




Millionaire: a hint-guided approach for crowdsourcing

Bo Han^{1,2}  · Quanming Yao³ · Yuangang Pan¹ · Ivor W. Tsang¹ · Xiaokui Xiao⁴ · Qiang Yang⁵ · Masashi Sugiyama^{2,6}

Received: 25 February 2018 / Accepted: 26 September 2018 / Published online: 19 December 2018
© The Author(s) 2018

Abstract

Modern machine learning is migrating to the era of complex models, which requires a plethora of well-annotated data. While crowdsourcing is a promising tool to achieve this goal, existing crowdsourcing approaches barely acquire a sufficient amount of high-quality labels. In this paper, motivated by the “Guess-with-Hints” answer strategy from the Millionaire game show, we introduce the hint-guided approach into crowdsourcing to deal with this challenge. Our approach encourages workers to get help from hints when they are unsure of questions. Specifically, we propose a hybrid-stage setting, consisting of the main stage and the hint stage. When workers face any uncertain question on the main stage, they are allowed to enter the hint stage and look up hints before making any answer. A unique payment mechanism that meets two important design principles for crowdsourcing is developed. Besides, the proposed mechanism further encourages high-quality workers less using hints, which helps identify and assigns larger possible payment to them. Experiments are performed on Amazon Mechanical Turk, which show that our approach ensures a sufficient number of high-quality labels with low expenditure and detects high-quality workers.

Keywords Game theory · Computational modeling · Crowdsourcing · Quality control · Human factors

1 Introduction

Huge and complex models are popularly used in today’s machine learning applications, since they can take advantage of big data to get better performance. Indeed, they have significantly boosted performance of many important tasks, such as image classification (Russakovsky et al. 2015), speech recognition (Hinton et al. 2012), dialogue systems (Sordoni et al. 2015) and autonomous driving (Bojarski et al. 2016). More recently, they even beat a human champion by a large margin in the game Go (Silver et al. 2016). However, a primary question arises: how can we provide sufficient fuel (a plethora of annotated data) to propel our rocket (complex models)? The most appealing way may be the crowdsourcing technology (Rus-

Bo Han and Quanming Yao have contributed equally to this work.

Editor: Yung-Kyun Noh.

Extended author information available on the last page of the article

sakovsky et al. 2015; Zhong et al. 2015; Wang and Zhou 2016; Wang et al. 2017), since the process of annotations is convenient and the cost of annotations is very cheap.

While crowdsourcing techniques (Li et al. 2016, 2017a, b) have been commonly used in many commercial platforms, such as Amazon Mechanical Turk (AMT), the quality of crowdsourced labels is not satisfactory (Ipeirotis et al. 2010). The reasons are that workers may not be domain experts (Vuurens et al. 2011; Yan et al. 2014; Rodrigues et al. 2014). For example, it is hard for an average person to distinguish some professional tasks, such as labeling bird images or medical data (Wais et al. 2010). Besides, some workers can just be spammers, who response questions with arbitrary answers (Difallah et al. 2012; Raykar and Yu 2012). Such low-quality labels inevitably degenerates the performance of subsequent learning models (Natarajan et al. 2013; Sukhbaatar et al. 2015; Han et al. 2016). For instance, noisy labels degrade the accuracy of deep neural networks by 20% to 40% (Patrini et al. 2017; Yu et al. 2017a, b).

The previous efforts have extensively focused on statistical inferences, which aggregate crowdsourced labels when they are already collected (Karger et al. 2011; Liu et al. 2012; Chen et al. 2013; Zhou et al. 2014; Tian and Zhu 2015; Zhang et al. 2016b). However, as crowdsourced labels are intrinsically noisy, statistical inferences are hard to guarantee that aggregated labels are reliable. In order to improve the label quality, recently, many researchers resort to a complementary direction, namely proposing approaches controlling the process of label collection (Singla et al. 2015; Litman et al. 2015; Chen et al. 2016; Pennock et al. 2016; Zheng et al. 2015b; Fan et al. 2015; Han et al. 2017).

These approaches aim to encourage workers to provide more reliable labels at the stage of collection. For example, the skip-based approach encourages workers to skip uncertain tasks. However, if many tasks are difficult, the label requester may collect only a few labels, which are not enough for subsequent learning models (Shah and Zhou 2015; Ding and Zhou 2017). The self-corrected approach encourages workers to check whether they need to correct their answers after looking at references. However, references consisting of responses from other workers are noisy, which may mislead workers. Moreover, this approach is not realized on crowdsourcing platforms (Shah and Zhou 2016). Therefore, existing approaches fail to acquire a sufficient number of high-quality labels on real tasks (Table 1). Besides, these approaches cannot detect and give potentially larger payment to high-quality workers. However, this is very important as these workers are always preferred by crowdsourcing platforms, thus they should be identified and more paid (Ipeirotis et al. 2010).

