Development on an Information Eliciting Model

Shuixiunan Zhang
IIIS, Yao Class 14
Tsinghua University
2021011832
zsxn21@mails.tsinghua.edu.cn

Weiliang Wang
IIIS, Yao Class 14
Tsinghua University
2021012928
wangweil21@mails.tsinghua.edu.cn

Abstract

The truth does not always align with the majority. When we understand the diverse perspectives people hold on a certain issue, how can we uncover the true answer to the question from their response? Previous researchers have proposed the concept of thinking hierarchy to address this issue. It suggests that people's perspectives are not on the same plane, and some individuals' answers occupy higher positions. Different answers exist at different levels within this hierarchy. By relying solely on people's responses without any prior knowledge, we can discover the thinking hierarchy within a population and effectively find the truth. This paper extends this viewpoint, develops a thinking hierarchy model, and presents a series of methods to measure the hierarchy among different answers. From this standpoint, we further propose specific metrics for measuring hierarchy, namely the hierarchical score and hierarchical rate.

1 Introduction

1.1 Background

Information eliciting refers to the process of gathering information from individuals in order to make informed decisions or to learn more about a particular phenomenon. It is an important concept in fields such as market research, public opinion polling, and data science.

The significance of information eliciting lies in the fact that accurate and reliable information is often crucial for making good decisions. By eliciting information from individuals, we can gain insights into their beliefs, preferences, behaviors, and attitudes, and use this information to guide our actions.

However, the process of information eliciting can be complex and fraught with challenges. For example, individuals may be reluctant to share certain types of information, or they may provide inaccurate or biased responses. Thus, it is important to carefully design and implement methods on an information eliciting model in order to maximize the quality and usefulness of the information obtained.

One issue in information elicitation is how to uncover the truth from the perspectives of a population. Specifically, when collecting questionnaires from individuals on certain issues, how can we find the correct answers directly from their responses without relying on prior knowledge? It is tempting to rely on the choices of the majority, but as stated, "The truth does not always align with the majority."

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

This phenomenon, known as thinking hierarchy[2], occurs when someone believes that A is correct, while the majority of people believe that B is correct. An investigation into thinking hierarchy could provide us with powerful tools for eliciting truthful responses from questionnaires. With this in mind, our main objective is to survey and explore potential mechanisms for measuring and utilizing hierarchical information. By integrating thinking hierarchy with truth telling, we can extract more trustworthy information from collected data

1.2 Related work

1.2.1 Crowd wisdom problem

Prelec et al.[4] explores the limitations of democratic voting methods in extracting accurate information from a crowd and proposes an alternative approach. By selecting the answer that is more popular than people predict, the authors demonstrate that this principle outperforms traditional methods based on popularity or confidence. The significance of this research lies in providing a more reliable and effective way to tap into the collective wisdom of a crowd, enabling informed decision-making across various domains, including scientific, artistic, legal, and historical contexts. This approach expands the potential application domain beyond machine learning and psychometric methods, which require data across multiple questions.

1.2.2 Thinking hierarchy

Kong et al.[2] proposes a new model and approach to address the issue of ranking answers when the majority may make systematic mistakes. They argue that ranking answers solely based on popularity can be dangerous, and propose a thinking hierarchy model that considers the sophistication level of the respondents. The paper presents two algorithms for learning the thinking hierarchy and a crowd-sourcing approach based on the proposed framework.

The results show that their approach outperforms plurality voting in terms of accuracy of top-ranking answers, especially in questions where the correct answer is not supported by the majority. However, there is still a series of deficiencies in this model.

The formula used in the model is too rough, and it fails to make use of all the information available, leading to some information being overlooked. In particular, the correctness of prediction of a participant is not considered. This can lead to inaccurate predictions and can affect the overall efficiency of the model.

In their research, the questions are considered independently, without taking into account the hierarchy level of the participants. In reality, the hierarchy level of a participant is fixed, and their intelligence appears similar on different questions. This means that some participants may be more skilled than others and can make more accurate predictions. However, the current model does not account for this.

