Cognitive Bias and Reassignment: Who Can Contribute High Quality LLM Data

Yunfan Gao[1], Yun Xiong[2], Zhongyuan Hu[3], Yiming Zhang[3], Meng Wang[3], Haofen Wang[3]*

[1]Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University
[2]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[3]College of Design and Innovation, Tongji University
{gaoyunfan1602, carter.whfcarter}@gmail.com

Abstract

The rapid development of Large Language Models has highlighted the urgent need for large-scale, high-quality, and diverse data. We have launched an LLM data co-creation platform aimed at bringing together a wide range of participants to contribute data. Within six months, the platform has attracted over 10,000 participants who contributed more than 150,000 data entries across more than 200 tasks. An observable user cohort was constructed around the question, "Who is the best data contributor?" along with sub-questions concerning user preferences, task competence, and more. Through a detailed analysis of data contributors, this paper reveals several data collection patterns related to human factors. It reveals that contributors who provide high-quality data often do not meet initial expectations, as their behavior exhibits typical characteristics of the Dunning-Kruger effect. This paper examined the cognitive bias between users' self-assessment and actual abilities, where individuals tend to overestimate their capabilities in certain tasks, leading to a decreased willingness to continue contributing and a consequent waste of human resources. To address this issue, we propose a task reassignment method based on multi-task fine-tuning of small language models (SLMs) to better align user groups with appropriate task types. After the reallocation, we observed a significant increase in user engagement and platform benefits, along with improved overall platform efficiency. The versatility of this method makes it applicable to broader data collection scenarios.

Introduction

In the evolution of Large Language Models (LLMs), data plays an indispensable role. Following Scaling Law (Kaplan et al. 2020), the capabilities of LLMs have been significantly enhanced, where high-quality data has a decisive impact (Zhou et al. 2024). In the data-centric era of artificial intelligence, the collection of data has always been a critical issue. Typically, high-quality question-and-answer datasets constructed from existing materials (Huang et al. 2024; Gunasekar et al. 2023; Longpre et al. 2024) are often constrained by the singularity of task types, which makes it difficult to meet the diverse application needs. Moreover, the collection process of historical materials also reveals significant differences from piratical application scenarios. Another approach is manual construction, such as establishing a crowdsourcing platform to collect manually annotated data, e.g., Scale.AI and Dynaboard (Ma et al. 2021). Most existing crowdsourcing platforms mainly relying on participants' responses to questions or annotations of given texts. However, they fail to fully utilize the experience accumulated by ordinary users in daily use of LLMs and their ability to identify questions that LLMs find difficult to answer.

To fully leverage this collective intelligence, we initiated a data co-creation platform[1] with the goal of "collecting high-quality LLM evaluation data". The data collection process is shown in Figure 1. The platform posts relevant data tasks to attract users to pose questions around specific topics and to evaluate the performance of different LLMs, selecting the best model to provide feedback or manually fill in corresponding answers. The submitted data is reviewed by experts to ensure its quality and practicality. Focusing on typical LLM tasks such as logical reasoning, text generation and personal assistant , over 200 specific scenario tasks were refined, attracting more than 10,000 participants who contributed a total of over 150,000 deduplicated data entries.

The extensive accumulation of data has provided us with rich material to explore "Who are the best data contributors?". However, we need to further analyze the following two research questions (RQ) first:

RQ1: What user characteristics significantly contribute to the quality of data?

RQ2: What kind of task is suitable for what kind of users?

To answer these questions, we start from the perspective of user behavior, constructing an observable user cohort, and conducting a comprehensive analysis of user behavior preferences and task completion rates. We found that users' cognitive levels and professional backgrounds, including age, education level, occupation, and professional skills, affect their understanding in different task types, leading to significant differences in ability when handling specific tasks. Nevertheless, some interesting findings that contradict initial expectations has also been revealed. In particular, users tend to overestimate their abilities or preferences in certain tasks, which do not match their actual areas of expertise.