To address these issues, we are inspired by the “Guess-with-Hints” answer strategy from the Millionaire game show,¹ where a challenger has opportunities to request hints from the show host when he/she feels unsure of the questions. By this strategy, we introduce a hint-guided approach to improve the quality of crowdsourced labels. This approach encourages workers to get help from auxiliary hints when they answer questions that they are unsure of. To be specific, we introduce a hybrid-stage setting, which consists of the main stage and the hint stage. In the main stage, for each question, workers answer it directly when they feel confident or jump into the hint stage when they feel uncertain. Once they enter the hint stage, they are allowed to look up hints before making any answer to this unsure question. The less number of times workers enter the hint stage, the higher quality they are estimated to be. To realize this setting, we provide an explicit “? & Hints” button (the bottom panel in Fig. 1) for each question. For example, when the worker is unsure of the question in Fig. 1, he/she can click this button and answer the question under the help of hints (the gray sentence).

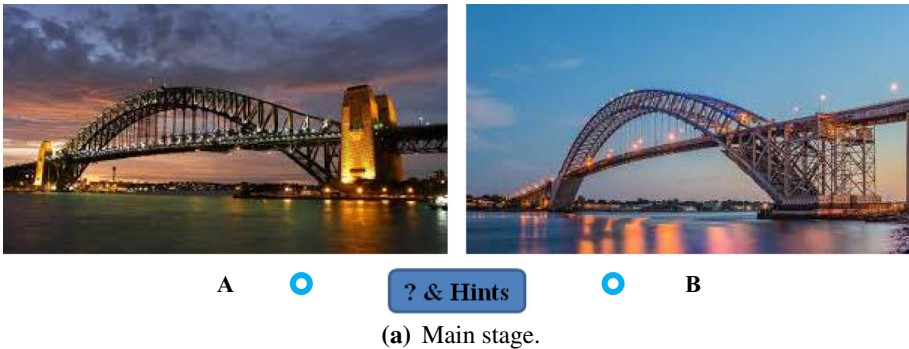
¹ [https://en.wikipedia.org/wiki/Who_Wants_to_Be_a_Millionaire_\(U.S._game_show\)](https://en.wikipedia.org/wiki/Who_Wants_to_Be_a_Millionaire_(U.S._game_show)).

Table 1 Comparison of related approaches and our hint-guided approach (in bold)

Perspective	Metric	Baseline	Skip-based (Shah and Zhou 2015; Ding and Zhou 2017)	Self-corrected (Shah and Zhou 2016)	Hint-guided (proposed)
Requester	Large label quantity	✓	✗	✓	✓
	High label quality	✗	✓	–	✓
Worker	Quality detection	✗	✗	✗	✓
	Spammer prevention	✗	✓	✓	✓
Platform	Low money cost	✗	✓	–	✓
	Realization	✓	✓	✗	✓

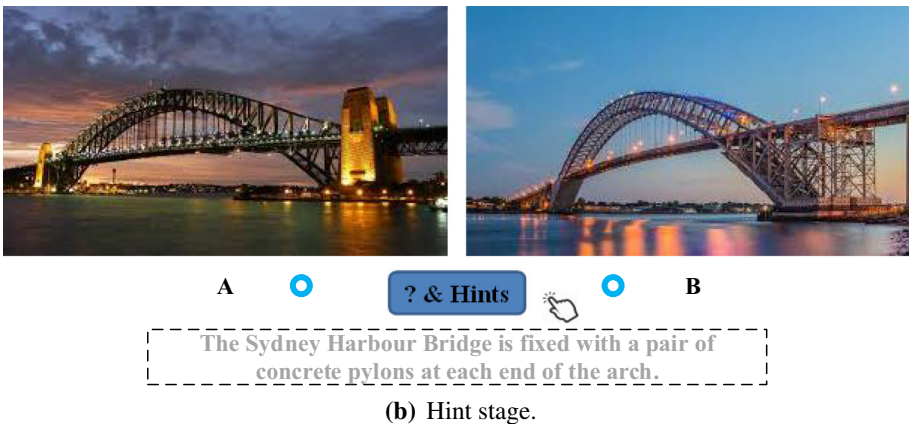
Baseline is the approach used by AMT (details are in “Appendix B”). Note that, the self-corrected approach is designed theoretically, but barely realized for real tasks (denoted as “✗”), so its metrics of “label quality” and “money cost” cannot be evaluated (denoted as “–”). However, since its payment mechanism has a multiplicative form, it prevents spammers theoretically

Which one is the Sydney Harbour Bridge ?



