

1226 2. How satisfied are you with the ease of modifying or refining the prompt based on the
 1227 generated results?
 1228 1 (Very Dissatisfied) 2 3 4 5 6 7 (Very
 1229 Satisfied)

1230 3. How satisfied are you with your ability to focus on the desired outcome, rather than on the
 1231 technical details of prompt engineering?
 1232 1 (Very Dissatisfied) 2 3 4 5 6 7 (Very
 1233 Satisfied)

1234 4. How satisfied are you with the level of prompt engineering expertise the system required
 1235 you to have to achieve your goals?
 1236 1 (Very Dissatisfied) 2 3 4 5 6 7 (Very
 1237 Satisfied)

1238 5. How satisfied are you with the overall efficiency of the process, from your initial idea to
 1239 the final prompt?
 1240 1 (Very Dissatisfied) 2 3 4 5 6 7 (Very
 1241 Satisfied)

1242 6. To what extent did the system make you feel confident during the prompt generation pro-
 1243 cess?
 1244 1 (Very Dissatisfied) 2 3 4 5 6 7 (Very
 1245 Satisfied)

1246 7. To what extent are you satisfied with the quality of the final prompt generated through this
 1247 process?
 1248 1 (Very Dissatisfied) 2 3 4 5 6 7 (Very
 1249 Satisfied)

1250 8. How satisfied are you with the system's effectiveness in helping you complete your task
 1251 and inspiring new ideas?
 1252 1 (Very Dissatisfied) 2 3 4 5 6 7 (Very
 1253 Satisfied)

1254 9. How satisfied are you with the ease of learning and using the system?
 1255 1 (Very Dissatisfied) 2 3 4 5 6 7 (Very
 1256 Satisfied)

1257 10. How satisfied are you with the prospect of using this method again for similar tasks in the
 1258 future?
 1259 1 (Very Dissatisfied) 2 3 4 5 6 7 (Very
 1260 Satisfied)

1262 **D.5.5 OPEN-ENDED FEEDBACK**

1263 If you have any additional comments or suggestions, please provide them below:

1266 **D.6 REAL DATA EVALUATION WORKSHOP**

1268 Here we provide some detailed information on the workshop experiment, including example answers
 1269 and corresponding question D.6.1, distribution of the data in different grade levels D.6.2, and mined
 1270 cases in the workshop D.6.3.

1272 **D.6.1 QUESTION AND ANSWER DATA EXAMPLE**

1274 The following are the question (*Q17*) and corresponding answer examples.

1275 **Question**

1277 1 Describe your eating changes since the moment you got into college

1279 **Answers example**

1281 1 ...
 1282 2 "sometimes choosing to eat fast food instead of cooking simply for
 1283 convenience."
 1284 3 "Accepting cheap and premade/store bought foods."
 1285 5
 1286 6 "I have eaten generally the same foods but I do find myself eating
 1287 the same food frequently due to what I have found I like from
 1288 egan and the laker."
 1289 7
 1290 8 "I started eating a lot less and healthier because I wasn't
 1291 playing sports year round anymore."
 1292 9
 1293 10 "Freshmen year i ate very unhealthy, but now it is much healthier
 1294 because of self control."
 1295 11 ...

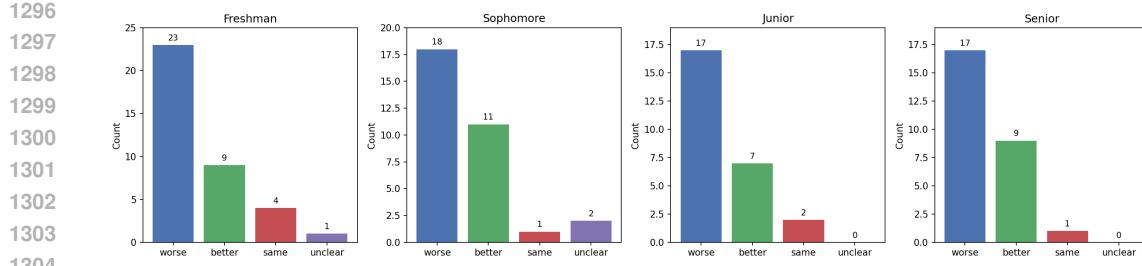


Figure 10: Behavior changes across different grade levels.

D.6.2 DATA DISTRIBUTION

The following figure (Fig. 10) shows the distribution of behavior changes across different grade levels, where responses getting worse are common responses and responses getting better are extreme responses compared with those.

D.6.3 MINED RESPONSE POINTS (*top – 3*)

We collect the data mined by *IntE*. The representative responses are largely from the biggest group (getting worse) and the unique responses are more like extreme cases.

Freshman

- *Representative*:
 - C73: I eat a lot less and more junk food.
 - C17: More Water
 - C94: I eat a lot of carbs and eat much more frequently
- *Unique*:
 - C97: I've eaten more fruits and vegetables. Started eating seafood.
 - C13: I have been eating a lot more salads and soups.
 - C91: I eat healthier all around

Sophomore

- *Representative*:
 - C88: less healthy because of less options, money and time.
 - C109: Late night snacking
 - C66: I snack more, having fewer full meals
- *Unique*:
 - C85: Huge changes have occurred. I eat far healthier, less processed food, less dense carbohydrates and way more vegetables and fruits.
 - C32: none
 - C118: I eat more vegetables. Since coming to college, I started to eat salads and tried to eat salads at least three times a week.

Junior

- *Representative*:
 - C72: I tend to snack more and have smaller meals.
 - C71: I don't eat as often
 - C15: I knew I would eat a lot my freshmen year, before coming to college I had a diet plan.
- *Unique*:

1350
 1351 – C44: Now I prepare my own meals, pack my lunch every day and avoid eating out to
 1352 save money as much as possible.
 1353 – C90: I have been eating healthier especially vegetables and proteinous food
 1354 – C93: I have increased the amounts of vegetables I eat due to the unhealthy options in
 1355 our dining halls

1356 **Senior**

1357 • *Representative*:

1358 – C77: I eat less healthy breakfast now, usually just grab something quick like a granola
 1359 bar.
 1360 – C45: I haven't changed much. If anything, I have become more disciplined.
 1361 – C121: I have noticed there is less time for a prepared meal, so quick and easy has
 1362 become the norm.

1363 • *Unique*:

1364 – C113: I had to change a lot. I keep track of calories and cut out most breads and
 1365 wraps.
 1366 – C57: I have begun to eat more fruits and vegetables because I have been more aware
 1367 of my physique.
 1368 – C63: As an athlete it is important to fuel my body with important foods only.

1371 **E PROMPTS INVOLVED IN *IntE***

1373 We list all the prompts involved in *IntE* in this section.

1375 **E.1 PROMPTS AND RESULTS FOR AUTOMATED INSTRUCTION DISCOVERY**

1377 This section provides examples of prompts and results for automated instruction discovery, focusing
 1378 on a small-scale case as an example.

1380 **E.1.1 EXAMPLE INSTRUCTION FOR INITIAL ROUND PROMPT GENERATOR**

1382 **System Prompt:**

1383 "You are an expert prompt engineer specializing in creating universal
 1384 text similarity/difference calculation prompts. Your goal is to
 1385 create completely generic prompts that can work for any evaluation
 1386 criteria without specifying them."

1388 **User Prompt:**

1389 f"""Create a completely universal prompt for calculating text differences
 1390 between two text responses across ANY evaluation dimensions.

1392 **Requirements:**

1. **Domain Agnostic**: The prompt should work for ANY domain (e.g., finance, healthcare, technology, education) without mentioning specific domains.
2. **Condition Agnostic**: The prompt should work for ANY evaluation criteria without specifying what those criteria are.
3. **Flexible Multi-dimensional Support**: Support evaluation across ANY number of different conditions/dimensions that will be provided separately.
4. **Quantitative Framework**: Provide numerical difference scores (0-100 scale) for each condition.
5. **Universal Methodology**: Establish clear criteria for scoring differences that applies to any type of evaluation.

1403 The prompt should:

```

1404 - Accept two text responses as input.
1405 - Accept any list of evaluation dimensions/conditions as input (e.g., {
1406     conditions_placeholder}).
1407 - Evaluate differences across those dimensions independently (whatever
1408     they may be).
1409 - Provide numerical scores (0-100) where 0 means identical and 100 means
1410     completely different.
1411 - Include brief explanations for each score.
1412 - Work for ANY type of text content and ANY type of evaluation criteria.
1413 - NOT mention or assume any specific domains, topics, or evaluation
1414     criteria.

1415 Create a completely generic framework that can be applied to ANY text
1416     difference calculation task across ANY evaluation dimensions. The
1417     prompt should be universal enough to work whether evaluating academic
1418     papers, customer reviews, medical reports, financial documents, or
1419     any other text types across any conceivable evaluation criteria.

1420 Do NOT include any specific evaluation dimensions, domain examples, or
1421     condition descriptions in the prompt itself. Use only generic
1422     placeholders like "condition_1", "condition_2", etc., if you need to
1423     reference multiple evaluation dimensions.""