1.2.3 Peer prediction to enable truth-telling

In situations where the ratings cannot be independently verified and no observable objective truth exists, a major problem arises: non-truthful equilibrium may have higher payoff for the respondents than truth-telling.

In order to address this issue, Kong et al. [1] introduces peer prediction, in part by specially selecting the proper scoring rule it is based on, to make the truth-telling equilibrium focal (i.e. truth-telling has a higher expected payoff than any other equilibrium).

The main technical tool used to achieve this is a best response plot, which is a graphical representation of the expected payoffs of different equilibria given different levels of signals from other respondents. By analyzing the best response plot, the most favorable strategy for each respondent can be determined, and they ensured that the equilibrium in which all respondents truthfully report their signals is the most attractive option.

The proposed solution of tweaking peer prediction and selecting the proper scoring rule to make the truth-telling equilibrium focal is promising. The use of a best response plot as a technical tool to achieve this is also a valid approach, as it provides a graphical representation of the expected payoffs of different equilibria given different levels of signals from other respondents.

However, the research could be further strengthened by providing more detailed information on how the best response plot is constructed and analyzed. Additionally, the effectiveness of the proposed solution could be tested in real-world scenarios to assess its practicality and applicability.

1.2.4 Bayesian truth serum

Prelec [3] proposed a Bayesian-based incentive method to elicit truthful answers in subjective questions. Specifically, he introduced an information score to quantify the degree of "surprisingly common" and "surprisingly uncommon" answers. This score has a heuristic interpretation suggesting that individuals tend to perceive common answers as distinct from their own preferences. By using the information score as a utility in the game, he demonstrates that truth-telling constitutes a Nash equilibrium based on this measurement. This method is potential to effectively elicit truthful responses, and providing a methodological reassurance for social science. The idea of score quantization may be extended on objective problems by combining Bayesian statistic with hierarchy information.

2 Main

2.1 Basic concepts in thinking hierarchy

Respondents' types Respondents are individuals who provide answers and predictions in response to a question. p_t denotes the probabilities of respondents belonging to each thinking type t. These probabilities indicate the distribution of respondents across different thinking types and satisfy the condition of $\sum_t p_t = 1$.

Oracles of thinking types An oracle refers to a mapping that associates a question q with a random answer from the answer set A. Each $O_t(q)$ corresponding to a specific thinking type t, representing the distribution of answers produced by the oracle for a given question q. The oracle matrix W has dimensions $|T| \times |A|$, where each row t corresponds to the distribution $w_t \in \Delta A$, representing the probabilities of answers generated by the oracle $Ot(q)$ of thinking type t.

Type distribution matrix The probability $p_{t \leftarrow t'}$ denotes the probability that a respondent, being of type t, predicts another respondent to be on distribution $p_{t'}$ in a fixed question q. This probability is captured by the type distribution matrix Λ with dimensions $|T| \times |T|$, where $\Lambda_{t,t'} = p_t p_{t \leftarrow t'}$.

Answer-prediction joint distribution matrix For a fixed question q, the answer-prediction joint distribution matrix M^* is a $|A| \times |A|$ matrix where $M^*_{a,g}$ is the probability an respondent answers a and predicts g. Let the answer-prediction matrix $\bar{M}$ which is obtained from the reality be the approximation of M^*. Let the answer-prediction matrix M obtained from the observed data, serve as an approximation of the true distribution matrix M^*.

Statement By the above definition, we have $M^* = W\Lambda W^T$ for strict answer-prediction joint distribution matrix M^*.

The paper by Kong et. al.[2] gives a formal proof that when W is restricted to semi-orthogonal matrix, for $\min_{W,\Lambda} ||M - W^T \Lambda W||_F^2$ is equivalent to solving $\max_W \sum_{i \le j} (WMW^T)_{i,j}^2$, with an assumption that Λ is set to an upper-triangular matrix. Base on this, they introduced an answer-ranking algorithm $\Pi^* \leftarrow \arg\max_{\Pi \in \mathcal{P}} \sum_{i \le j} (\Pi M \Pi^T)_{i,j}^2$, where $\mathcal{P}$ is the set of all $|A| \times |A|$ permutation matrices.