(a) Main stage.

Which one is the Sydney Harbour Bridge ?



(b) Hint stage.

Fig. 1 A task that requires workers to answer the question “Which one is the Sydney Harbour Bridge?”. Top panel: the proposed interface under the hybrid-stage setting, consists of two options (“A” and “B”) and a “? & Hints” button. Bottom panel: when workers feel unsure of this question and click the button, the content of hints (gray) is visible, which guides workers to make a choice

Nevertheless, only hybrid-stage setting is not enough to address all issues. For example, if hints are freely available in the hint stage, even high-quality workers may abuse free hints for higher accuracy and rewards. This issue causes failure in the detection of high-quality workers. Under the hybrid-stage setting, we develop a hint-guided payment mechanism, which aims to incentivize workers to use the hints properly. Specifically, our mechanism penalizes workers who use the hints. Therefore, high-quality workers will answer most of the questions directly (without hints) for higher rewards. Then, our mechanism assists our setting to detect the high-quality workers effectively. Moreover, we prove that our mechanism is unique under the hybrid-stage setting. Since our mechanism has a multiplicative form, it prevents spammers as well. Our contributions are summarized as follows.

- In crowdsourcing, our hint-guided approach is the first attempt to improve the quality of labels by auxiliary hints, and detect the high-quality workers. Our approach is different in both setting and payment mechanism from existing approaches such as self-corrected and skip-based approaches.
- We introduce a hybrid-stage setting. Under this setting, we propose a hint-guided payment mechanism, which incentivizes workers to use hints properly instead of abusing them. Moreover, we prove the uniqueness of our mechanism under the proposed setting.
- We further give some general rules for task requester, which helps them easily design hints for their own tasks.
- Unlike many machine learning papers in crowdsourcing, which do not or cannot perform experiments on real datasets, we conduct comprehensive real-world experiments on AMT platforms. Empirical results on three real tasks show that, the proposed approach reaches an excellent performance on the adequate collection of high-quality labels using low expenditure. Meanwhile, our approach prevents spammers and detects high-quality workers as well.

The remainder of this paper is organized as follows. In Sect. 2, related literature is presented. Section 3 introduces the novel setup in crowdsourcing, namely the hybrid-stage setting. In Sect. 4, we propose a hint-guided payment mechanism under this setting. In Sect. 5, we provide the experiment setup and empirical results related to three real-world tasks. The conclusions are given in Sect. 6.

2 Related literature

2.1 Post-processed approach

In crowdsourcing, the statistical inference (post-processed) approach is popularly used to improve the quality of labels (Zheng et al. 2015a, 2016, 2017, Zhang et al. 2016a). Such approach tries to find the correct label for each question only after noisy labels being collected from the platform. Many methods have been developed under this approach.

For example, Raykar et al. (2010) presented a two-coin probabilistic model, where each worker's labels are generated by flipping the ground-truth labels with a certain probability. Yan et al. (2010) extended this two-coin model by setting the dynamic flipping probability associated with samples. Kajino et al. (2012) formulated a probabilistic multi-task model, where each worker is considered as a task. Zhou et al. (2012) proposed a minimax entropy model. Bi et al. (2014) employed a mixture probabilistic model for worker annotations, which learns a prediction model directly. Tian and Zhu (2015) extended weighted Majority Voting by the max-margin principle, which provides a geometric interpretation of crowdsourcing

margin. However, as labels are intrinsically noisy, it is hard for this type of approach to obtain a sufficient amount of correct labels with statistical inference.

2.2 Pre-processed approach

While previous efforts have extensively focused on several statistical inferences, pre-processing approach has been recently developed as an alternative way to improve label quality. Namely, the crowdsourced setting is coupled with the payment mechanism, which incentivizes workers to provide more reliable labels at the stage of label collection. Thus, unlike post-processed approach, pre-processed one can directly reduce the noise in obtained labels. Moreover, post-processed approach can be used to further reduce the noise in labels after they are obtained by the pre-processed approach.