*Corresponding Author

[1]https://ai-ceping.com

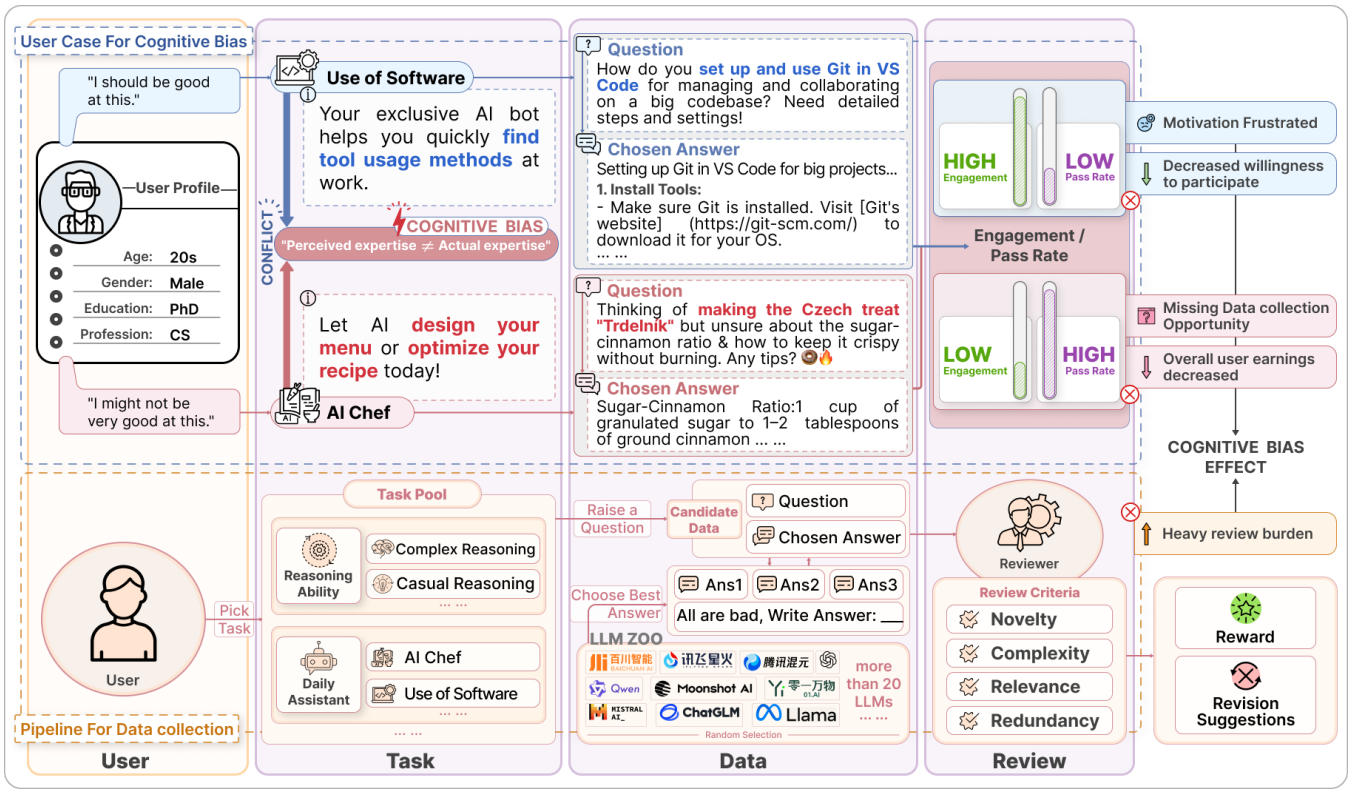

Figure 1: Data collection and user submission process. (Lower part) Users begin by selecting a task from a task pool and posing a question to a LLM. The system then randomly selects three models from an LLM zoo to provide answers. Users evaluate these responses and choose the best one. If none are satisfactory, users need supply their own answer. The user's question and selected preferences become part of a candidate data set submitted for reviewer evaluation. Successful submissions earn the user a reward. (Upper part) During data collection, we noticed a cognitive bias among users. For example, doctoral candidates showed high engagement in "software use" tasks but had a quite low success rate. Interestingly, tasks with less participation may had higher success rates, highlighting a disconnect between skills and self-perception.

This phenomenon is consistent with the Dunning-Kruger effect in cognitive psychology (Dunning 2011). It usually lead to users failing in tasks they think they are good at, which not only diminishes their enthusiasm for continued contribution but also impeded the operation of the platform.

To address this challenge, we propose a **G**enerative **D**e-**B**iased **M**atching method (GDBM) aimed at refining the alignment between users and tasks. Specifically, by establishing instruction dataset for three scenarios—"user-task matching", "user-to-task" (assigning tasks to a given user), and "task-to-user" (matching users to a given task)—we fine-tune a general small language model to alleviate biases that various demographic groups may exhibit when handling specific tasks. Experiments have shown that through task reassignment, the user's overall pass rate has increased by 35.78%, and their earnings have also seen a corresponding increase of 20.68%, ensuring a mutually beneficial outcome for both users and platform's sustainable growth. The main contributions are summarized as follows:

- This paper explores the relationship between user profiles and LLM data collection tasks, along with the cognitive biases users exhibit during the process. It specifically highlights the human factors, the mismatch between user competitiveness and task requirements, and, drawing on the Dunning-Kruger effect, constructs an observable cohort for social research in the era of LLMs.
- In addressing cognitive biases, we introduced three specific user-task matching tasks during data collection and developed a generative de-biasing and matching method. By multi-task fine-tuning on small language models, it effectively reassigned users to tasks, enhancing the data collection process.
- The proposed method's effectiveness was validated using real-world data. The experiments showed dual benefits post-reallocation: an enhancement in both individual gains and platform operational efficiency, achieving a Pareto improvement.