In our study, we extend the scope of our analysis to encompass a broader range of scenarios, thereby relaxing the assumption of Λ being an upper-triangular matrix. Instead, we consider a type matrix Λ that adheres to certain properties, contingent upon specific assumptions. By doing so, we establish a comprehensive framework to accommodate a wider array of situations and enable a more robust investigation.

2.2 Framework of thinking hierarchy model

To formalize the process of raising the best recommended answer, we introduce a ranking rule that enables the systematic determination of the order in which the answers are ranked.

Ranking value Let M be the answer-prediction joint distribution matrix and Π be a permutation matrix. Define R' as a mapping from $\mathbf{R}^{|T|\times|T|}$ to $\mathbf{R}^{+}$, which represents the scoring or evaluation metric for the permutation of M.

Note that each permutation represents a linear ordering π, where $\Pi_{i,\pi(i)} = 1$ indicates that element i in the linear ordering corresponds to position $\pi(i)$. Thus, we express the ranking value R as a function of the answer-prediction matrix M and the permutation π: $R(M, \pi) = R'(\Pi M \Pi^T)$.

Ranking rule The framework of ranking rule essentially seeks to find the permutation π^* that yields the highest possible ranking value $R(M, \pi)$ among all possible permutations, indicating the optimal ordering of the elements in the answer-prediction matrix M.

$$\pi^* \leftarrow \arg\max_{\pi} R(M, \pi)$$

We can also extend beyond providing a single recommended answer and offer a ranked sequence of all the answers based on the ranking values.

Within our framework, the specific definition of the ranking value relies on the chosen evaluation criteria. Consequently, we have successfully transferred the ranking rule to providing the explicit formula that governs the ranking value.

In specific, the index $\pi^*(0)$ corresponds to the best recommended answer.

Now we will introduce several assumptions regarding the hierarchical structure of thinking. These assumptions will serve as the foundation for the subsequent methods presented to predict the optimal answer.

2.2.1 Majority

Assumption 0 *Majority voting provides reasonable answer.*

The majority assumption serves as the baseline method. We formalize it into our model framework, and the ranking value is defined as follows:

$$R(M, \pi) = \sum_{i,j} (n - i) M_{\pi(i),\pi(j)}$$

Here n denotes the number of answers for this question, and i, j ranges from 1 to n respectively.

The ranking value is computed as the sum of the elements for each row of M, weighted by the difference in indices. It simply ranks answers by the number of supporters, completely ignoring the information of prediction.

2.2.2 One-way oracle

Assumption 1 *People of less sophisticated level can never run the oracles of higher hierarchy.*

The deduced ranking value is as follows:

$$R(M, \pi) = \sum_{i \le j} M^2_{\pi(i),\pi(j)}$$

The formula focuses on the upper triangular portion of the matrix M, i.e., elements where $i \le j$. This emphasis reflects the assumption that individuals of lower hierarchical levels cannot access or utilize the oracles of higher hierarchical levels.

This ranking rule is equivalent to the default ranking algorithm in Kong et. al.'s paper [2]. We successfully reproduced the result in our study (see Section 3) that considering thinking hierarchy works better than the baseline (Assumption 0).

However, as critically analysed in Section 1.2.2, when taking the same weight for each entry in the upper triangular, it leads to some information in the structure overlooked. The next part optimizes this problem to some extend.

2.2.3 Cluster estimation

Assumption 2 *People are grouped based on their intelligence. They tend to believe that the majority is close to their thinking hierarchical score.*

We deduced two variations of ranking value, namely the beta-triangular method and band method:

$$R(M, \pi) = \sum_{i \le j} \beta^{j-i} M^2_{\pi(i),\pi(j)},\ \beta \in (0, 1)$$

The beta-triangular method calculates the ranking value based on M, where the square of the elements are still on the upper-triangular, but is multiplied by a weight that decreases exponentially with the distance between the diagonal, controlled by the parameter $\beta < 1$.