In this paper, we target the pre-processed approach from the perspective of machine learning (Buhrmester et al. 2011; Singla and Krause 2013; Goel et al. 2014; Ho et al. 2015; Lambert et al. 2015; Shah and Zhou 2015; Ding and Zhou 2017). The most related works are the skip-based (Shah and Zhou 2015; Ding and Zhou 2017) and self-corrected approaches (Shah and Zhou 2016). In the skip-based approach, workers are allowed to select a skip option based on their confidence for each question. However, this in turn leads to insufficient label quantity. A two-stage setting is used in the self-corrected approach. Workers firstly answer all questions in the first stage, and then they are allowed to correct their first-stage answers after looking at a reference in the second stage. However, references consisting of responses from other workers are noisy, which may mislead workers to providing incorrect labels. Besides, as a reference needs to be set for each task, such a setting is not supported by the AMT platform and only simulation results are reported in Shah and Zhou (2016). Finally, neither the skip-based nor self-corrected approaches can identify worker quality as our approach.

The pre-processed approach was also considered in the database area, but the focus is different. Normally, their research is to dynamically assign the optimal K ($\leq N$) problems to each worker by his/her work quality, where N is the total number of problems to be annotated (Zheng et al. 2015b; Fan et al. 2015; Hu et al. 2016). Thus, worker quality control plays a fundamental role in the quality of crowdsourcing from the viewpoint of database.

2.3 Worker quality control

As workers' quality has huge impact on the obtained labels, many researchers tried to improve label quality by offering better control over workers' quality. For example, Raykar and Yu (2012) considered detecting spammers or adversarial behavior, and tried to eliminate them in the following iterations or phases. However, this method does not consider how to detect high-quality workers. Then, Joglekar et al. (2013) devised techniques to generate confidence intervals for worker error rate estimates, thereby enabling a better evaluation of worker quality. However, this method is complex to be deployed. For our hybrid-stage setting, the less number of times workers enter the hint stage, the higher quality they are estimated to be.

3 Problem setup

Inspired by the "Guess-with-Hints" answer strategy, we introduce the hint-guided approach to improve the quality of crowdsourced labels and detect the high-quality workers at the same

time. This approach encourages workers to get help from the useful hints when they answer uncertain questions (Fig. 1). Specifically, we realize this approach in Sect. 3.1, including the hybrid-stage setting and the payment mechanism. Then, easy usage of hints is discussed in Sect. 3.2. Finally, the rationality of our design is discussed in Sect. 3.3.

3.1 Hint-guided approach

Here, we describe our hint-guided approach from the following four aspects.

3.1.1 Hybrid-stage setting

We first set up definitions for the hybrid-stage setting that consists of the main stage and the hint stage. To model our setting, let us consider a simple example: each worker answers N binary-valued (objective) questions, and each question has precisely one correct answer, either “A” or “B”. Therefore, for every question $i \in \{1, \dots, N\}$, a worker chooses an answer matching his/her own belief under the following hybrid-stage setting.

- The main stage (Fig. 1a): For question i , he/she should be incentivized to select the option that he/she feels confident. When he/she feels unsure and clicks the “? & Hints” button, he/she jumps into the hint stage formalized by the “H” option, namely,

$$\text{select} \begin{cases} \text{“A”} & \text{if } P_{A,i} \in [\frac{1}{2} + \epsilon, 1), \\ \text{“B”} & \text{if } P_{A,i} \in (0, \frac{1}{2} - \epsilon], \\ \text{“H”} & \text{otherwise,} \end{cases}$$

where $\epsilon \in [0, \frac{1}{2})$ models the worker’s uncertainty degree in this stage, $P_{A,i}$ is the probability of the worker’s belief that the answer to the i th question is “A” (i.e., the probability that the worker believes “A” is the correct answer for the i th question).

- The hint stage (Fig. 1b): When he/she feels unsure of the question, the worker clicks the “? & Hints” button. This means that he/she enters the hint stage. Then, the worker picks up “A” or “B” according to