Related Work

LLM data collection. In response to the current challenges of data quality and diversity in LLM, researchers have embarked on multifaceted explorations. On one hand, many efforts are focused on data synthesis (Wang et al. 2022), yet concerns about the potential negative impact of

synthetic data on model performance have garnered attention (Shumailov et al. 2024). On the other hand, establishing crowdsourcing platforms has proven to be an effective solution (Wang et al. 2024; Shmueli et al. 2021). However, in the actual data collection process, efforts are often data-centric, rarely considering user dimensions such as background, skills, and task alignment. During the data collection process, bias directly affects the efficiency and ethics of the model, leading to discriminatory or unfair behavior in LLMs when dealing with certain tasks (Gallegos et al. 2024; Shmueli et al. 2021). Particularly, biases in data collection may lead to LLMs generating discriminatory content, potentially exacerbating social inequalities and propagating stereotypes and prejudices(Wan et al. 2023).

Data collection de-biasing. Researchers employ various strategies to ensure diversity in data collection (Zhuang et al. 2015). Traditional bias mitigation methods rely on statistical means, such as setting fairness objectives through techniques like threshold balancing and data augmentation, which can effectively address the under-representation of specific groups (Chen et al. 2023). However, these methods are applied after data collection, which has limitations and may lead to the wastage of resources during the process. In the early stages of data collection, especially for instruction data, the current common practice is to strive for balance among contributors of different races (Ouyang et al. 2022; Touvron et al. 2023). However, this approach fails to thoroughly consider the backgrounds, skills, and task performance of data contributors. It relies on preset rules and lacks the ability to flexibly adjust collection strategies based on users' dynamic performance.

LLMs have demonstrated potential in matching individuals with specific tasks. Zheng et al., (Zheng et al. 2023) explored using LLMs for aligning resumes with job descriptions. However, their research did not fully address the inherent biases that LLMs might possess regarding user roles. Moreover, research has found that larger parameter LLMs are more likely to adhere to their inherent biases compared to SLMs (Hsia et al. 2024), while SLMs can more easily mitigate this phenomenon through fine-tuning (Hu et al. 2024). Therefore, to fully leverage the potential of LLMs and improve efficiency and fairness during the data collection process, a mechanism is needed that can both alleviate the inherent biases of LLMs and dynamically adjust task matching based on user performance.

Cognitive Bias in Data Collection

This section will first provide an overview and empirical analysis of the data collected. Then, through data analysis, we reveal several key associations between user characteristics and data quality. Finally, we delve into a discussion on cognitive biases of users during the data collection process.

Data Collection and Basic Information

To enhance the diversity of data contribution forms and the variety of data types, the platform offers a range of tasks, such as single data entry contribution, dataset submission, multi-modal tasks, and adversarial attacks against LLMs. This study utilizes data collected from single data submission tasks during the platform's operation from March 2024 to April 2024. In data cleaning, we filtered out data with missing user information, data pending review, data from the novice guidance, and users with a total submission of less than 30 entries. The total amount of valid data entries is 46,224, with 371 valid users. For each user $u \in \mathcal{U}$, where $\mathcal{U}$ represents the set of all users, the platform's statistical information includes the following attributes:

- Gender: $G \in \{\text{Male}, \text{Female}, \text{Not Disclosure}\}$
- Age: $A \in \{\text{Under 18}, \text{18-29}, \text{30-39}, \text{40-50}, \text{50+}\}$
- Education Level (E): Including $E \in \{\text{High School}, \text{Undergraduate}, \text{Postgraduate}, \text{Ph.D.}\}$
- Major (M): Clustered into eight major categories represented by initials, specifically:Sports and Health (S&H), Medicine (Med), Philosophy and Social Sciences (P&S), Engineering and Technology (E&T), Literature (Liter), Economics and Management (E&M), Natural Sciences (NS), Arts and Design (A&D).

For the task pool T, this paper involves a total of 43 types. Each task t in the pool is characterized by a set of attributes $T(t) = \{\text{Task Name}, \text{Task Description}, \text{Examples}\}$. An example is :

Task: Emotional Response

Description: Your personal AI assistant is exclusively designed to cater to your needs.

- The AI assistant does not possess emotions, and it is your task to guide its development towards becoming an emotionally intelligent system.
- Please provide a specific scenario.
- The AI assistant is required to respond to this scenario, demonstrating appropriate emotional responses and vocabulary.
- Types of emotions to consider include sadness, frustration, happiness, excitement, anger, etc.

Example: Xiaoming had a dispute with his boss in the office because Xiaoming resisted working overtime and explicitly stated that he wants to avoid overtime; otherwise, he would use arbitration to protect his rights. What emotion is the boss most likely to use in response to Xiaoming?

Data Analysis

Different background users exhibit significant differences in data preferences and professional fields, which are intuitively reflected in the user participation in various tasks. Each user group has its areas of expertise, which is reflected in the differences in the approval rates, further confirming the purpose and value of leveraging collective wisdom. In response to RQ1: Which user characteristics significantly affect the quality of contributed data. We conducted data exploration and analysis on users' intrinsic characteristics

In terms of gender characteristics, there is a significant difference in the pass rate of data collection, with female participants consistently outperforming male participants. The average pass rate for females is 55%, while for males it is

45%, and the distribution of pass rates for males is also more concentrated.