This method incorporates the notion that, people believe the majority is closer to their own thinking. By exponentially decaying the weight based on the distance in the hierarchy, the method captures the varying levels of influence and relevance attributed to different hierarchical positions.

$$R(M, \pi) = \sum_{0 \le j-i \le 1} M^2_{\pi(i),\pi(j)}$$

The band method calculates the ranking value on M, where only the elements within a band of width 2 near the diagonal are considered, which captures the local relationships and importance of nearby positions, and thus also aligns with the assumption.

Among the statistical results of our sample questionnaire, however, the performance of both methods is relatively poor. This implies that the assumption here does not fit the reality well. So we move to the contrary assumption next.

2.2.4 Debasement on hierarchy

Assumption 3 *People tend to underestimate the hierarchical score of others.*

We deduced the ranking value here:

$$R(M, \pi) = \sum_{i \le j} \beta^{j-i} M^2_{\pi(i),\pi(j)},\ \beta > 1$$

This method is almost the same as the beta-triangular method discussed in Assumption 2, except that β is greater than 1 to control the weights assigned to different positions in the ranking. This allows the method to better capture the perceived importance of positions and align it with individuals' underestimation tendencies.

It reaches high performance in our statistical result, addressing the underestimation bias of individuals.

2.2.5 Rational guesses

Assumption 4 *People who knows the question well tend to make right guesses.*

We deduced the ranking value here:

$$R(M, \pi) = \sum_{i,j} (n - i) \cdot \frac{M_{\pi(i),\pi(j)}}{\sum_k M_{\pi(i),k}} \cdot (\sum_k M_{\pi(j),k})^2$$

The prediction accuracy method considers the weighted position, the relative importance of options, and the amplifying effect of probabilities.

This method captures the local information at position $\pi(i)$ by examining the proportion of probability assigned to option $\pi(j)$ within that position. Simultaneously, the term $(\sum_k M_{\pi(j),k})^2$ accounts for the global information by considering the overall probability associated with option $\pi(j)$ across all positions.

It incorporates these intuitions and ideas to rank the answers accordingly, leveraging the assumption that individuals who possess more knowledge about the question are more likely to make correct guesses, and it successfully outperforms the baseline (Assumption 0) and the naive upper triangular method (Assumption 1).

2.2.6 Random guess by lower hierarchy

Assumption 5 *People of high hierarchy know majority's answer well, while lower level guesses randomly.*

We deduced the ranking value here:

$$R(M,\pi) = \sum_{i,j}(n-i)\cdot M_{\pi(i),\pi(j)}\cdot Var_k(\frac{M_{\pi(i),k}}{\sum_r M_{\pi(i),r}})$$

where $Var_k(\frac{M_{\pi(i),k}}{\sum_r M_{\pi(i),r}})$ denotes the variance:

$$Var_k(\frac{M_{\pi(i),k}}{\sum_r M_{\pi(i),r}}) = \sum_k(\frac{M_{\pi(i),k}}{\sum_r M_{\pi(i),r}} - \frac{1}{n})^2$$

This variance method aims to capture the diversity in lower-level guesses while taking into account the confidence associated with higher-level positions' understanding of the majority's answer.

This is a heuristic method, since it does not require that the different answers in a question forms a strict hierarchical structure. This suits the reality especially on non-logical questions. For instance, on the question "What country is Chopin from?", "Germany" and "UK" are both wrong answers, and it's hard to tell which is in a strictly higher hierarchical level.

Though with a more relaxed requirement, this method still reaches high accuracy, and outperforms the baseline (Assumption 0) and the naive upper triangular method (Assumption 1).

2.2.7 Rational belief on distribution

Assumption 6 *People has a rational belief on the distribution of those in lower hierarchy, i.e., their belief on lower types $M_{i,j} \sim p(j|i)$ aligns well with the true distribution $\sum_j M_{i,j} \sim q(i)$.*

This results in a low Kullback-Leibler (KL) divergence:

$$D_{KL}(p||q)_i = -\sum_j p(j|i)\log(\frac{q(j)}{p(j|i)})$$

Note that $D_{KL} \geq 0$, we deduced the ranking value here:

$$R(M,\pi) = \sum_i \frac{1}{D_{KL}(p||q)_i^{\pi} + 1}$$

By minimizing D_{KL} in the ranking rule, this Bayesian method seeks to capture the rationality of individuals in higher hierarchy, by considering the prevalence of options and the discrepancies between beliefs and reality.