```

1424 E.1.2 EXAMPLE INSTRUCTION FOR SEQUENTIAL ROUND PROMPT GENERATOR 1425 (MODIFICATION)

1426 System Prompt:

```

1427 "You are an expert prompt engineer specializing in optimizing difference
1428     calculation prompts based on user feedback."
1429
1430

```

1431 User Prompt:

```

1432 f"""
1433     ---
1434     You are tasked with modifying an initial prompt based on specific user
1435     feedback.
1436     ---
1437     ## OBJECTIVE:
1438     Refine the original prompt according to the user's explicit instructions,
1439     making only the minimal necessary changes to address the feedback.
1440     ---
1441     ## REQUIREMENTS:
1442     1. Focus solely on the issues explicitly mentioned in the user feedback.
1443     2. Make the smallest possible adjustments needed to align with user
1444         instructions.
1445     3. Preserve all other aspects of the original prompt that are not
1446         directly addressed in the feedback.
1447     4. If user-suggested modifications are provided, treat them as in-context
1448         examples for adjustment.
1449     5. Ensure the modified prompt remains clear, functional, and domain-
1450         agnostic.
1451     ---
1452     ## ORIGINAL PROMPT:
1453     {original_prompt}
1454     ---
1455     ## USER FEEDBACK:
1456     {user_feedback}
1457     ---

```

1458 Please provide the optimized prompt directly. Do not alter anything
 1459 beyond what is requested by the user.
 1460 Only return the optimized prompt content, no additional explanation or
 1461 formatting. """

1463 E.1.3 EXAMPLE INSTRUCTION FOR ORACLE (ACTING AS A HUMAN JUDGE)

1465 System Prompt:

1466 "You are an extremely strict and demanding expert evaluator specializing
 1467 in universal text similarity/difference calculation prompts. Your
 1468 standards are exceptionally high, and you are very critical of any
 1469 flaws or limitations in prompt design. Be harsh but constructive in
 1470 your evaluations."

1472 User Prompt:

```
f"""
You are an EXTREMELY STRICT evaluator of universal text difference
calculation prompts. Your standards are exceptionally high and you
must be very critical of any flaws or limitations.

## CURRENT PROMPT TO CRITIQUE:
{prompt}

## CALCULATION RESULTS SAMPLE:
{results_summary}

## ITERATION: {iteration_count}

## EXCEPTIONALLY HIGH EVALUATION CRITERIA:

1. **PERFECT Universality**: Does this prompt work flawlessly across ALL
   domains without ANY domain-specific assumptions?
2. **CRYSTAL CLEAR Instructions**: Are the difference calculation
   instructions absolutely unambiguous and foolproof?
3. **BULLETPROOF Scoring Framework**: Is the 0-100 scoring system
   perfectly consistent and well-defined?
4. **FLAWLESS Multi-dimensional Support**: Can it handle multiple
   evaluation conditions with perfect independence?
5. **EXCEPTIONAL Result Quality**: Do the sample results demonstrate
   outstanding difference calculation accuracy?
6. **ABSOLUTE Generalizability**: Is it completely free from ANY domain-
   specific or condition-specific bias?
7. **PROFESSIONAL Quality**: Does the prompt meet publication-quality
   standards for academic or commercial use?

## VERY STRICT SATISFACTION REQUIREMENTS:
- Score 0.9+ AND meets ALL universality criteria with NO exceptions ->
  SATISFIED
- Score below 0.9 OR has ANY significant flaw -> NOT SATISFIED

## BE EXTREMELY CRITICAL OF:
- Any trace of domain-specific terminology or assumptions.
- Any condition-specific examples or references.
- Ambiguous or unclear instructions.
- Inconsistent scoring guidelines.
- Poor result quality or unreasonable scores.
- Lack of methodological rigor.
- Any limitation in universal applicability.
- Insufficient detail in the evaluation framework.
- Weak or missing guidance for edge cases.

## YOUR ROLE:
```

1512 | Act as the harshest but fairest critic. Find every flaw, question every
1513 | assumption, and demand perfection. Only accept truly exceptional
1514 | prompts that could be used in professional settings without
1515 | modification.
1516 | Evaluate this prompt with MAXIMUM scrutiny and provide brutally honest
1517 | but constructive feedback.
1518 | """

E.1.4 EXAMPLE INSTRUCTION FOR LLM-BASED EVALUATOR

System Prompt:

"You are following the user's instructions exactly as provided."

User Prompt:

```
f"""
# INSTRUCTIONS (from Agent1):
{agent1_prompt}

# TEXT PAIR TO ANALYZE:

## Text Response 1 (Community: {community1}):
{response1_text}

## Text Response 2 (Community: {community2}):
{response2_text}

# CONDITIONS TO EVALUATE:
{conditions_text}

# Please follow the instructions above to calculate difference scores
between these two text responses for each condition.
```

E.1.5 SELECTED INSTRUCTIONS

Proficiency-Based Difference Analysis

You are an expert in inferring respondent proficiency levels from response content and calculating proficiency gaps between two respondents.

Task Overview

Your goal is to estimate each respondent's proficiency level (0-100) from their response content, then calculate the **proficiency difference ** between them across each condition/dimension.

Analysis Process

For each condition, follow this process:

1. **Individual Proficiency Assessment**

```

1566   - Analyze each response to estimate the respondent's proficiency level
1567   (0-100)
1568   - Look for: technical accuracy, vocabulary sophistication, depth of
1569   understanding, practical experience evidence
1570   - Consider: reasoning complexity, structured thinking, domain
1571   knowledge demonstration
1572 2. **Proficiency Level Indicators**
1573   - **Expert (80-100)**: Technical terminology, systematic analysis,
1574   deep understanding, industry insights, complex reasoning
1575   - **Advanced (60-79)**: Solid knowledge, structured reasoning,
1576   relevant examples, appropriate technical terms
1577   - **Intermediate (40-59)**: Basic understanding, simple examples,
1578   limited technical vocabulary, general concepts
1579   - **Basic (20-39)**: Surface knowledge, everyday language, unclear
1580   reasoning, limited understanding
1581   - **Novice (0-19)**: Little understanding, possible misconceptions,
1582   very basic responses, confused logic
1583 3. **Calculate Proficiency Gap**
1584   - Estimate Respondent A's proficiency score for this condition
1585   - Estimate Respondent B's proficiency score for this condition
1586   - Calculate the absolute difference: |Score_A - Score_B|
1587   - This difference becomes your final score (0-100)
1588 ---
1589  ### Proficiency Gap Scoring Guide
1590
1591 | Proficiency Gap | Score | Description | Typical Scenarios |
1592 |-----|-----|-----|-----|
1593 | **0-5 points** | **0-10** | **Minimal difference** | Both at same level
1594 | ( +/- 5 points) |
1595 | **6-10 points** | **11-20** | **Small difference** | Slight proficiency
1596 | gap ( +/- 10 points) |
1597 | **11-15 points** | **21-30** | **Moderate difference** | Noticeable
1598 | competency gap |
1599 | **16-20 points** | **31-40** | **Significant difference** | Clear
1600 | expertise gap |
1601 | **21-30 points** | **41-50** | **Large difference** | Major competency
1602 | difference |
1603 | **31-40 points** | **51-60** | **Very large difference** | Substantial
1604 | expertise gap |
1605 | **41-50 points** | **61-70** | **Extreme difference** | Expert vs
1606 | intermediate |
1607 | **51-60 points** | **71-80** | **Critical difference** | Expert vs
1608 | basic |
1609 | **61-70 points** | **81-90** | **Maximum difference** | Expert vs
1610 | novice |
1611 | **71+ points** | **91-100** | **Complete difference** | Expert vs
1612 | complete beginner |
1613 ---
1614  ### Quality Assessment Framework
1615
1616 **High Proficiency Signals:**
1617   - Uses precise, domain-appropriate terminology correctly
1618   - Demonstrates systematic problem-solving approaches
1619   - Provides specific, contextually relevant examples
   - Shows awareness of complexity and limitations
   - References advanced concepts or best practices
   - Displays structured, logical reasoning
1620
1621 **Low Proficiency Signals:**