In the age stratification analysis, the findings contradict our initial hypothesis. We had anticipated that younger individuals, who are generally perceived as being more adept at utilizing LLMs and more receptive to new technologies, would excel in formulating high-quality questions. However, the data indicate that individuals aged 30 to 39 performed best in passing tests or assessments, with a median pass rate approaching 75%. In contrast, the 18 to 29 age group had the lowest overall pass rate among the four age categories, with approximately 61% of individuals passing the tests or assessments. These results suggest that groups with more work experience may be more suitable subjects for LLM crowdsourced data collection.

In the realm of educational backgrounds, an unexpected revelation is that users with a high school education often excel at formulating questions of superior quality, outperforming those with advanced degrees. The median pass rate for high school diploma holders is approximately 70%, with a generally higher distribution. In contrast, the median pass rate for bachelor's degree holders is around 45%, with a lower overall distribution. This phenomenon suggests that an intuitive understanding of task requirements may be a key factor in their success. It is important to note that this finding indicates that contributing to LLM data is not solely a skill reserved for individuals with high levels of education. Users who may not have a profound understanding of LLM principles can still provide valuable high-quality data contributions. Of course, this conclusion is based on overall performance. In specific professional fields, recruiting professionals from those fields remains the preferred option. Furthermore, this insight provides a valuable perspective on cost management in LLM data collection, as hiring individuals with higher levels of education typically requires greater financial investment.

In terms of professional background, there is a significant variation in the ability of different user groups to develop high-quality LLM examination questions. Specifically, users in the fields of Social Sciences and Humanities (S&H) and Natural Sciences (NS) achieve the highest pass rates, with medians approaching 0.7 and 0.65, respectively. This indicates that users in these domains not only perform well but also exhibit a relatively consistent level of proficiency in the task. In contrast, users in the fields of Medicine (Med) and Arts and Design (A&D) have the lowest pass rates, with medians of approximately 0.3. This suggests that these user groups face greater challenges in developing high-quality questions and show a wider range of performance variability. These findings highlight the crucial role of users' professional expertise in the quality of the data they contribute. It implies that more targeted approaches are needed when recruiting users for LLM data collection. Otherwise, there is a risk of significant waste of human resources during the data collection phase, given the substantial differences in the ability of users from different professional backgrounds to contribute high-quality data. This discovery is of great significance for optimizing the LLM data collection process and improving data quality. It also suggests that the characteristics and needs of users from different professional backgrounds should be considered when designing data collection tasks.

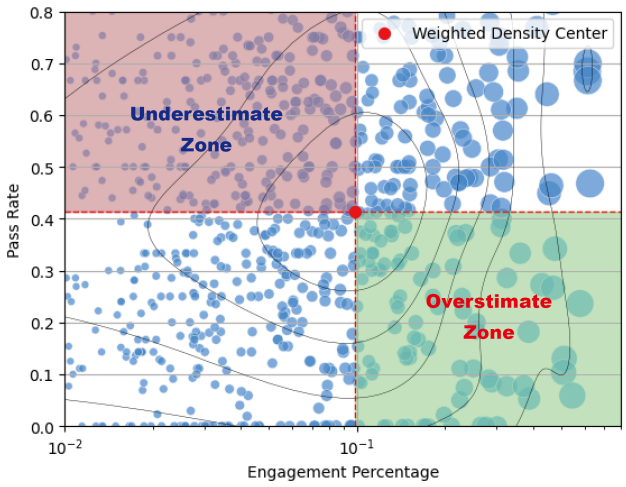

Figure 2: Cognitive bias in data collection. The engagement and actual performance across different tasks are represented by individual points. In the underestimate zone (upper left), there is a dense scattering of small points, indicating that although users have a high pass rate in these tasks, their level of participation is quite low. Conversely, in the overestimate zone (lower right), there are a greater number of larger points, suggesting that despite not excelling in these tasks, users exhibit an extremely high level of engagement and an overabundance of confidence.

Cognitive Bias

Above discoveries have piqued our curiosity to delve deeper into the alignment between user profiles, their competencies, and the data collection tasks they are suited for. In response to RQ2, which inquires about the optimal data tasks for different user types, we initiated by clustering the user base according to their profiles. Subsequently, we conducted an analysis of their engagement levels and actual performance across various tasks. A notable observation was the mismatch between the self-assessed capabilities and the actual proficiency of numerous user groups. For instance, in the task of "Weak Correlation Reasoning," individuals with doctoral degrees were disproportionately represented, constituting 81.75% of the participants, significantly exceeding their average task engagement rate of 14.4%. However, their pass rate for this task stood at a mere 42.92%, which is notably lower than their overall average pass rate of 50.77%.

This discrepancy underscores a stark contrast between the self-perceived competence of data contributors and their empirical performance, suggesting that tasks users are confident in are not necessarily their areas of genuine expertise. This observation aligns with the Dunning-Kruger effect, a cognitive bias where individuals with limited skills often overestimate their abilities, while those with greater skills may undervalue their competencies. The effect highlights

that individuals deficient in certain skills are frequently oblivious to their shortcomings due to the absence of the cognitive apparatus required for self-assessment.