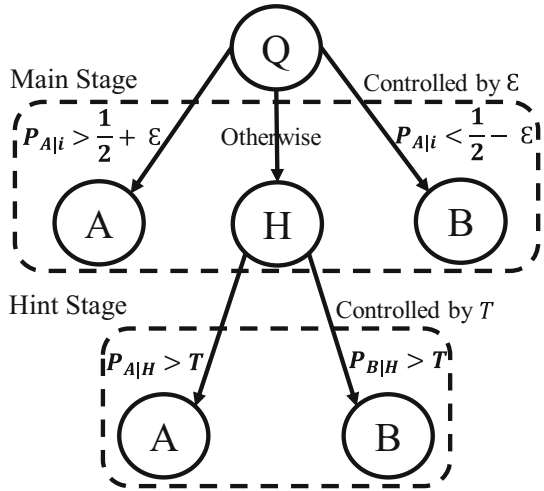
$$\text{select} \begin{cases} \text{“A”} & \text{if } P_{A|H,i} \in [T, 1), \\ \text{“B”} & \text{if } P_{B|H,i} \in [T, 1), \end{cases}$$

where $T \in (\frac{1}{2}, 1)$ is the predefined threshold value of the worker’s belief in the hint stage, $P_{A|H,i}$ is the probability of the worker’s belief that the answer to the i th question is “A” given hints, and $P_{B|H,i}$ is the probability of the worker’s belief that the answer to the i th question is “B” given hints ($P_{B|H,i} = 1 - P_{A|H,i}$).

The above modeling of the decision process is also summarized in Fig. 2. As we can see, ϵ controls the decision in the main stage and the hint stage depends on T . When ϵ is large, i.e., $\epsilon \rightarrow \frac{1}{2}$, more workers need hints to make their decision for each question. When ϵ is smaller, i.e., $\epsilon \rightarrow 0$, fewer workers need hints to make their decision for each question. Once the worker enters the hint stage, when T is set to a large value, i.e., $T \rightarrow 1$, he/she will become more confident to make his/her final decision for each question. When T is set to a small value, i.e., $T \rightarrow \frac{1}{2}$, he/she will be less confident to make his/her final decision for each question.

Note that, ϵ is decided by T according to Proposition 2 in Sect. 4.2, and T is controlled by a mechanism designer. The choice of T is based on different applications and given to us. In

Fig. 2 Mathematical model of the decision process under our hybrid-stage setting



the experiments, we empirically choose $T = 0.75$ due to the qualitative psychology (Smith 2007).

3.1.2 Model assumption

Based on the hybrid-stage setting, we will introduce the corresponding payment mechanism, where it is rooted in the following assumption.

- Assumption 1 (A)** There are G “gold standard” questions ($1 \leq G \leq N$), of which answers are known to the requester, uniformly distributed at random positions among all N questions;
- (B)** Each worker aims to maximize his/her expected payment for N questions;

Assumption 1 is a standard one in analyzing pre-processed approaches for crowdsourcing (Shah and Zhou 2015, 2016; Zhang et al. 2016a). Specifically, as answers to the “gold standard” questions are known to the requester in advanced, workers’ responses to them can be used to evaluate workers’ performance and decide payment for workers. This is the functionality of Assumption 1(A). Then, Assumption 1(B) is a must for analyzing workers’ performance. It originates from game theory (Nisan et al. 2007), and means that each work wants to maximize its revenue.

Next, we make the following Assumption 2, which specifies our usage of hints here. It is motivated by the educational psychology (Koedinger and Aleven 2007), and means that the hints are useful enough to guide workers to making final decisions.

Assumption 2 Workers have enough confidence to make a final decision after acquiring useful hints, i.e., $T \in (\frac{5}{8}, 1)$ in the hint state.

Note that the confidence of a random guess is $T = \frac{1}{2}$, thus $T > \frac{5}{8}$ means that the worker’s confidence to pick up an answer is high after looking at the hint. This value ($5/8$) is related to the proof of Corollary 1. As an illustration, let us see Fig. 1a. Workers outside Australia may not know which one is the Sydney Harbour Bridge. However, after reading the hints (grey) in Fig. 1b, workers should have enough confidence to make a final decision “A” as the pylons structure is very obvious. When T approaches to 1, the beliefs from the hint are maximal, or equivalently, the hint provides the worker with a certain answer.

3.1.3 Payment mechanism

According to the model assumption, we are ready to introduce our payment mechanism based on the hybrid-stage setting. Specifically, after the worker answers all N questions in the hybrid-stage setting, his/her performance is evaluated by his/her responses to G ($\leq N$) questions. Namely, his/her choice for each question in the gold standard gets evaluated to one of four states, denoted by $\{\mathbb{D}_+, \mathbb{D}_-, \mathbb{H}_+, \mathbb{H}_-\}$. We define the four states as follows.

- \mathbb{D}_+ : answer in the main stage and correct;
- \mathbb{D}_- : answer in the main stage and incorrect;
- \mathbb{H}_+ : answer in the hint stage and correct;
- \mathbb{H}_- : answer in the hint stage and incorrect.