Compared to Assumption 4, it reaches an even higher accuracy since it not only considers the correctness of prediction, but also takes a general consider on the distribution.

2.3 Hierarchical measures

In this subsection, we introduce two novel concepts, namely "hierarchical rate" and "hierarchical score," to measure the hierarchical level of respondents and questions in the context of our study. These concepts provide a theoretical framework for understanding and evaluating the relative importance and expertise of respondents and the complexity of questions being addressed.

In the preceding sections, all the concepts and variables were defined with respect to a particular fixed question denoted as q. To signify the variables specifically associated with question q, we employ subscripts, replacing the generic notation with the subscripted version. For instance, the variable previously denoted as M is now represented as M_q.

Hierarchical rate *For a question q on the optimizing framework $\pi_q^* \leftarrow \arg\max_{\pi_q} R(M_q, \pi_q)$, define its hierarchical rate to be:*

$$\rho_h(q) = \log \frac{R(M_q, \pi_q^*)}{\frac{1}{n_q!} \sum_{\pi_q} R(\pi_q)}$$

The hierarchical rate captures the relative improvement achieved by the optimal permutation compared to the average performance across all permutations, it is a measure that quantifies the performance of a specific question q in an optimizing framework.

The hierarchical rate provides a quantitative measure of the relative performance of the optimizing framework for a given question. It can be used to compare the performance of different questions within the model. By analyzing the hierarchical rates of different questions, one can gain insights into the effectiveness of the trends in the performance across different question types, which is discussed later in Section 3.3.

Hierarchical score *For an respondent r on the optimising framework $\pi_q^* \leftarrow \arg\max_{\pi_q} R(M_q, \pi_q)$ and a question set $\mathcal{P}$, define the agent's hierarchical score to be:*

$$S_h(r) = \frac{1}{|\mathcal{P}|} \sum_{q \in \mathcal{P}} \rho_h(q) \cdot (\sum_{k_q} X_{q,k_q}^r (1 - \frac{\pi_q^*(k_q)}{n_q})$$

where X_{q,k_q}^r is whether agent r choose answer k_q in a question q.

A naive way is to define hierarchical score as $\frac{1}{|\mathcal{P}|} \sum_{q \in \mathcal{P}} X_q^r \cdot e_{\pi^*(0)}$, which is the accuracy for a respondent r in the question set $\mathcal{P}$, where the correct answer is produced by the chosen model without any prior knowledge.

However, this naive definition ignores the information of hierarchical level of different answers, and the hierarchical rate of different questions.

Our definition takes into account both the hierarchical rates of the questions, and the respondent's deviations from the optimal rankings. A higher hierarchical score indicates a stronger alignment between the respondent's choices and the optimal rankings.

The hierarchical score provides a quantitative measure of the respondent's performance in the optimizing framework. It can help identify respondents who consistently make choices that align with the optimal rankings, indicating a higher level of understanding and decision-making ability. The score can also be used to provide feedback to respondents, guiding them towards better decision-making and improving their performance.

3 Study

3.1 Data collection

The questionnaire we designed consists of 36 questions, where each pair of consecutive questions belonged to the same question. Odd-numbered questions required respondents to select their own options, while even-numbered questions involved predicting the answers of others. Each question provided 3 or 4 options to choose from. We send outs questionnaire through We-Chat and receive 80 valid responses.

3.2 Model accuracy

We conducted a comparison of model accuracy on the test data, considering various assumptions, as depicted in Figure 1.

Our code to test the model is posted on:

`https://git.tsinghua.edu.cn/wangweil21/information-elicitaiton-project`

Among the different approaches, our best-performing method is the Bayesian approach.

To illustrate its effectiveness, we present a case study in Figure 2. This figure provides statistical insights into the following question: "*What is the maximum number of natural numbers that can be summed up to equal 100?*" The options given are *A) 1, B) 13, C) 14, and D) NaN (not a number)*. The "Real distribution" represents the frequency of choices obtained from our questionnaire.