Guided by this theory, we further explored the quadrant chart composed of pairs $(\mathbf{u}, t)$ of user group $\mathbf{u}$ and task t. This chart comprehensively measures users' engagement and actual performance across different tasks within various user groups, where they often engage in tasks they are not good at and perform less in tasks they excel in. As shown in Figure 2, the x-axis represents the participation ratio of users, and the y-axis represents the task pass rate. The central point is the KDE mean centroid weighted by quantity, along with the corresponding contour lines, where the size of the points represents the number of specific user task pair.

The most ideal state is the area traversed by the right diagonal line, that is, the upper right and lower left quadrants, where there is a simultaneous increase/decrease in output in tasks with higher/lower pass rates. After several attempts and receiving feedback of rejection, one should reduce attempts in tasks they are not proficient in. However, in reality, a large number of small points are scattered in the upper left area, which we call the "Underestimate zone". Despite being proficient, users show low enthusiasm for participation, missing many opportunities. On the other hand, larger points are clearly scattered in the lower right area, which we call the "Overestimate zone". This indicates that although users are not proficient in these tasks, they have a high level of participation enthusiasm and excessive confidence, and the negative review results of the data have not affected them. In the absence of guidance, underestimation will lead to insufficient data collection for some related tasks, while overestimation will lead to an excess of low-quality data, increasing the burden of audit and blocking the smooth operation of the platform. Therefore, a dynamic matching mechanism is needed to improve this mismatch between cognition and skills, guiding users to fill in what they are good at and avoid mismatched tasks.

Method

To address the mismatch between user skills and task requirements, this paper introduces a generative de-biased matching method called GDBM. The overall framework is delineated in Figure 3. Specifically, the process commences with clustering based on group characteristics, to extract a set of representative user groups $\mathbf{u}$. Subsequently, we establish pairings between these groups and various tasks t, while meticulously filtering out any group with a submission count less than 30 to uphold the integrity of the dataset.

Building upon the foundation of group-task alignment, we have adeptly crafted three specialized matching tasks to address the pivotal question: "From which individuals and what types of data should we collect?"

- $M_{\text{match}}(\mathbf{u}, t)$ User-Task Matching Task. Determining whether a particular set of users is well-matched with a given task.
- $M_{\text{assign}}(t, \mathbf{u})$ Task-User Assignment Task. For a given task, assign the most suitable user group to address the task based on its requirements.

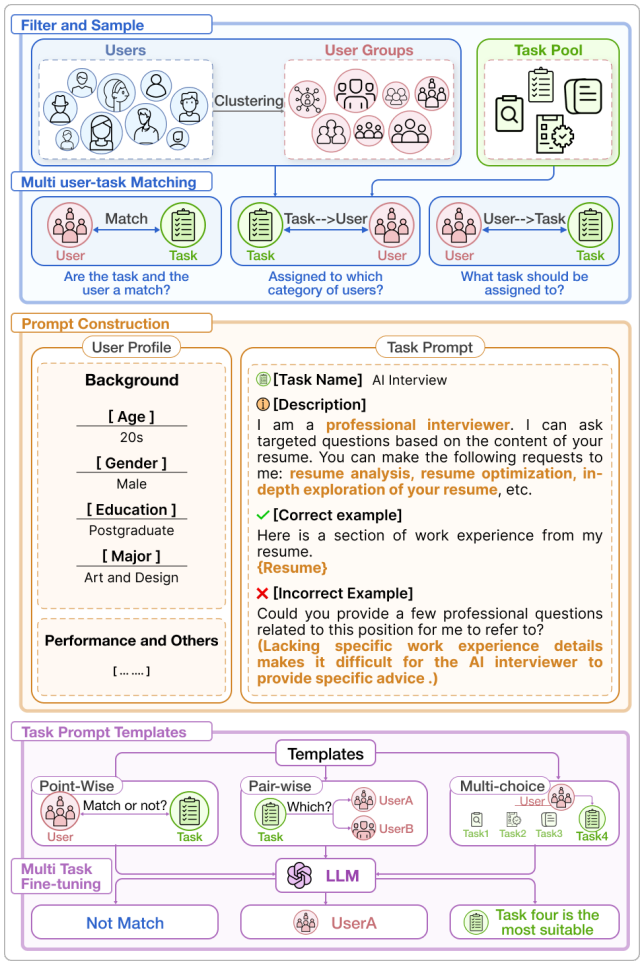

Figure 3: The framework of generative de-biased matching method.

- $M_{\text{assign}}(\mathbf{u}, t)$ User-Task Assignment Task. For a user group, identifying task that are aligned with their competencies.

Building upon the work (Wu et al. 2024; Gao et al. 2023), we have further developed three refined instruction templates tailored to these tasks as follows. It should be noted that the original data collection and instruction prompts are in Chinese. For ease of understanding, they are described here in English.

Point-Wise Matching: As an LLM data collection expert, you are recruiting users to complete high-quality data collection tasks. Based on the current {user profile} and {task requirements} please assess whether there is a match between the two. Please respond with "Yes" or "No."