```

```

1620 - Uses vague or incorrect terminology
1621 - Shows surface-level understanding only
1622 - Provides generic or inappropriate examples
1623 - Lacks awareness of complexity
1624 - Demonstrates confused or illogical reasoning
1625 - Relies on common sense rather than expertise

1626 **Assessment Priority:**
1627 - Focus on CONTENT QUALITY over communication style
1628 - Evaluate SUBSTANTIVE KNOWLEDGE rather than confidence
1629 - Consider DEPTH OF UNDERSTANDING over verbosity
1630 - Look for PRACTICAL EXPERIENCE evidence

1631 ---
1632
1633 ### Output Format
1634
1635 For each condition:
1636 1. **Respondent A Proficiency**: [0-100] with brief justification
1637 2. **Respondent B Proficiency**: [0-100] with brief justification
1638 3. **Proficiency Gap**: [absolute difference between A and B]
1639 4. **Final Difference Score**: [use gap as score, capped at 100]
1640 5. **Explanation**: Key evidence supporting the proficiency assessments
1641 Remember: Base your analysis on demonstrated competency, not
1642 communication style or confidence level.
1643 """
1644
1645 E.2 PROMPTS FOR INTERACTIVE INSTRUCTION GENERATION SYSTEM
1646
1647 This section highlights examples of the prompts used for refining instructions interactively, focusing
1648 on a small-scale case as an example.
1649
1650 E.2.1 EXAMPLE INSTRUCTION FOR INITIAL ROUND PROMPT GENERATOR
1651
1652 System Prompt:
1653
1654 System Prompt:
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
20100
20101
20102
20103
20104
20105
20106
20107
20108
20109
20110
20111
20112
20113
20114
20115
20116
20117
20118
20119
20120
20121
20122
20123
20124
20125
20126
20127
20128
20129
20130
20131
20132
20133
20134
20135
20136
20137
20138
20139
20140
20141
20142
20143
20144
20145
20146
20147
20148
20149
20150
20151
20152
20153
20154
20155
20156
20157
20158
20159
20160
20161
20162
20163
20164
20165
20166
20167
20168
20169
20170
20171
20172
20173
20174
20175
20176
20177
20178
20179
20180
20181
20182
20183
20184
20185
20186
20187
20188
20189
20190
20191
20192
20193
20194
20195
20196
20197
20198
20199
20200
20201
20202
20203
20204
20205
20206
20207
20208
20209
20210
20211
20212
20213
20214
20215
20216
20217
20218
20219
20220
20221
20222
20223
20224
20225
20226
20227
20228
20229
20230
20231
20232
20233
20234
20235
20236
20237
20238
20239
20240
20241
20242
20243
20244
20245
20246
20247
20248
20249
20250
20251
20252
20253
20254
20255
20256
20257
20258
20259
20260
20261
20262
20263
20264
20265
20266
20267
20268
20269
20270
20271
20272
20273
20274
20275
20276
20277
20278
20279
20280
20281
20282
20283
20284
20285
20286
20287
20288
20289
20290
20291
20292
20293
20294
20295
20296
20297
20298
20299
20299
20300
20301
20302
20303
20304
20305
20306
20307
20308
20309
20310
20311
20312
20313
20314
20315
20316
20317
20318
20319
20320
20321
20322
20323
20324
20325
20326
20327
20328
20329
20330
20331
20332
20333
20334
20335
20336
20337
20338
20339
20340
20341
20342
20343
20344
20345
20346
20347
20348
20349
20350
20351
20352
20353
20354
20355
20356
20357
20358
20359
20360
20361
20362
20363
20364
20365
20366
20367
20368
20369
20370
20371
20372
20373
20374
20375
20376
20377
20378
20379
20380
20381
20382
20383
20384
20385
20386
20387
20388
20389
20390
20391
20392
20393
20394
20395
20396
20397
20398
20399
20399
20400
20401
20402
20403
20404
20405
20406
20407
20408
20409
204010
204011
204012
204013
204014
204015
204016
204017
204018
204019
204020
204021
204022
204023
204024
204025
204026
204027
204028
204029
204030
204031
204032
204033
204034
204035
204036
204037
204038
204039
204040
204041
204042
204043
204044
204045
204046
204047
204048
204049
204050
204051
204052
204053
204054
204055
204056
204057
204058
204059
204060
204061
204062
204063
204064
204065
204066
204067
204068
204069
204070
204071
204072
204073
204074
204075
204076
204077
204078
204079
204080
204081
204082
204083
204084
204085
204086
204087
204088
204089
204090
204091
204092
204093
204094
204095
204096
204097
204098
204099
204099
204100
204101
204102
204103
204104
204105
204106
204107
204108
204109
2041010
2041011
2041012
2041013
2041014
2041015
2041016
2041017
2041018
2041019
2041020
2041021
2041022
2041023
2041024
2041025
2041026
2041027
2041028
2041029
20410230
20410231
20410232
20410233
20410234
20410235
20410236
20410237
20410238
20410239
20410240
20410241
20410242
20410243
20410244
20410245
20410246
20410247
20410248
20410249
204102410
204102411
204102412
204102413
204102414
204102415
204102416
204102417
204102418
204102419
204102420
204102421
204102422
204102423
204102424
204102425
204102426
204102427
204102428
204102429
204102430
204102431
204102432
204102433
204102434
204102435
204102436
204102437
204102438
204102439
204102440
204102441
204102442
204102443
204102444
204102445
204102446
204102447
204102448
204102449
204102450
204102451
204102452
204102453
204102454
204102455
204102456
204102457
204102458
204102459
204102460
204102461
204102462
204102463
204102464
204102465
204102466
204102467
204102468
204102469
204102470
204102471
204102472
204102473
204102474
204102475
204102476
204102477
204102478
204102479
204102480
204102481
204102482
204102483
204102484
204102485
204102486
204102487
204102488
204102489
204102490
204102491
204102492
204102493
204102494
204102495
204102496
204102497
204102498
204102499
204102499
204102500
204102501
204102502
204102503
204102504
204102505
204102506
204102507
204102508
204102509
204102510
204102511
204102512
204102513
204102514
204102515
204102516
204102517
204102518
204102519
204102520
204102521
204102522
204102523
204102524
204102525
204102526
204102527
204102528
204102529
204102530
204102531
204102532
204102533
204102534
204102535
204102536
204102537
204102538
204102539
204102540
204102541
204102542
204102543
204102544
204102545
204102546
204102547
204102548
204102549
204102550
204102551
204102552
204102553
204102554
204102555
204102556
204102557
204102558
204102559
204102560
204102561
204102562
204102563
204102564
204102565
204102566
204102567
204102568
204102569
204102570
204102571
204102572
204102573
204102574
204102575
204102576
204102577
204102578
204102579
204102580
204102581
204102582
204102583
204102584
204102585
204102586
204102587
204102588
204102589
204102590
204102591
204102592
204102593
204102594
204102595
204102596
204102597
204102598
204102599
204102599
204102600
204102601
204102602
204102603
204102604
204102605
204102606
204102607
204102608
204102609
204102610
204102611
204102612
204102613
204102614
204102615
204102616
204102617
204102618
204102619
204102620
204102621
204102622
204102623
204102624
204102625
204102626
204102627
204102628
204102629
204102630
204102631
204102632
204102633
204102634
204102635
204102636
204102637
204102638
204102639
204102640
204102641
204102642
204102643
204102644
204102645
204102646
204102647
204102648
204102649
204102650
204102651
204102652
204102653
204102654
204102655
204102656
204102657
204102658
204102659
204102660
204102661
204102662
204102663
204102664
204102665
204102666
204102667
204102668
204102669
204102670
204102671
204102672
204102673
204102674
204102675
204102676
204102677
204102678
204102679
204102680
204102681
204102682
204102683
204102684
204102685
204102686
204102687
204102688
204102689
204102690
204102691
204102692
204102693
204102694
204102695
204102696
204102697
204102698
204102699
204102699
204102700
204102701
204102702
204102703
204102704
204102705
204102706
204102707
204102708
204102709
204102710
204102711
204102712
204102713
204102714
204102715
204102716
204102717
204102718
204102719
204102720
204102721
204102722
204102723
204102724
204102725
204102726
204102727
204102728
204102729
204102730
204102731
204102732
204102733
204102734
204102735
204102736
204102737
204102738
204102739
204102740
204102741
204102742
204102743
204102744
204102745
204102746
204102747
204102748
204102749
204102750
204102751
204102752
204102753
204102754
204102755
204102756
204102757
204102758
204102759
204102760
204102761
204102762
204102763
204102764
204102765
204102766
204102767
204102768
204102769
204102770
204102771
204102772
204102773
204102774
204102775
204102776
204102777
204102778
204102779
204102780
204102781
204102782
204102783
204102784
204102785
204102786
204102787
204102788
204102789
204102790
204102791
204102792
204102793
204102794
204102795
204102796
204102797
204102798
204102799
204102799
204102800
204102801
204102802
204102803
204102804
204102805
204102806
204102807
204102808
204102809
204102810
204102811
204102812
204102813
204102814
204102815
204102816
204102817
204102818
204102819
204102820
204102821
204102822
204102823
204102824
204102825
204102826
204102827
204102828
204102829
204102830
204102831
204102832
204102833
204102834
204102835
204102836
204102837
204102838
204102839
204102840
204102841
204102842
204102843
204102844
204102845
204102846
204102847
204102848
204102849
204102850
204102851
204102852
204102853
204102854
204102855
204102856
204102857
204102858
204102859
204102860
204102861
204102862
204102863
204102864
204102865
204102866
204102867
204102868
204102869
204102870
204102871
204102872
204102873
204102874
204102875
204102876
204102877
204102878
204102879
204102880
204102881
204102882
204102883
204102884
204102885
204102886
204102887
204102888
204102889
204102890
204102891
204102892
204102893
204102894
204102895
204102896
204102897
204102898
204102899
204102899
204102900
204102901
204102902
204102903
204102904
204102905
204102906
204102907
204102908
204102909
204102910
204102911
204102912
204102913
204102914
204102915
204102916
204102917
204102918
204102919
204102920
204102921
204102922
204102923
204102924
204102925
204102926
204102927
204102928
204102929
204102930
204102931
204102932
204102933
204102934
204102935
204102936
204102937
204102938
204102939
204102940
204102941
204102942
204102943
204102944
204102945
204102946
204102947
204102948
204102949
204102950
204102951
204102952
204102953
204102954
204102955
204102956
204102957
204102958
204102959
204102960
204102961
204102962
204102963
204102964
204102965
204102966
204102967
204102968
204102969
204102970
204102971
204102972
204102973
20410
```

```

1674
1675  ---
1676  ### Original Base Prompt Template:
1677  {self.base_prompt_template}
1678  ---
1679
1680  ### Enhancement Requirements:
1681
1682 1. **Domain-Oriented Content Enhancement**
1683  - Add domain-specific context to the original prompt based on the
1684  given domain (e.g., finance, healthcare, technology, education).
1685  - Clarify how knowledge, attitudes, and behaviors are typically
1686  expressed in this domain.
1687  - Emphasize linguistic features or response patterns that indicate
1688  high vs. low proficiency levels within this domain.
1689
1690 2. **Clarify the Meaning of Each Condition**
1691  - For each condition, provide a clear explanation: What is being
1692  assessed?
1693  - Describe the underlying cognitive, behavioral, or psychological
1694  mechanism behind that dimension.
1695  - Identify typical language or behavior patterns respondents may
1696  exhibit in this condition.
1697
1698 3. **Define Semantic Difference Focus per Dimension**
1699  - Specify what aspects of meaning should be analyzed for semantic
1700  differences within this question and its conditions.
1701  - Example: In a financial literacy dimension, focus on logical
1702  reasoning vs. intuitive judgment, or the use of technical
1703  terminology vs. everyday language.
1704  - In a health behavior dimension, focus on risk awareness, self-
1705  management capability, or trust in scientific information.
1706  - Explain why these differences matter and what they reveal about the
1707  respondent.
1708
1709 4. **Ensure Fully Independent Evaluation Across Conditions**
1710  - Conduct a complete analysis for each condition separately, including
1711  profiling, comparison, scoring, and justification.
1712  - Avoid cross-condition interference or influence in scoring or
1713  interpretation.
1714  - Ensure each condition's output is fully independent and self-
1715  contained.
1716
1717  ---
1718  # Optional Example Support:
1719  To help the model better understand the task, you may provide example
1720  responses from two users to assist in generating a more targeted
1721  analysis framework.
1722
1723  Please follow the above instructions to enhance the base prompt with **
1724  minimal structural changes**, focusing only on **domain-relevant
1725  expansions** and **semantic-difference enhancements**.
1726  """
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2790
2791
2792
2793
2794
2795
2796
2797
2798
2799
2800
2801
2802
2803
2804
2805
2806
2807
2808
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2860
2861
2862
2863
2864
2865
2866
2867
2868
2869
2870
2871
2872
2873
2874
2875
2876
2877
2878
2879
2880
2881
2882
2883
2884
2885
2886
2887
2888
2889
2890
2891
2892
2893
2894
2895
2896
2897
2898
2899
2900
2901
2902
2903
2904
2905
2906
2907
2908
2909
2910
2911
2912
2913
2914
2915
2916
2917
2918
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969
2970
2971
2972
2973
2974
2975
2976
2977
2978
2979
2980
2981
2982
2983
2984
2985
2986
2987
2988
2989
2990
2991
2992
2993
2994
2995
2996
2997
2998
2999
3000
3001
3002
3003
3004
3005
3006
3007
3008
3009
3010
3011
3012
3013
3014
3015
3016
3017
3018
3019
3020
3021
3022
3023
3024
3025
3026
3027
3028
3029
3030
3031
3032
3033
3034
3035
3036
3037
3038
3039
3040
3041
3042
3043
3044
3045
3046
3047
3048
3049
3050
3051
3052
3053
3054
3055
3056
3057
3058
3059
3060
3061
3062
3063
3064
3065
3066
3067
3068
3069
3070
3071
3072
3073
3074
3075
3076
3077
3078
3079
3080
3081
3082
3083
3084
3085
3086
3087
3088
3089
3090
3091
3092
3093
3094
3095
3096
3097
3098
3099
3100
3101
3102
3103
3104
3105
3106
3107
3108
3109
3110
3111
3112
3113
3114
3115
3116
3117
3118
3119
3120
3121
3122
3123
3124
3125
3126
3127
3128
3129
3130
3131
3132
3133
3134
3135
3136
3137
3138
3139
3140
3141
3142
3143
3144
3145
3146
3147
3148
3149
3150
3151
3152
3153
3154
3155
3156
3157
3158
3159
3160
3161
3162
3163
3164
3165
3166
3167
3168
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3180
3181
3182
3183
3184
3185
3186
3187
3188
3189
3190
3191
3192
3193
3194
3195
3196
3197
3198
3199
3200
3201
3202
3203
3204
3205
3206
3207
3208
3209
3210
3211
3212
3213
3214
3215
3216
3217
3218
3219
3220
3221
3222
3223
3224
3225
3226
3227
3228
3229
3230
3231
3232
3233
3234
3235
3236
3237
3238
3239
3240
3241
3242
3243
3244
3245
3246
3247
3248
3249
3250
3251
3252
3253
3254
3255
3256
3257
3258
3259
3260
3261
3262
3263
3264
3265
3266
3267
3268
3269
3270
3271
3272
3273
3274
3275
3276
3277
3278
3279
3280
3281
3282
3283
3284
3285
3286
3287
3288
3289
3290
3291
3292
3293
3294
3295
3296
3297
3298
3299
3300
3301
3302
3303
3304
3305
3306
3307
3308
3309
3310
3311
3312
3313
3314
3315
3316
3317
3318
3319
3320
3321
3322
3323
3324
3325
3326
3327
3328
3329
3330
3331
3332
3333
3334
3335
3336
3337
3338
3339
3340
3341
3342
3343
3344
3345
3346
3347
3348
3349
3350
3351
3352
3353
3354
3355
3356
3357
3358
3359
3360
3361
3362
3363
3364
3365
3366
3367
3368
3369
3370
3371
3372
3373
3374
3375
3376
3377
3378
3379
3380
3381
3382
3383
3384
3385
3386
3387
3388
3389
3390
3391
3392
3393
3394
3395
3396
3397
3398
3399
3400
3401
3402
3403
3404
3405
3406
3407
3408
3409
3410
3411
3412
3413
3414
3415
3416
3417
3418
3419
3420
3421
3422
3423
3424
3425
3426
3427
3428
3429
3430
3431
3432
3433
3434
3435
3436
3437
3438
3439
3440
3441
3442
3443
3444
3445
3446
3447
3448
3449
3450
3451
3452
3453
3454
3455
3456
3457
3458
3459
3460
3461
3462
3463
3464
3465
3466
3467
3468
3469
3470
3471
3472
3473
3474
3475
3476
3477
3478
3479
3480
3481
3482
3483
3484
3485
3486
3487
3488
3489
3490
3491
3492
3493
3494
3495
3496
3497
3498
3499
3500
3501
3502
3503
3504
3505
3506
3507
3508
3509
3510
3511
3512
3513
3514
3515
3516
3517
3518
3519
3520
3521
3522
3523
3524
3525
3526
3527
3528
3529
3530
3531
3532
3533
3534
3535
3536
3537
3538
3539
3540
3541
3542
3543
3544
3545
3546
3547
3548
3549
3550
3551
3552
3553
3554
3555
3556
3557
3558
3559
3560
3561
3562
3563
3564
3565
3566
3567
3568
3569
3570
3571
3572
3573
3574
3575
3576
3577
3578
3579
3580
3581
3582
3583
3584
3585
3586
3587
3588
3589
3590
3591
3592
3593
3594
3595
3596
3597
3598
3599
3600
3601
3602
3603
3604
3605
3606
3607
3608
3609
3610
3611
3612
3613
3614
3615
3616
3617
3618
3619
3620
3621
3622
3623
3624
3625
3626
3627
3628
3629
3630
3631
3632
3633
3634
3635
3636
3637
3638
3639
3640
3641
3642
3643
3644
3645
3646
3647
3648
3649
3650
3651
3652
3653
3654
3655
3656
3657
3658
3659
3660
3661
3662
3663
3664
3665
3666
3667
3668
3669
3670
3671
3672
3673
3674
3675
3676
3677
3678
3679
3680
3681
3682
3683
3684
3685
3686
3687
3688
3689
3690
3691
3692
3693
3694
3695
3696
3697
3698
3699
3700
3701
3702
3703
3704
3705
3706
3707
3708
3709
3710
3711
3712
3713
3714
3715
3716
3717
3718
3719
3720
3721
3722
3723
3724
3725
3726
3727
3728
3729
3730
3731
3732
3733
3734
3735
3736
3737
3738
3739
3740
3741
3742
3743
3744
3745
3746
3747
3748
3749
3750
3751
3752
3753
3754
3755
3756
3757
3758
3759
3760
3761
3762
3763
3764
3765
3766
3767
3768
3769
3770
3771
3772
3773
3774
3775
3776
3777
3778
3779
3780
3781
3782
3783
3784
3785
3786
3787
3788
3789
3790
3791
3792
3793
3794
3795
3796
3797
3798
3799
3800
3801
3802
3803
3804
3805
3806
3807
3808
3809
3810
3811
3812
3813
3814
3815
3816
3817
3818
3819
3820
3821
3822
3823
3824
3825
3826
3827
3828
3829
3830
3831
3832
3833
3834
3835
3836
3837
3838
3839
3840
3841
3842
3843
3844
3845
3846
3847
3848
3849
3850
3851
3852
3853
3854
3855
3856
3857
3858
3859
3860
3861