The "Prediction distribution" showcases type i's belief on answer prediction, indicated in the northeast region of each curve. C) is the correct answer, and interestingly, our method did indeed select C).

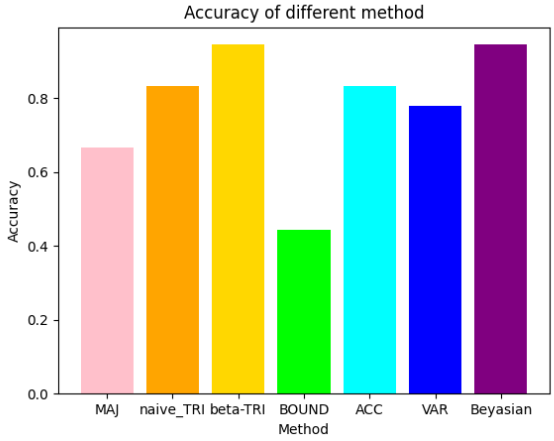


Figure 1: Prediction accuracy (**MAJ**: Assumption 0, majority voting for answer. **naive_TRI**: Assumption 1, maximize summation of the upper triangular values. **beta_TRI**: Assumption 2 with $\beta > 1$. **BOUND method**: A discarded assumption that only focusing on options near the diagonal of answer-prediction matrix. **ACC**: Assumption 4, based on prediction accuracy of each respondent types. **VAR**: Assumption 5, based on prediction variance of each respondent types, **Beyasian**: Assumption 6, Bayesian estimation of the prediction rationality.

Hierarchical score for each problem is presented in figure 3. In this figure, we can observe the subtle differences between the first nine questions and the last nine questions, which we will further explain then.

3.3 Comparison between math and general knowledge questions

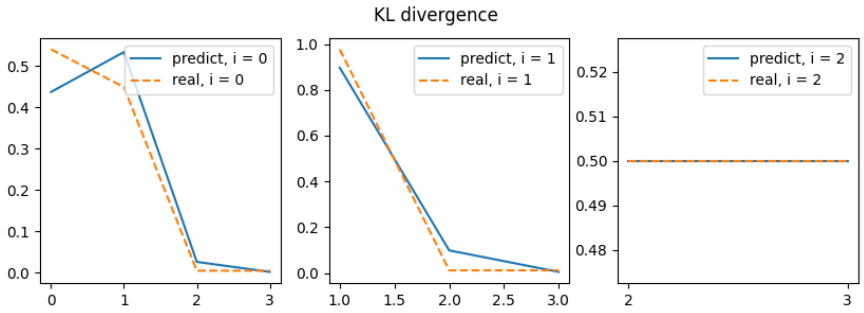
Just like investigating whether a shrewd bargainer in grocery shopping buys lottery tickets, our study focuses on assessing the disparities in difficulty between mathematical questions and general knowledge questions, as well as examining variations in rationality levels among individual respondents. The average scores, as presented in Table 1, demonstrate that the math questions included in our questionnaire generally exhibit a higher level of hierarchical difficulty compared to the general knowledge questions.

Furthermore, we calculate the hierarchical scores for each respondent separately for both math and general knowledge questions. Table 2 displays the top ten answers based on their hierarchical scores. Notably, only respondents 10, 20, and 35 appear in both leaderboards, while no individual manages to secure a position in both top-5 lists.

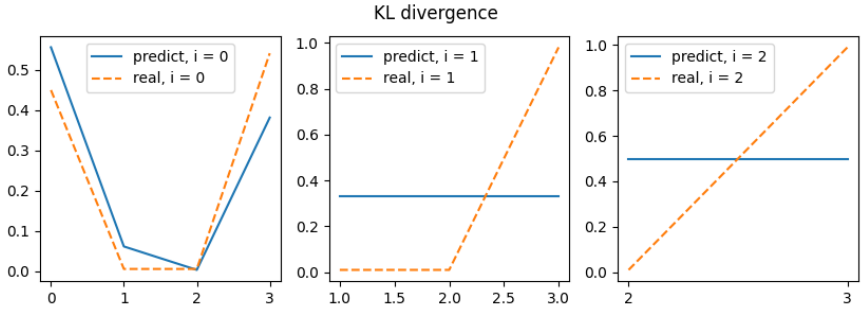
By delving deeper into these findings, we gain a more comprehensive understanding of the nuanced distinctions in difficulty and respondent rationality levels across the math and general knowledge domains.