Pair-wise Assignment: As an LLM data collection expert, you are now recruiting users for the {task}. For {candidate A} and {candidate B} , decide who is more suitable based on their background information and the task's demands. Please respond with "A" or "B."

Multi-choice Assignment: As an LLM data collection expert, given the {user profile} and {tasks A, B, C, etc.},

please determine the most suitable task for the current user. Please respond with the name of the task.

After creating the instruction dataset, we train the small language model in the supervised fine-tuning way, enabling it to learn the complex matching relationships between actual user characteristics and task skills. By leveraging continuously updated data from the platform's operations, it can be dynamically and periodically adjusted to fit user preferences. The selection of SLM is advantageous for several reasons. On one hand, relevant research (Zheng et al. 2023) has demonstrated that LLMs tend to retain their inherent knowledge and biases more stubbornly compared to SLMs. Conversely, SLMs are more amenable to adjustments through fine-tuning, allowing for greater flexibility and adaptability in response to new data and tasks. Specially, for generator $\mathcal{G}$, given the user group $\mathbf{u}$, the task description t, and the prompt template $\mathcal{T}$, objective is to optimize the negative log-likelihood to generate correct responses:

$$\begin{aligned}\mathcal{L}_{\text{total}} &= \mathcal{L}_{M_{\text{match}}(\mathbf{u},t)} + \mathcal{L}_{M_{\text{assign}}(t,\mathbf{u})} + \mathcal{L}_{M_{\text{assign}}(\mathbf{u},t)} \\ &= -(logPr(\mathbf{u},t|\mathcal{T},\mathcal{G}) + logPr(t|\mathbf{u},\mathcal{T},\mathcal{G}) \\ &\quad + logPr(\mathbf{u}|t,\mathcal{T},\mathcal{G}))\end{aligned} \tag{1}$$

where $logPr(t|\mathbf{u},\mathcal{T},\mathcal{G})$ denotes the generation probability for $\mathbf{u}$ of the generator model $\mathcal{G}$ given the task t and the prompt template $\mathcal{T}$.

Experiment

Dataset and setting. We conducted random sampling from a pool of authentic users. When constructing the candidate tasks and users, we first determined quantiles based on the actual pass rate and then performed random sampling N times within each segment to ensure the distinguishability of each candidate option. For task $M_{\text{match}}(\mathbf{u},t)$, which is constructed under Point-wise template, there are a total of 715 data entries, denoted as $\mathcal{D}_{match}$. For the task $M_{\text{assign}}(t,\mathbf{u})$, we used a pair-wise prompt template to construct a total of 925 data entries, which are labeled as $\mathcal{D}_{t-pair}$. For the task $M_{\text{assign}}(\mathbf{u},t)$, we used pair-wise and multi-choice templates to construct 1,580 and 495 data entries, respectively, labeled as $\mathcal{D}_{u-pair}$ and $\mathcal{D}_{u-multi}$. We split all the datasets into a training set and a validation set in an 8:2 ratio, with a consistent random seed of 42 for reproducibility. For each fine-tuning task, we employ the Parameter-Efficient Fine-Tuning (PEFT) method, specifically, LORA (Hu et al. 2021). Each task is trained for 3 epochs. The experimental setup utilizes 4 Nvidia GeForce RTX 3090 GPUs.

Evaluation Metrics. We initially establish the ground truth by identifying the tasks with the highest pass rates among various groups and the groups with the highest pass rates for specific tasks. Accuracy (Acc.) is utilized to evaluate the model's predicted values against the actual label. During the data collection process, the pass rate is not the sole metric; different levels of difficulty, innovation, and other dimensions are considered to further categorize the data submitted, with varying cash rewards and platform points awarded accordingly—points that can be redeemed for virtual or real-world benefits. To further quantify the alignment between user skills and task assignments, and to reflect the benefits to users post-task reallocation from their perspective, we have developed the Average Reward per Task (ART) ratio:

$$\text{ART}_i = C_i + P_i \times \alpha \tag{2}$$

and the Expected Average Reward per Task (EART) ratio:

$$\begin{aligned}\text{EART}_i &= \text{ART}_i \times Pr_i \\ &= (C_i + P_i \times \alpha) \times Pr_i\end{aligned} \tag{3}$$

Where C_i represent the direct cash bonus for the i-th task, P_i the points reward, α is the hyperparameter for converting points to cash (e.g., 0.1) , and Pr_i is the pass rate of the i-th task. For instance, when EART equals 1, it indicates that the expected earning for the user of per submission in this task is one.

Baseline. To evaluate the matching effectiveness between user groups and specific tasks, we selected several baseline methods to compare with our method, as follow:

- **Embedding-based Method**. Leveraging dense encoding models to transform user characteristics and task requirements from prompts into vectors, subsequently matching the best results through similarity calculations. We select BERT and GTE model (Li et al. 2023).
- **Closed-source LLMs**. Representing the most powerful models currently available, we have utilized GPT-4-turbo and GPT-4-mini for our comparisons.
- **Open-source SLMs**: Smaller and open-source language models that have not undergone SFT for multiple tasks. We have selected mainstream models such as Llama 3.1-8B-Instruct and Qwen2-7B-Instruct (Yang et al. 2024),.