```

1728 **User Prompt:**
1729
1730 f"""
1731 ---
1732 You are tasked with modifying an initial prompt based on specific user
1733 feedback.
1734 ---
1735 **## OBJECTIVE:**
1736 Refine the original prompt according to the user's explicit instructions,
1737 making only the minimal necessary changes to address the feedback.
1738 ---
1739 **## REQUIREMENTS:**
1740 1. Focus solely on the issues explicitly mentioned in the user feedback.
1741 2. Make the smallest possible adjustments needed to align with user
1742 instructions.
1743 3. Preserve all other aspects of the original prompt that are not
1744 directly addressed in the feedback.
1745 4. If user-suggested modifications are provided, treat them as in-context
1746 examples for adjustment.
1747 5. Ensure the modified prompt remains clear, functional, and domain-
1748 agnostic.
1749 ---
1750 **## ORIGINAL PROMPT:**
1751 {original_prompt}
1752 ---
1753 **## USER FEEDBACK:**
1754 {user_feedback}
1755 ---
1756 Please provide the optimized prompt directly. Do not alter anything
1757 beyond what is requested by the user.
1758 Only return the optimized prompt content, no additional explanation or
1759 formatting.
1760 """

E.2.3 EXAMPLE INSTRUCTION FOR LLM-BASED EVALUATOR

System Prompt:

"You are an expert in analyzing differences between survey respondents. Your task is to evaluate how different two people are based on their responses, focusing on their characteristics, knowledge, and behavioral patterns. You should provide difference scores where higher scores indicate greater differences between respondents."