Note that “answer in the main stage” means that he/she feels confident in the main stage and answers directly; “answer in the hint stage” means that he/she feels unsure in the main stage and answers with hints in the hint stage. “correct” or “incorrect” denotes whether the worker’s selection matched with the standard answer in G questions or not.

Therefore, under the hybrid-stage setting, we can formulate any payment mechanism as function

$$f : \{\mathbb{D}_+, \mathbb{D}_-, \mathbb{H}_+, \mathbb{H}_-\}^G \rightarrow [\mu_{\min}, \mu_{\max}], \quad (1)$$

where $\min f(\cdot) = \mu_{\min}$ and $\max f(\cdot) = \mu_{\max}$. We reserve the rights to set μ_{\min} and μ_{\max} , where $0 \leq \mu_{\min} \leq \mu_{\max}$. In our paper, the goal is to design f such that its expected payment for each worker is strictly maximized under the above setting.

3.1.4 Difference from previous approaches

The most related approach to ours is the self-corrected approach (Shah and Zhou 2016), since both of us have two phases in the setting. However, they are totally different in probabilistic modeling. The self-corrected approach builds up the two-stage setting, and workers are necessarily required to enter the second stage to check the reference answer of every question, whereas, our approach builds up the hybrid-stage setting, and workers are not necessary to enter the hint stage related to confident questions. Besides, since each payment mechanism is customized based on some designed goals (examples are in Sect. 4.1) under its corresponding setting, our hint-guided payment mechanism is also different from the one used in the self-corrected approach.

It would be also interesting to discuss the advantages of the proposed approach over active learning (Yan et al. 2011) for crowdsourcing. There are two points to be highlighted. First, compared with active learning, hints in our approach may not be as strong as querying the ground-truth label; the hint only guides the worker to make a choice. Second, active learning is constrained to query which data sample should be labeled next and which annotator should be queried to benefit the learning model. However, our approach is free of these restrictions.

3.2 General rules of hints

Motivated by instructional hints in the educational psychology (Koedinger and Alevan 2007), to make the hints useful and reduce interface designers’ workloads, we offer three general rules here:

- (A) The hints should be easily accessible to interface designers;
- (B) The hints should be discriminative and concise for workers; and
- (C) The hints should be irrelevant to the number of annotated samples in each task.

We adopt the three rules in designing our hints in experiments. We take three practical datasets in our experimental setup (Sect. 5.1) to justify that these requirements are reasonable in the real world. First, for *Sydney Bridge*, as an interface designer, we easily acquire the content of hints from Wikipedia, which includes discriminative and concise phrases, such as “concrete pylons” and “around Sydney Opera House”. Second, for *Stanford Dogs*, we build a lookup table as hints, which includes the characteristics of four breeds of dogs, such as prick ears for Norwich Terrier. It means that, the hints in this dataset should be irrelevant to the number of annotated samples, but relevant to the number of classes. Third, for *Speech Clips*, the tool is freely available online to roughly recognize each speech clip and save the concise keywords (≤ 4) as the hints.

3.3 Needs of hybrid-stage setting

It is also noted that, in designing the pre-processing mechanism, the high-quality worker detection is very important for collecting a sufficient number of high-quality labels. If the tasks can be assigned to each worker by his/her work quality, the annotation quality will be increased accordingly. Also, if we can detect the high-quality workers and give more weights on his/her annotations, we can acquire the better label aggregation. Here, we show that it may not be achieved by a single-stage setting with hints (i.e., only Fig. 1b and no Fig. 1a). Later, we also empirically demonstrate this point in Sect. 5.2.3.

Specifically, by our Assumption 2, if we want to collect more correct labels, it is more naturally to directly assign visible hints for every single question. This removes the necessity to have a hybrid-stage setting as we design here. However, high-quality workers are always preferred by crowdsourcing platforms, thus they should be identified and more paid. Such a fundamental goal may not be achieved by a simple single-stage setting with visible hints. The reason is explained as follows. Under the single-stage setting, both high-quality and low-quality workers can easily read the visible hints to answer questions. Thus, we cannot make a difference between them. However, under the hybrid-stage setting, high-quality workers may not read the hints frequently. Namely, the less number of times workers enter the hint stage, the higher quality they are estimated to be. Thus, we can track the high-quality workers by our setting.