Results and Analysis

The experimental results, as depicted in Table 1, demonstrate that Generative De-Biased Matching (GDBM) method effectively mitigates cognitive biases among users and facilitates a reallocation between user groups and data collection tasks. GDBM fine-tuned on Llama3.1-8B shows a 40.14% improvement in overall accuracy compared to the baseline, while the version fine-tuned on the Qwen2-7B achieved a 35.78% enhancement. Overall, the GDBM (Qwen) outperforms, which may be attributed to its larger proportion of Chinese-language corpora and, consequently, its stronger comprehension capabilities in Chinese. The improvements observed on both base models also demonstrate the generality and generalizability of our approach.

Furthermore, when it comes to matching and reallocating user skills to tasks, powerful closed-source LLMs, such as GPT-4-Turbo, do not exhibit superior performance. It has been observed that large models often have distinct stylistic preferences. For instance, in the point-wise matching task, GPT-4-Turbo demonstrates a high degree of permissiveness, evaluating users as competent in 84.62% of cases. On the other hand, while embedding-based models perform well in simpler matching tasks, they tend to underperform in more complex tasks. In contrast, SLMs can achieve results close to those of closed-source LLMs without requiring fine-tuning.

	Model	Overall		$\mathcal{D}_{match}$		$\mathcal{D}_{t-pair}$			$\mathcal{D}_{u-pair}$			$\mathcal{D}_{u-multi}$	
		Acc.	EART	Acc.	Acc.	ART	EART	Acc.	ART	EART	Acc.	ART	EART
Embedding Model	BERT-base	0.475	1.323	0.490	0.476	2.290	1.466	0.502	2.123	1.245	0.364	2.205	1.303
	GTE-base	0.432	1.219	0.392	0.460	2.300	1.417	0.470	2.067	1.165	0.313	1.795	1.022
Close-source LLM	GPT-4o-mini	0.455	1.239	0.441	0.470	2.203	1.339	0.520	2.011	1.214	0.242	1.901	1.132
	GPT-4-turbo	0.434	1.229	0.413	0.460	2.262	1.363	0.478	1.989	1.149	0.273	2.152	1.235
Open-Source SLM	Llama3.1-8B	0.416	1.186	0.406	0.454	2.235	1.357	0.472	1.989	1.133	0.182	1.860	1.033
	GDBM (Llama)	<u>0.583</u>	1.383	0.472	<u>0.524</u>	2.254	1.418	**0.693**	2.073	1.364	**0.505**	2.084	1.376
	Impr.	(0.167)	(0.197)	(0.067)	(0.070)	(0.019)	(0.062)	(0.222)	(0.083)	(0.231)	(0.323)	(0.224)	(0.343)
	Qwen2-7B	0.450	1.218	0.483	0.476	2.251	1.376	0.494	2.007	1.158	0.212	1.889	1.116
	GDBM (Qwen)	**0.611**	1.470	**0.569**	**0.660**	2.477	1.674	<u>0.661</u>	2.117	1.342	<u>0.424</u>	2.387	1.499
	Impr.	(0.162)	(0.252)	(0.087)	(0.184)	(0.226)	(0.298)	(0.168)	(0.110)	(0.183)	(0.212)	(0.498)	(0.383)

Table 1: The performance of our Generative De-Biased Matching (GDBM) method across three matching tasks. Two variants of our method were fine-tuned using the Llama3.1-8B-Instruct and Qwen2-7B-Instruct models, respectively. The top two results in terms of accuracy are highlighted with bold and underlined text for emphasis.

Coupled with their lightweight nature, ease of deployment, and ease of fine-tuning, as well as their high instruction compliance and conversational capabilities, they make an ideal base model for online reallocation tasks.

Based on the inherent mechanisms of SLMs and supported by experimental results, it is evident that in the online data collection process, particularly for user group and task matching, large language models are essential. This is especially true when fine-tuning data is limited, as these models possess strong instruction-following capabilities that embedding models lack. However, there is no need for excessively large models, as they incur higher inference costs and possess inherent inductive biases that are challenging to correct and adapt quickly to specific domain scenarios.

Post-reallocation, not only is there an enhancement in the overall pass rate of user groups, but the efficiency of data collection is also significantly improved, which greatly benefits the users involved. The EART metric indicates that, following reallocation, the GDBM (Llama/Qwen) have seen overall increases of 16.62% / 20.68%, respectively. This also exhibits that the user's own earning will also be greatly benefit from the reassignment. This aligns with the Pareto improvement theory (Pang, Deng, and Chiu 2015) in economics, which emphasizes maximizing the utilization of resources through their reallocation, ensuring a non-zero-sum outcome where the interests of the participants are not compromised. Such a win-win situation contributes to the sustainable and healthy operation of data collection activities.