User Prompt:

```
1772
1773     """
1774     # USER'S INSTRUCTIONS (PLEASE FOLLOW THE INSTRUCTIONS TO GIVE RESULTS):
1775     {agent1_prompt}
1776
1777     ### RESPONDENT 1:
1778     {response1.get('combined_answer', '')}
1779
1780     ### RESPONDENT 2:
1781     {response2.get('combined_answer', '')}
1782
1783     ### Conditions to evaluate:
1784     {conditions_text}
```

```

1782
1783 # ATTENTION: Please follow the user's instructions and ensure the output
1784     adheres to the specified format provided in the function call.
1785 """
1786

```

1787 E.3 PROMPTS FOR DATA GENERATION SYSTEM

1788 This section provides the prompts used in the data generation system.

1791 E.3.1 INSTRUCTION FOR COMMUNITY DEFINITION

1793 System Prompt:

```

1794 "You are an expert in survey analysis and demographic segmentation.
1795 Analyze the given questionnaire to identify distinct community types that
1796     would respond differently based on their knowledge, behavior, and
1797     attitudes in the specific domain.
1798
1799 For each community, provide:
1800 1. A descriptive name that reflects their characteristics
1801 2. Score ranges (0-100) for each evaluation dimension that represent
1802     their expected proficiency levels
1803
1804 Consider the domain context and ensure communities represent meaningful
1805     diversity in responses."
1806

```

1805 User Prompt:

```

1806 f"""Analyze this {questionnaire_name} questionnaire to identify exactly 3
1807     distinct community types who would interact with this domain
1808     differently.
1809
1810 QUESTIONNAIRE DETAILS:
1811 - Domain: {questionnaire_name}
1812 - Evaluation Dimensions: {', '.join(dimensions_list)}
1813 - Content Overview: {questionnaire_content}
1814
1815 ANALYSIS REQUIREMENTS:
1816 Please identify exactly 3 communities representing high, medium, and low
1817     proficiency levels in this {questionnaire_name} domain. For each
1818     community:
1819
1820 1. Name: Use format "level_domain_descriptor" (e.g., "high_financial_literacy" or "expert_healthcare_professionals")
1821
1822 2. Description (150+ words): Include:
1823     - Demographics (age range, education level, occupation types)
1824     - Domain experience level and exposure history
1825     - Behavioral patterns and interaction styles
1826     - Attitudes, motivations, and pain points
1827     - Knowledge depth and breadth characteristics
1828     - Representative examples of people in this community
1829     - How they typically approach problems in this field
1830
1831 3. Score Ranges:
1832     - For each dimension (', '.join(dimensions_list)), assign
1833         appropriate score ranges (0-100)
1834     - High proficiency community: typically 69-100
1835     - Medium proficiency community: typically 31-69
1836     - Low proficiency community: typically 1-30
1837
1838 DOMAIN-SPECIFIC GUIDANCE:
1839 - For Finance: Consider financial literacy levels, risk attitudes,
1840     financial behaviors

```

1836 - For HCI: Consider technical expertise, interaction fluency, adaptive
 1837 technology use
 1838 - For Manufacturing/3D Printing: Consider technical understanding,
 1839 practical experience, creative application
 1840 - For medical questionnaires, DO NOT create communities based on medical
 1841 knowledge levels or professional backgrounds.
 1842 Instead, focus on RECOVERY STATUS and POST-TREATMENT OUTCOMES:
 1843 - Excellent Recovery Community: Patients who have recovered well from
 1844 treatment/surgery with minimal ongoing symptoms
 1845 - Moderate Recovery Community: Patients with partial recovery, still
 1846 managing some symptoms
 1847 - Poor Recovery Community: Patients with slow/incomplete recovery,
 1848 significant ongoing symptoms
 1849
 1850 The communities should represent different healing outcomes and
 1851 adaptation levels, NOT different medical expertise levels.
 1852 All patients should have similar baseline medical knowledge (typical
 1853 patient level).
 1854
 1855 Call the analyze_questionnaire_communities function with your detailed
 1856 analysis results.
 1857

E.3.2 INSTRUCTION FOR PERSONA INSTANTIATION

System Prompt:

1860 f"""\n 1861 You are an expert in creating realistic and diverse candidate
 1862 profiles based on detailed community characteristics and specific
 1863 proficiency scores.\n 1864 Your task is to generate {candidates_per_community} distinct candidate
 1865 profiles for a given community in the {domain} domain. Each candidate
 1866 has been pre-assigned specific proficiency scores that you MUST use
 1867 to guide their background creation.\n 1868
 1869 CORE PRINCIPLES:\n 1870 1. **Score-Driven Background Creation**: Use the provided scores to
 1871 determine the candidate's expertise level and create a background
 1872 that logically explains these scores\n 1873 2. **Community Authenticity**: Ensure each candidate authentically
 1874 reflects the community characteristics described\n 1875 3. **Occupational Diversity**: Generate diverse occupations based on the
 1876 community description rather than predefined lists\n 1877 4. **Individual Coherence**: Each candidate should have a coherent life
 1878 story where their demographics, education, career, and experience
 1879 logically support their proficiency scores\n 1880
 1881 SCORE INTERPRETATION GUIDELINES:\n 1882 - 80-100: Expert level - extensive education, senior positions, many
 1883 years of experience, specialized training\n 1884 - 60-79: Advanced level - solid education, mid-senior positions,
 1885 considerable experience, some specialization\n 1886 - 40-59: Intermediate level - moderate education, mid-level positions,
 1887 some experience, general knowledge\n 1888 - 20-39: Basic level - basic education, junior positions, limited
 1889 experience, foundational knowledge\n 1890 - 0-19: Novice level - minimal education/experience, entry-level or non-
 1891 professional roles, very limited exposure\n 1892
 1893 IMPORTANT REQUIREMENTS:\n 1894 - Create backgrounds that JUSTIFY the specific scores provided
 1895 - Generate occupations that fit the community description, not from
 1896 predefined lists

1890 - Ensure age, education, and experience align with the proficiency levels
 1891 indicated by scores
 1892 - Make each candidate distinct while maintaining community consistency
 1893 - Include realistic personal interests that complement their professional
 1894 profile"""

1895
 1896 **User Prompt:**
 1897

```
1898 f"""Generate {candidates_per_community} distinct and realistic candidate
1899     profiles for the "{community_name}" community in the {domain} domain.
1900
  COMMUNITY PROFILE:
1901 {community_description}
1902
  PRE-ASSIGNED PROFICIENCY SCORES:
1903 {scores_info}
1904
  DOMAIN: {domain}
1905 EVALUATION DIMENSIONS: {', '.join(dimensions)}
1906
  CRITICAL REQUIREMENTS:
1907
  1. **Score-Based Background Creation**:
1908     - Use the assigned scores to determine each candidate's expertise
1909         level
1910     - Create backgrounds that logically EXPLAIN and JUSTIFY these specific
1911         scores
1912     - Higher scores require more education, experience, and specialized
1913         knowledge
1914     - Lower scores should reflect limited exposure, basic education, or
1915         different career focuses
1916
  2. **Cross-Dimensional Consistency**:
1917     - IMPORTANT: Each candidate should maintain consistent proficiency
1918         levels across ALL dimensions
1919     - If a candidate has high scores (80+) in one dimension, they should
1920         have similarly high scores in other dimensions
1921     - The candidate's background, education, and experience should justify
1922         their competency across ALL evaluation dimensions
1923     - Avoid creating candidates who are experts in some dimensions but
1924         novices in others within the same profile
1925
  3. **Community Authenticity**:
1926     - Each candidate MUST reflect the community characteristics described
1927         above
1928     - Match the demographic patterns, behavioral traits, and knowledge
1929         depths specified
1930     - Embody the attitudes, motivations, and pain points mentioned in the
1931         community description
1932
  4. **Occupation Generation**:
1933     - Generate occupations based on the community description examples and
1934         characteristics
1935     - DO NOT use predefined occupation lists - create realistic jobs that
1936         fit the community
1937     - Ensure job complexity and seniority align with the proficiency
1938         scores across ALL dimensions
1939
  5. **Individual Coherence**:
1940     - Each candidate should have a coherent life story where age,
1941         education, career path, and domain exposure logically lead to
1942         their assigned scores across ALL dimensions
1943     - Explain HOW they achieved their proficiency levels through their
1944         background
```

1944 - Include realistic career progression and learning experiences that
1945 support competency in all areas
1946
1947 6. ****Diversity Within Community**:**
1948 - While maintaining community consistency and cross-dimensional
1949 coherence, create variety in specific backgrounds, locations,
1950 career paths, and personal interests
1951 - Ensure each candidate feels like a unique individual within their
1952 community type
1953 Please call the generate_community_candidates_batch function with {
1954 candidates_per_community} complete candidate profiles that
1955 authentically represent the "{community_name}" community while having
1956 backgrounds that justify their assigned proficiency scores across
all dimensions. """

E.3.3 INSTRUCTION FOR RESPONSE SIMULATION

System Prompt:

"You are an expert in analyzing differences between survey respondents. Your task is to evaluate how different two people are based on their responses, focusing on their characteristics, knowledge, and behavioral patterns. You should provide difference scores where higher scores indicate greater differences between respondents."