Note that, only this setting may encounter a problem: if the hints are freely available in the hint stage, by Assumption 1(B), even high-quality workers may abuse free hints for higher accuracy and rewards. This issue causes failure in the detection of high-quality workers by the hybrid-stage setting. Therefore, under the hybrid-stage setting, we hope to develop a payment mechanism (Sect. 4), which incentivizes workers to use the hints properly. Specifically, this mechanism penalizes workers who use the hints. Then, high-quality workers will answer most of the questions directly for higher rewards. As a result, this mechanism helps our setting to detect the high-quality workers effectively.

4 Hint-guided payment mechanism

In Sect. 4.1, we first give two important definitions which help us to design a payment function. Then, the designed payment function is given in Sect. 4.2. Furthermore, we prove

that our incentive-compatible payment mechanism is also unique under the hybrid-stage setting. Finally, in Sect. 4.3, we clarify that more restrictive designing goals cannot be realized here.

4.1 Design principles

Incentive compatibility (Definition 1) and mild no-free-lunch axiom (Definition 2) are important to design a payment mechanism for pre-processed approaches, which are also popularly used by previous works (Shah and Zhou 2015, 2016).

Definition 1 (*Incentive compatibility*) A payment mechanism f is incentive-compatible only if the following two conditions are satisfied: (i) f gives an incentive to a worker to choose all answers by his/her belief; (ii) The expected payment, from the worker’s belief, is strictly maximized in both the main stage and the hint stage.

Definition 1, which is adapted from the standard game theoretical assumption (Nisan et al. 2007), describes incentive compatibility. Basically, it means that f should encourage a worker to select the option he/she believes most likely to be correct.

Definition 2 (*Mild no-free-lunch axiom*) If all answers attempted by a worker in “gold standard” questions are either wrong or based on hints, then the payment for the worker should be zero, unless all answers attempted by the worker are correct. More formally, $f(\mathbf{a}) = 0, \forall \mathbf{a} \in \{\mathbb{D}_-, \mathbb{H}_+, \mathbb{H}_-\}^G \setminus \{\mathbb{H}_+\}^G$.

Definition 2 is a variant of the no-free-lunch axioms for our hybrid-stage setting. It requires that f should not pay a worker who has bad performance on “gold standard” questions. This helps to reject spammers and keep high-quality workers, since answers to these questions are known to the platform and spammers are likely to give wrong answers while high-quality workers are not.

Our aim is to design the payment mechanism f , which is defined in Eq. (1), simultaneously satisfies the above definitions.

4.2 Proposed payment mechanism

In order to design a payment mechanism, we first consider the easiest case, i.e., for a single question, how the worker should get paid under our hybrid-stage setting. This helps us to find specific rules under Definition 1 for our setting under Assumption 1, such rules are given in Proposition 1 below, and its proof is in Appendix A.1.

Proposition 1 Let $f : \{\mathbb{D}_+, \mathbb{D}_-, \mathbb{H}_+, \mathbb{H}_-\} \rightarrow [0, \mu_{max}]$, $d_+ = f(\mathbb{D}_+)$, $d_- = f(\mathbb{D}_-)$, $h_+ = f(\mathbb{H}_+)$ and $h_- = f(\mathbb{H}_-)$. When $N = G = 1$, f satisfies Definition 1 if it meets the following pricing constraints:

- (A) $d_+ > d_-, h_+ > h_-, d_+ > h_+$.
- (B) $\frac{d_+ - d_-}{1 - 2\epsilon} \geq \frac{h_+ - h_-}{2\epsilon}$.
- (C) $d_+ - d_- \leq \frac{2T - 1}{1/2 - \epsilon} (h_+ - h_-)$.

Condition (A) highlights that, for each question, the payment h_+ from an indirect correct answer (after reading hints) should be less than d_+ from a direct correct answer. Condition (B) bridges the per unit income gap $\frac{d_+ - d_-}{1 - 2\epsilon}$ in the main stage and $\frac{h_+ - h_-}{2\epsilon}$ in the hint stage together, and the inequality encourages a worker to directly answer questions that he/she feels

confident about in the main-stage. Condition (C) incentivizes his/her to leverage the hints before answering questions that he/she is unsure of. Thus, conditions (A) and (B) encourage workers to directly answer questions without hints if they are confident enough; and when a worker has really low confidence, condition (C) encourages him/her to use hints.

Remark 1 In condition (A), we cannot set $d_+ = h_+$. If $d_+ = h_+$, even high-quality workers may abuse hints for higher accuracy and rewards. This issue fails the detection of high-quality workers by the hybrid-stage setting. Therefore, $d_+ > h_+$ ensures that high-quality workers will answer most of the questions directly for higher rewards. Under the hybrid-stage setting, whether hints are used can be taken as a criterion to detect the high-quality workers. Then, $d_+ > h_+$ assists our setting to detect the high-quality workers, which has been verified in experiments in Sect. 5.2.3.