Ablation Study. To verify the effectiveness of multi-task fine-tuning for GDBM, we conducted ablation experiments based on a version of GDBM (Qwen) and measured the outcomes using accuracy (Acc). The results, as shown in Table 2, validate the reasonableness of our experimental setup. Although fine-tuning for specific tasks can potentially enhance the performance of a particular task, it can also degrade overall performance. This is especially true for the most challenging multi-choice tasks, where performance significantly decreases.

	Overall	$\mathcal{D}_{match}$	$\mathcal{D}_{t-pair}$	$\mathcal{D}_{u-pair}$	$\mathcal{D}_{u-multi}$
w/o $\mathcal{D}_{match}, \mathcal{D}_{t-pair}$	0.422	0.264	0.129	0.671	0.404
w/o $\mathcal{D}_{u-pair}$	0.381	0.528	0.177	0.383	0.546
w/o $\mathcal{D}_{t-pair}$	0.503	0.542	0.157	0.684	0.515
w/o $\mathcal{D}_{match}$	0.553	0.208	0.643	0.674	0.495
w/o $\mathcal{D}_{u-multi}$	0.583	0.653	0.616	0.658	0.182
All	**0.611**	0.569	0.660	0.661	0.424

Table 2: The ablation study of GDBM (Qwen) on different matching tasks and datasets. The evaluation metric is Acc.

Conclusion

Focusing on the collection of high-quality LLM data, this paper initiated a collaborative data collection platform that engages a broad community of LLM users. It analyzes the relationship between user characteristics and the completion of data collection tasks, uncovering cognitive biases within data contributors during the data collection process. Utilizing a multi-task fine-tuning approach based on small language models, tasks were successfully reassigned to more suitable user groups, significantly mitigating these cognitive biases. This paper provides a solution to identifying high-quality contributors for LLMs and determining what types of data to collect from whom.

The collaboratively LLM data collection platform offers a rich playground for engaging with a broad community of users. While crowdsourcing remains the most important source of high-quality data in the short term, effectively leveraging it requires a human-centric approach, focusing not only on the data but also on understanding the contributors. For future research, we plan to expand our data collection efforts to include a wider array of tasks and participants, enhancing both diversity and representativeness. Additionally, we aim to rigorously test and refine our task reallocation algorithms in a live environment to validate their effectiveness under dynamic conditions and gather valuable feedback for future iterations.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (U23B2057, 62176185, 62276063), the National Key Research and Development Program of China (2022YFF0712400), and the Natural Science Foundation of Jiangsu Province (BK20221457).

References

Chen, J.; Dong, H.; Wang, X.; Feng, F.; Wang, M.; and He, X. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3): 1–39.

Dunning, D. 2011. The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*, volume 44, 247–296. Elsevier.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Gao, Y.; Sheng, T.; Xiang, Y.; Xiong, Y.; Wang, H.; and Zhang, J. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*.

Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C. C. T.; Del Giorno, A.; Gopi, S.; Javaheripi, M.; Kauffmann, P.; de Rosa, G.; Saarikivi, O.; et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Hsia, J.; Shaikh, A.; Wang, Z.; and Neubig, G. 2024. RAGGED: Towards Informed Design of Retrieval Augmented Generation Systems. *arXiv preprint arXiv:2403.09040*.

Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and Qi, P. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22105–22113.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Fu, Y.; et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; and Zhang, M. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Longpre, S.; Mahari, R.; Lee, A.; Lund, C.; Oderinwale, H.; Brannon, W.; Saxena, N.; Obeng-Marnu, N.; South, T.; Hunter, C.; et al. 2024. Consent in Crisis: The Rapid Decline of the AI Data Commons. *arXiv preprint arXiv:2407.14933*.

Ma, Z.; Ethayarajh, K.; Thrush, T.; Jain, S.; Wu, L.; Jia, R.; Potts, C.; Williams, A.; and Kiela, D. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *Advances in Neural Information Processing Systems*, 34: 10351–10367.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pang, R.-z.; Deng, Z.-q.; and Chiu, Y.-h. 2015. Pareto improvement through a reallocation of carbon emission quotas. *Renewable and Sustainable Energy Reviews*, 50: 419–430.

Shmueli, B.; Fell, J.; Ray, S.; and Ku, L.-W. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. *arXiv preprint arXiv:2104.10097*.

Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; and Gal, Y. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wan, Y.; Pu, G.; Sun, J.; Garimella, A.; Chang, K.-W.; and Peng, N. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Wang, X.; Kim, H.; Rahman, S.; Mitra, K.; and Miao, Z. 2024. Human-LLM collaborative annotation through effective verification of LLM labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–21.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Wu, L.; Qiu, Z.; Zheng, Z.; Zhu, H.; and Chen, E. 2024. Exploring large language model for graph data understanding in online job recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 9178–9186.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zheng, Z.; Qiu, Z.; Hu, X.; Wu, L.; Zhu, H.; and Xiong, H. 2023. Generative job recommendations with large language model. *arXiv preprint arXiv:2307.02157*.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Zhuang, H.; Parameswaran, A.; Roth, D.; and Han, J. 2015. Debiasing crowdsourced batches. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1593–1602.