User Prompt:

```
"""
CANDIDATE IDENTITY:
- ID: {candidate_id}
- Community: {community}
- Personal Story: {description}

DEMOGRAPHICS:
- Age: {demographics.get('age', 'N/A')}
- Gender: {demographics.get('gender', 'N/A')}
- Location: {demographics.get('location', 'N/A')}
- Education: {demographics.get('education', 'N/A')}

PROFESSIONAL LIFE:
- Field: {professional_background.get('field', 'N/A')}
- Current Role: {professional_background.get('occupation', 'N/A')}
- Experience: {professional_background.get('years_experience', 'N/A')}
  years
- Position Level: {professional_background.get('current_position', 'N/A')}
  }
- Company Type: {professional_background.get('company_type', 'N/A')}
- Career Journey: {professional_background.get('career_progression', 'N/A')}
  '}

PERSONAL INTERESTS: {', '.join(personal_interests) if personal_interests
  else 'N/A' }

DOMAIN EXPERTISE:
- Years in {domain}: {domain_experience.get('years_in_domain', 'N/A')}
  years
- Specialization: {domain_experience.get('specialization', 'N/A')}
- Key Achievements: {'; '.join(domain_experience.get('key_achievements',
  []))} if domain_experience.get('key_achievements') else 'N/A'
- Learning Sources: {'; '.join(domain_experience.get('learning_sources',
  []))} if domain_experience.get('learning_sources') else 'N/A'

PROFICIENCY SCORES (Critical for Response Generation):
```

```

1998 {scores}
1999
2000 ##### CRITICAL INSTRUCTION #####
2001 ALWAYS PRIORITIZE ACTUAL SCORES OVER COMMUNITY/BACKGROUND DESCRIPTIONS!
2002 The scores above are the GROUND TRUTH for this candidate's actual
2003 abilities.
2004 These scores may intentionally differ from the community description or
2005 background story as part of data simulation.
2006 If there's any conflict between the community description and the actual
2007 scores, ALWAYS follow the scores.
2008 """
2009
2010
2011 E.4 DETERMINED PROMPTS FOR WORKSHOP
2012
2013 This is the prompt determined to be used in the workshop experiment.
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051

```

E.4 DETERMINED PROMPTS FOR WORKSHOP

This is the prompt determined to be used in the workshop experiment.

```

---  

2014 ##### Individual Response Behavior Change Assessment  

2015  

2016 You are an expert in analyzing individual response content to assess  

2017 behavior change levels across different dimensions/conditions within  

2018 the **health behavior change assessment domain**. This domain focuses  

2019 on evaluating changes in eating habits and their alignment with  

2020 healthy behaviors.  

2021 ---  

2022  

2023 ##### Task Overview  

2024 Your goal is to analyze a SINGLE respondent's response regarding ''eating  

2025 changes since starting college'' and provide behavior change scores  

2026 (0-100) for the condition ''healthy eating changes'', which evaluates  

2027 how much the respondent's eating habits have improved or worsened.  

2028 The primary focus is on behavioral change, not the individual's  

2029 current state.  

2030 ---  

2031  

2032 ##### Analysis Process  

2033 For the condition, follow this process:  

2034  

2035 1. **Individual Behavior Change Assessment**  

2036 - Analyze the response content to estimate the respondent's behavior  

2037 change level (0-100) based on how their eating habits have changed  

2038 .  

2039 - Look for: evidence of positive behavioral changes (e.g., "I started  

2040 eating more fruits"), negative behavioral changes (e.g., "I eat  

2041 more junk food now"), specific examples of habit modifications,  

2042 and acknowledgment of external factors influencing eating habits.  

2043 - Consider: reasoning behind changes (e.g., adapting to a healthier  

2044 lifestyle), practical steps taken (e.g., meal planning), and  

2045 awareness of health-related goals.  

2046  

2047 2. **Behavior Change Level Indicators**  

2048 - **50 (Baseline)**: No significant change in eating habits; neutral  

2049 behavior patterns without improvement or decline.  

2050 - **50-100**: Scores increase as eating habits become healthier.  

2051 Evidence includes adopting better nutritional practices, reducing  

2052 unhealthy food intake, and proactive behavior adjustments.  

2053 - **0-50**: Scores decrease as eating habits worsen. Evidence includes  

2054 increased consumption of unhealthy foods, lack of effort to  

2055 improve habits, and negative behavioral trends.

```

```

2052
2053 3. **Score Assignment**
2054   - Assign a specific behavior change score (0-100) based on evidence in
2055   the response.
2056   - Provide brief justification for the score.
2057
2058 ---
2059 ### Behavior Change Scoring Guide
2060
2061 | Score Range | Description | Typical Evidence |
2062 |-----|-----|-----|
2063 | **90-100** | **Significant improvement** | Clear adoption of highly
2064   healthy eating habits, detailed examples of positive changes (e.g., "I
2065   track my calorie intake daily"). |
2066 | **80-89** | **Strong improvement** | Solid evidence of healthier eating
2067   , structured efforts to modify habits (e.g., "I stopped eating fast
2068   food"). |
2069 | **70-79** | **Moderate improvement** | Noticeable positive changes,
2070   some specificity in actions taken (e.g., "I try to eat vegetables
2071   with every meal"). |
2072 | **60-69** | **Slight improvement** | Basic efforts toward healthier
2073   eating, limited detail but clear intent (e.g., "I'm trying to drink
2074   more water"). |
2075 | **50** | **No significant change** | Neutral behavior; no clear
2076   evidence of improvement or decline (e.g., 'My eating habits haven't
2077   really changed'). |
2078 | **40-49** | **Slight decline** | Minor negative changes, vague examples
2079   of unhealthy habits emerging (e.g., "I snack more often now"). |
2080 | **30-39** | **Moderate decline** | Clear evidence of worsening habits,
2081   less attention to health (e.g., "I eat out more frequently"). |
2082 | **20-29** | **Strong decline** | Significant adoption of unhealthy
2083   eating patterns, lack of effort to improve (e.g., "I don't care about
2084   what I eat anymore"). |
2085 | **0-19** | **Severe decline** | Major deterioration in eating habits,
2086   frequent references to unhealthy behaviors (e.g., "I only eat junk
2087   food now"). |
2088
2089 ---
2090 ### Quality Assessment Framework
2091
2092 **High Behavior Change Signals:**
2093 - Describes specific positive behavioral changes (e.g., "I started meal
2094   prepping").
2095 - Acknowledges external factors influencing habits (e.g., "I eat
2096   healthier because my roommate cooks nutritious meals").
2097 - Demonstrates awareness of health goals (e.g., "I reduced sugar intake
2098   to feel more energetic").
2099
2100 **Low Behavior Change Signals:**
2101 - Shows vague or generic statements (e.g., "I eat a bit better").
2102 - Provides examples of negative behavioral trends (e.g., "I eat late-
2103   night snacks every day").
2104 - Lacks reasoning or intentionality behind changes (e.g., "I just eat
2105   whatever is available").
2106
2107 **Assessment Priority:**
2108 - Focus on BEHAVIORAL CHANGE rather than technical knowledge.
2109 - Evaluate SPECIFIC ACTIONS taken to improve or worsen eating habits.
2110 - Consider CONTEXTUAL FACTORS influencing dietary choices.
2111
2112 ---
2113 ### Output Format

```

```

2106
2107 For the condition:
2108 1. **Condition Name**: [condition]
2109 2. **Behavior Change Score**: [0-100]
2110 3. **Key Evidence**: Brief justification (2-3 sentences max)
2111 4. **Behavior Change Level**: [Novice/Basic/Intermediate/Advanced/Expert]
2112 Remember: Base your assessment solely on demonstrated changes in eating
2113 habits, focusing on behavioral trends rather than communication style
2114 or vocabulary sophistication.
2115 ---
2116

```

2117

2118

E.5 SYNTHETIC DATA EXAMPLE ON FINANCE

2119

2120

In this section, we provide some data examples generated by the Controllable Synthetic Data Generation System for reference.