Type	Math	Gen
rate	0.286	0.251

Table 1: Average hierarchical rate of math and general knowledge questions



(a)



(b)

Figure 2: Illustration of Bayesian method. (a) is the predicted and real distribution in a permutation with high ranking score (C B A D), and (b) is distribution of a low-score distribution (B A D C). In each image, Left, Center, and Right represent the predicted distributions and the actual distributions of rows 0, 1, and 2 in the answer-prediction matrix. $0, 1, 2, 3$ represents choices A,B,C,D respectively.

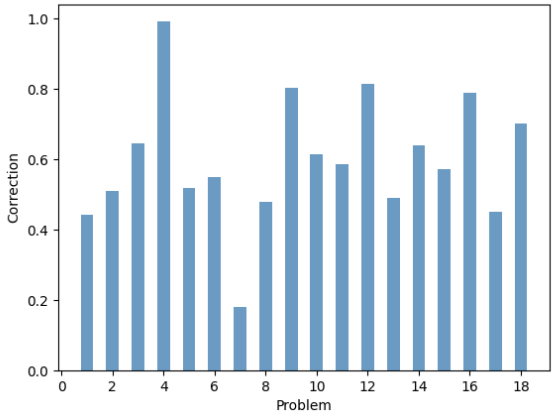


Figure 3: Smoothed hierarchical rate of math and general knowledge questions. x-axis is the index of question, y-axis is smoothed value of hierarchical rate (rate r_i for question i is transformed to $(r_i - \min_j r_j)/(\max_j r_j - \min_k r_k)$).

Rank	Math top-10	General top-10
1	33	7
2	73	65
3	36	35
4	55	49
5	10	24
6	3	20
7	20	38
8	35	79
9	34	80
10	64	10

Table 2: Top 10 respondents of math and general knowledge questions respectively.

4 Discussion

Our approach presents a novel and interpretable method for finding answers, as demonstrated by the specific cases examined. However, it is important to note that this interpretability is contingent upon real cases, and we seek to further explore the game-theoretical foundations of our proposed method. For instance, we aim to investigate potential connections between our approach and Nash equilibrium and Bayesian equilibrium. By delving into these theoretical underpinnings, we can gain deeper insights into the strategic decision-making processes involved in our methodology.

Furthermore, we acknowledge the need for scalability and generalizability of our approach. To address this, we aspire to enhance our method by leveraging larger datasets and incorporating advanced techniques like deep learning. By automating the process of answer extraction, we can improve the practicality and efficiency of our approach. This advancement would enable its application in a wide range of real-world scenarios, such as students' psychological conditions or surveys investigating people's opinions on current events and hot topics. Ultimately, our goal is to bridge the gap between theory and application, allowing our method to be effectively employed in various domains.

5 Others

Both two contribute equally in the project.

References

[1] Yuqing Kong, Grant Schoenebeck, and Katrina Ligett. *Putting Peer Prediction Under the Micro(economic)scope and Making Truth-telling Focal*. 2016. arXiv: 1603.07319 [cs.GT].

[2] Yuqing Kong et al. *Eliciting Thinking Hierarchy without a Prior*. 2022. arXiv: 2109.10619 [cs.GT].

[3] Drazen Prelec. "A Bayesian Truth Serum for Subjective Data". In: *Science (New York, N.Y.)* 306 (Nov. 2004), pp. 462–6. DOI: 10.1126/science.1102081.

[4] Dražen Prelec, Hyunjune Seung, and John McCoy. "A solution to the single-question crowd wisdom problem". In: *Nature* 541 (Jan. 2017), pp. 532–535. DOI: 10.1038/nature21054.