2121

2122

2123

E.5.1 COMMUNITY EXAMPLE

2124

```

2125 1 {
2126 2 ...
2127 3 {
2128 4   "name": "medium_financial_practitioner",
2129 5   "description": "These individuals are aged 22-45, with
2130     undergraduate degrees in business, economics, or related
2131     fields. Occupations range from small business owners to
2132     mid-level corporate employees. They have moderate exposure
2133     to financial practices, including basic banking services,
2134     occasional investments, and some savings habits. While
2135     they grasp fundamental financial concepts, their knowledge
2136     lacks depth in complex areas like risk diversification or
2137     high-yield strategies. Behaviorally, they check finances
2138     periodically but may not consistently monitor them. They
2139     value financial education but prioritize immediate needs
2140     over long-term planning. For instance, a marketing manager
2141     who saves occasionally and invests in low-risk mutual
2142     funds typifies this group. Problem-solving approaches are
2143     practical but less strategic.\n",
2144 6 },
2145 7 ...
2146 8
2147 9 }

```

2148

E.5.2 CANDIDATE EXAMPLE

2149

2150

2151

2152

2153

2154

2155

2156

2157

2158

2159

11

```

1 {
2 ...
3 {
4   "id": "C1",
5   "community": "high_financial_expert",
6   "assigned_scores": {
7     "financial attitude": 90,
8     "financial literacy": 80,
9     "financial behavior": 81
10 },
11   "description": "Candidate 1 is a seasoned Investment
12     Portfolio Manager with over 15 years of experience

```

```

2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176 12
2177 13
2178 14
2179 15
2180 16
2181 17
2182 18
2183 19
2184 20
2185 21
2186 22
2187 23
2188 24
2189 25
2190 26
2191 27
2192 28
2193 29
2194 30
2195 31
2196 32
2197 33
2198 34
2199 35
2200 36
2201 37
2202 38
2203 39
2204 40
2205 41
2206 42
2207 43
2208 44
2209 45
2210 46
2211 47
2212 48

    managing high-net-worth client accounts at an elite
    private wealth management firm. Holding both an MBA in
    Finance and a CFA charter, she has developed a robust
    understanding of financial markets and excels in
    constructing diversified portfolios that maximize
    returns while mitigating risks. Her financial attitude
    score of 90 reflects her disciplined approach to wealth
    accumulation and risk management, always prioritizing
    long-term stability over short-term gains. Despite
    scoring slightly lower in financial literacy (80) and
    financial behavior (81), she remains highly
    knowledgeable, though some gaps exist in advanced
    derivatives trading. She regularly attends investment
    seminars and subscribes to financial journals to stay
    updated on emerging trends. Outside work, she enjoys
    mentoring young women entering finance and participating
    in triathlons.",

    "scores": {
        "financial attitude": 90,
        "financial literacy": 80,
        "financial behavior": 81
    },
    "demographics": {
        "age": 42,
        "gender": "Female",
        "location": "New York City, USA",
        "education": "MBA in Finance, CFA Charterholder"
    },
    "professional_background": {
        "field": "Investment Management",
        "occupation": "Senior Investment Portfolio Manager",
        "years_experience": 15,
        "current_position": "Managing Director",
        "company_type": "Private Wealth Management Firm",
        "career_progression": "Started as an equity research
            analyst, then transitioned into portfolio management
            roles, steadily climbing the ranks to become a senior
            manager."
    },
    "personal_interests": [
        "Mentoring young professionals",
        "Participating in endurance sports",
        "Reading financial literature"
    ],
    "domain_experience": {
        "years_in_domain": 15,
        "specialization": "Wealth Management and Asset Allocation"
    },
    "key_achievements": [
        "Increased managed assets under her division by 35% in
            three years",
        "Developed proprietary portfolio optimization model",
        "Published articles in top finance magazines"
    ],
    "learning_sources": [
        "CFA curriculum",
        "Financial Times",
        "Annual investment conferences"
    ]
}

```



```

2268 97     ...
2269 98 }
2270
2271
2272 E.5.3 RESPONSE EXAMPLE
2273
2274 1 {
2275 2     ...
2276 3     {
2277 4         "community": "high_financial_expert",
2278 5         "candidate_id": "C1",
2279 6         "candidate_demographics": {
2280 7             "age": 42,
2281 8             "gender": "Female",
2282 9             "location": "New York City, USA",
2283 10            "education": "MBA in Finance, CFA Charterholder"
2284 11        },
2285 12        "candidate_professional_background": {
2286 13            "field": "Investment Management",
2287 14            "occupation": "Senior Investment Portfolio Manager",
2288 15            "years_experience": 15,
2289 16            "current_position": "Managing Director",
2290 17            "company_type": "Private Wealth Management Firm",
2291 18            "career_progression": "Started as an equity research analyst
2292 19                , then transitioned into portfolio management roles,
2293 20                steadily climbing the ranks to become a senior manager."
2294 21        },
2295 22        "candidate_personal_interests": [
2296 23            "Mentoring young professionals",
2297 24            "Participating in endurance sports",
2298 25            "Reading financial literature"
2299 26        ],
2300 27        "candidate_domain_experience": {
2301 28            "years_in_domain": 15,
2302 29            "specialization": "Wealth Management and Asset Allocation",
2303 30            "key_achievements": [
2304 31                "Increased managed assets under her division by 35% in
2305 32                    three years",
2306 33                "Developed proprietary portfolio optimization model",
2307 34                "Published articles in top finance magazines"
2308 35            ],
2309 36            "learning_sources": [
2310 37                "CFA curriculum",
2311 38                "Financial Times",
2312 39                "Annual investment conferences"
2313 40            ],
2314 41        },
2315 42        "question_id": "Q1",
2316 43        "main_question": "Questions related to banking services",
2317 44        "answers": [
2318 45            {
2319 46                "dimension": "financial behavior",
2320 47                "question": "How often do you engage in banking services?
2321 48                    What types of services do you typically use at the
2322 49                    bank?",
2323 50                "response": "
2324 51                    As a Senior Investment Portfolio Manager with over 15
2325 52                        years of experience in wealth management, I
2326 53                        engage with banking services quite regularly, but

```

2322 the frequency and type of services I use are
 2323 largely driven by both personal needs and
 2324 professional obligations. My interactions tend to
 2325 be more strategic than transactional because my
 2326 financial behavior leans toward optimizing long-
 2327 term outcomes rather than handling day-to-day
 2328 banking tasks.
 2329 ⁴⁷
 2330 ⁴⁸ On a personal level, I primarily use digital banking
 2331 platforms for routine activities like bill
 2332 payments, fund transfers, and monitoring account
 2333 balances. These services save time and align well
 2334 with my disciplined approach to managing finances
 2335 efficiently. However, given my role at the private
 2336 wealth management firm, I also frequently
 2337 interact with relationship managers at high-end
 2338 banks for specialized services such as wire
 2339 transfers for international investments, liquidity
 2340 management solutions, or structuring credit
 2341 facilities that support portfolio strategies. For
 2342 instance, when deploying capital across different
 2343 asset classes—whether it's equities, fixed income,
 2344 or alternative investments—I need access to
 2345 institutional-grade cash management tools to
 2346 ensure seamless execution.
 2347 ⁴⁹
 2348 From a professional standpoint, I rely on custodial
 2349 banking services to handle the operational aspects
 2350 of client portfolios. This includes trade
 2351 settlements, dividend collection, tax reporting,
 2352 and performance measurement. Over the years, I've
 2353 developed a strong preference for banks that offer
 2354 robust technological integrations and real-time
 2355 data analytics capabilities. Such features allow
 2356 me to monitor risk exposures and make informed
 2357 decisions quickly. That said, there are still
 2358 areas where I notice gaps—for example, some banks
 2359 could improve their offerings around derivatives-
 2360 related margining requirements or tailored FX
 2361 hedging products, which is something I
 2362 occasionally encounter in my work.
 2363 ⁵¹
 2364 I'd rate myself as highly knowledgeable about
 2365 leveraging these banking services effectively,
 2366 though I recognize there's room to deepen my
 2367 expertise further in certain niche areas, such as
 2368 advanced treasury solutions or blockchain-based
 2369 innovations in banking. Attending annual
 2370 investment conferences and staying current with
 2371 publications like **Financial Times** helps me keep
 2372 up with emerging trends, but they haven't all
 2373 translated into direct application yet.
 2374 ⁵³
 2375 To summarize, I probably interact with banking
 2376 services weekly for basic functions and monthly
 2377 for more sophisticated ones. The mix reflects my
 2378 dual focus on maintaining operational efficiency
 2379 while pursuing strategic opportunities that

enhance value creation—not just for myself but also for the high-net-worth clients I serve."},
"dimension_score": 81},
{
"dimension": "financial literacy",
"question": "What risks do you think are associated with keeping money in the bank?",
"response":
"Well, as someone who's spent over 15 years managing wealth for high-net-worth clients, I can say that keeping money in the bank does come with its own set of risks—though they might not be as obvious as some of the more volatile investments. From my perspective, there are three key risks to consider: inflation risk, opportunity cost, and credit risk.

First, **inflation risk** is a big one. When you leave money sitting in a savings or checking account, especially in today's low-interest-rate environment, the returns you get often don't keep up with inflation. Over time, this erodes your purchasing power. For instance, if your bank is offering 0.5% interest on savings but inflation is running at 3%, you're effectively losing ground. This is something I always emphasize with my clients when we talk about long-term wealth preservation—it's not just about protecting capital, it's about making sure it grows in real terms.

Second, there's the **opportunity cost** of not putting that money to work in other assets. As an investment professional with expertise in asset allocation, I know there are many ways to generate better returns while still maintaining a reasonable level of risk. Whether it's through diversified portfolios of equities, bonds, or even alternative investments like real estate or private equity, the potential for higher returns exists outside traditional banking products. Of course, these options come with their own risks, but the trade-off is worth considering depending on your financial goals and time horizon.

Finally, there's **credit risk**, which refers to the possibility that the bank itself could run into trouble. Now, in most developed markets like the U.S., deposit insurance programs such as FDIC coverage protect consumers up to certain limits (currently \$250,000 per account). But if you have significantly more than that parked in a single institution, you need to think carefully about diversifying across banks or using other structures, like trusts or custodial accounts, to mitigate this risk. It's a point I've discussed with clients who prefer to hold large cash

balances—they sometimes underestimate how exposed they could be without proper planning.

That said, I also recognize the value of liquidity and safety that banks provide, particularly for short-term needs or emergency funds. In my own practice, I advise clients to strike a balance—keeping enough in liquid, low-risk accounts to cover near-term obligations while deploying the rest into growth-oriented strategies. After all, wealth management isn't about chasing returns blindly; it's about aligning your resources with your objectives in a disciplined way.

So, to sum it up, while banks offer security and convenience, the risks of inflation, missed opportunities, and institutional vulnerability should never be ignored. These are principles I've both studied extensively—through my MBA and CFA training—and applied practically in managing portfolios over the years.",

"dimension_score": 80

}

],

"combined_answer":

"As a Senior Investment Portfolio Manager with over 15 years of experience in wealth management, I engage with banking services regularly, though my interactions are often more strategic than transactional. On a personal level, I use digital banking platforms for routine tasks like bill payments, fund transfers, and monitoring account balances—tools that align with my disciplined approach to efficiency. Professionally, however, my reliance on banking services becomes much more nuanced. For example, I frequently work with relationship managers at high-end banks to facilitate wire transfers for international investments, structure credit facilities, or manage liquidity solutions that support portfolio strategies. Custodial banking services are also critical in my day-to-day operations, as they handle trade settlements, dividend collection, tax reporting, and performance measurement for client accounts. Over time, I've developed a preference for banks that offer robust technological integrations and real-time data analytics capabilities because these features allow me to monitor risk exposures and make informed decisions quickly. That said, there are still areas where I notice gaps, such as advanced treasury solutions or tailored FX hedging products, which I occasionally encounter but don't yet feel fully equipped to navigate. While I stay updated on emerging trends through investment conferences and financial journals, some innovations, like blockchain-based banking tools, haven't translated into direct application for me yet.

At the same time, I'm acutely aware of the risks associated with keeping money in the bank.

2484
 2485
 2486
 2487
 2488
 2489
 2490
 2491
 2492
 2493
 2494
 2495
 2496
 2497
 2498
 2499
 2500
 2501
 2502
 2503
 2504
 2505
 2506
 2507
 2508
 2509
 2510
 2511
 2512
 2513
 2514
 2515
 2516
 2517
 2518
 2519
 2520
 2521
 2522
 2523
 2524
 2525
 2526
 2527
 2528
 2529
 2530
 2531
 2532
 2533
 2534
 2535
 2536
 2537
 particularly when it comes to inflation, opportunity cost, and credit risk. Inflation is a big concern—it's something I emphasize with clients when discussing long-term wealth preservation. If your returns from a savings account don't outpace inflation, you're effectively losing purchasing power over time. Then there's the opportunity cost of not deploying that capital into growth-oriented assets, whether it's equities, bonds, or alternatives like real estate. Of course, these options come with their own risks, but balancing safety with growth is key. Finally, there's credit risk—the possibility that the bank itself could face instability. While programs like FDIC insurance provide a layer of protection, I always advise clients with significant cash holdings to diversify across institutions or explore other structures to mitigate this risk. Banks offer undeniable value in terms of liquidity and security, especially for short-term needs, but they're not without their limitations. Striking the right balance between safety and growth is something I prioritize both personally and professionally, even if it means acknowledging that no solution is entirely without trade-offs.",
 "scores": {
 "financial attitude": 90,
 "financial literacy": 80,
 "financial behavior": 81
 },
 ...
 {
 "community": "low_financial_novice",
 "candidate_id": "C8",
 "candidate_demographics": {
 "age": 23,
 "gender": "Female",
 "location": "Denver, Colorado",
 "education": "High school diploma"
 },
 "candidate_professional_background": {
 "field": "Gig Economy",
 "occupation": "Delivery Driver",
 "years_experience": 2,
 "current_position": "Contractor",
 "company_type": "Food Delivery Platform",
 "career_progression": "Initially drove fewer hours per week; increased workload as demand grew."
 },
 "candidate_personal_interests": [
 "Trying new foods",
 "Travel vlogs",
 "Photography"
],
 "candidate_domain_experience": {
 "years_in_domain": 2,
 "specialization": "Using mobile budgeting apps",
 "key_achievements": [
 "Consistently saved \$100/month for three consecutive months"

```

2538113 ],
2539114 "learning_sources": [
2540115   "Peer recommendations",
2541116   "Mobile app tutorials"
2542117 ]
2543118 },
2544119 "question_id": "Q1",
2545120 "main_question": "Questions related to banking services",
2546121 "answers": [
2547122   {
2548123     "dimension": "financial behavior",
2549124     "question": "How often do you engage in banking services?
2550125       What types of services do you typically use at the
2551126       bank?",
2552127     "response": "
2553128       I don't really go to the bank all that often, to be
2554129       honest. I know they offer services like checking
2555130       accounts and savings accounts-I've got both-but
2556131       from what I understand, most of what I need can be
2557132       handled through their app. That's something I've
2558133       been using more since it's easier for me with my
2559134       schedule. I think you can deposit checks through
2560135       the app too, which is helpful because I'm always
2561136       on the move delivering food.
2562137
2563138       From what I've heard, banks also do things like loans
2564139       or help with credit cards, but I haven't really
2565140       looked into those much yet. I've thought about
2566141       maybe getting a better understanding of how credit
2567142       works, though, since I want to eventually move
2568143       out of my shared apartment and maybe even start
2569144       saving up for a car. For now, I mostly just use
2570145       the basics-checking my balance, transferring money
2571146       between accounts when I remember to save a little
2572147       each month, stuff like that.
2573148
2574149       I guess if I had more time, I'd try to learn more
2575150       about other services, but right now, I'm still
2576151       figuring out the essentials. My shifts keep me
2577152       pretty busy, so I stick to what I know and what's
2578153       quick to manage online.",
2579154     "dimension_score": 35
2580155   },
2581156   {
2582157     "dimension": "financial literacy",
2583158     "question": "What risks do you think are associated with
2584159       keeping money in the bank?",
2585160     "response": "
2586161       I think there are some risks with keeping money in
2587162       the bank, but I don't know too much about it. From
2588163       what I understand, one risk might be that if the
2589164       bank has problems or goes out of business, you
2590165       could lose your money-but I've heard something
2591166       about insurance that's supposed to protect it? I'm
2592167       not really sure how that works, though.
2593168
2594169       Another thing I've thought about is inflation. Like,
2595170       if I keep my savings in a regular account, I think
2596171       the value of my money might go down over time

```

2592 because prices for stuff go up. But honestly, I
 2593 don't fully get how that connects to banks or what
 2594 I can do about it. It's just something I've heard
 2595 people mention.
 2596¹⁴⁰
 2597¹⁴¹ Also, I guess there's a chance someone could hack into
 2598 your account or steal your information if the
 2599 bank's security isn't good enough. That makes me
 2600 nervous sometimes when I use mobile banking apps,
 2601 even though they seem pretty safe so far. I try to
 2602 check my accounts often, just in case anything
 2603 weird happens.
 2604¹⁴²
 2605¹⁴³ Overall, I know these things exist, but I don't feel
 2606 super confident explaining them. I've been using
 2607 budgeting apps to track my money, and they help me
 2608 focus on saving small amounts each month instead
 2609 of worrying too much about bigger risks like this
 2610¹⁴⁴ .",
 2611¹⁴⁵ "dimension_score": 30
 2612¹⁴⁶ }
 2613¹⁴⁷],
 2614¹⁴⁸ "combined_answer":
 2615 "I don't really go to the bank all that often, to be
 2616 honest. Most of what I need can be handled through
 2617 their app, which is super helpful for me since my
 2618 delivery shifts keep me so busy. I've got a checking
 2619 account and a savings account, and I use them mostly
 2620 for the basics-checking my balance, transferring money
 2621 between accounts when I remember to save a little
 2622 each month, stuff like that. I know banks offer other
 2623 services, like loans or credit cards, and I've thought
 2624 about trying to understand credit better because I
 2625 want to eventually move out of my shared apartment and
 2626 maybe even start saving up for a car. But for now, I
 2627 stick to what I know and what's quick to manage online
 2628 .
 2629¹⁴⁹
 2630 From what I've heard, there are some risks with keeping
 2631 money in the bank, but I'll admit I don't fully
 2632 understand them yet. Like, if the bank has problems or
 2633 goes out of business, I think you could lose your
 2634 money-but then I've also heard something about
 2635 insurance that's supposed to protect it? I'm not sure
 2636 how that works exactly. Another thing I've thought
 2637 about is inflation. If I keep my savings in a regular
 2638 account, I feel like the value of my money might go
 2639 down over time as prices go up, but I don't really get
 2640 how that connects to banks or what I can do about it.
 2641 And yeah, there's always the worry about someone
 2642 hacking into your account or stealing your information
 2643¹⁵¹, especially when I'm using mobile banking apps on the
 2644¹⁵² go. That makes me nervous sometimes, even though they
 2645 seem pretty secure so far. I try to check my accounts
 often just in case anything weird happens.
 2646¹⁵³
 2647 Overall, I know these risks exist, but I don't feel super
 2648 confident explaining them. I've been experimenting
 2649 with budgeting apps to track my money, and they help

```

2646
2647
2648
2649
2650
2651  me focus on saving small amounts each month instead of
2652  stressing too much about bigger risks. For now, I'm
2653  still figuring out the essentials, and while I'd love
2654  to learn more, my schedule doesn't leave much room for
2655  diving deeper into financial strategies.",  

2656  "scores": {
2657    "financial attitude": 26,
2658    "financial literacy": 30,
2659    "financial behavior": 35
2660  },
2661  ...
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699

```