From Proposition 1, we can see that f relies on workers' uncertainty degree ϵ in the main stage and their confidence T in the hint stage. When ϵ is set to a large value, more workers need hints to make their decision for each question. The disadvantage of large ϵ is that the overall payments for workers may be low due to leveraging too many hints. When ϵ is set to a small value, fewer workers need hints to make their decision for each question. The disadvantage of small ϵ is that the quality of crowdsourced labels may be poor since more workers avoid hints for higher payments. Thus, we need to find ϵ to achieve a good tradeoff such that most workers are balanced, neither too cautious nor too careless.

However, Proposition 1 only makes use of Assumption 1 to find rules for f and does not specify the relationship between ϵ and T . Below Proposition 2 helps to connect ϵ and T , and shows a lower-bound of ϵ . Its proof is in Appendix A.2.

Proposition 2 *Under Assumption 1, f satisfies both Definitions 1 and 2 if $\epsilon \in [\epsilon_{\min}, 1/2)$ where $\epsilon_{\min} = T - \sqrt{T^2 - 1/4}$.*

Moreover, based on above Proposition, we can derived following Corollary which is based on Assumption 2. Its proof is in "Appendix A.3".

Corollary 1 *Under Assumption 2, $(1/2 - \epsilon_{\min}) < (2T - 1)$.*

Finally, we show when $\epsilon = \epsilon_{\min}$, i.e., the boundary condition in Proposition 2 is achieved, a hint-guided payment mechanism f can be designed (Algorithm 1). The function g , which sets how a single question should be paid, is defined at step 1 in Algorithm 1. Note that $g(\mathbb{H}_+) < g(\mathbb{D}_+) = 1$ due to Corollary 1, which is also in consistent with condition (a) in Proposition 1. Responses from workers on "gold standard" questions are collected in step 2, and the budget is set in step 3. A multiplicative form of g is adopted in step 4, which is inspired by Shah and Zhou (2015). It incentivizes workers to use hints properly and also helps to make the smallest payment to spammers. The reasons are highlighted in Remark 2.

Remark 2 The benefits of using the multiplicative form is detailed as follows. For example, a spammer will respond to a question with an arbitrary answer, thus he/she will get the minimum payment once any answers in "gold standard" are wrong. Then, for a normal worker, if he/she tries to get the highest payment, he/she is encouraged to use hints as less as possible. The reason is that the payment for a correct answer after using hints is $g(\mathbb{H}_+)$ which is smaller than 1, i.e., $g(\mathbb{D}_+)$ (Corollary 1). Thus, more hints are used, the maximum payment for a worker will get smaller. Besides, such a multiplicative form also helps us to identify and pay more for high-quality workers, as those workers will naturally user less hints.

The design of Algorithm 1 is further supported by the following Theorem 1. Its proof is in "Appendix A.4". Thus, our algorithm is the unique one to satisfy both Definitions 1 and

Algorithm 1 Hint-guided payment mechanism**Inputs:**

1. Define a function $g : \{\mathbb{D}_+, \mathbb{D}_-, \mathbb{H}_+, \mathbb{H}_-\} \rightarrow \mathbb{R}_+$, which sets how a single question should be paid, and, where $g(\mathbb{D}_+) = 1$, $g(\mathbb{D}_-) = 0$, $g(\mathbb{H}_+) = \frac{1/2 - \epsilon_{\min}}{(2T-1)}$ and $g(\mathbb{H}_-) = 0$;
2. $a_1, \dots, a_G \in \{\mathbb{D}_+, \mathbb{D}_-, \mathbb{H}_+, \mathbb{H}_-\}$ are the state evaluations of the answers to the G gold standard questions;
3. Set the minimum payment μ_{\min} and maximum payment μ_{\max} properly;

The payment is:

4. $f([a_1, \dots, a_G]) = \beta \prod_{i=1}^G g(a_i) + \mu_{\min}$ where $\beta = \mu_{\max} - \mu_{\min}$.

2, and $\epsilon = \epsilon_{\min}$ is also a must choice here. Note that, in practice, the algorithm makes the minimum payment μ_{\min} instead of 0 in Definition 2, if one or more attempted answers in the gold standard are wrong. This operation is without any loss of generality.

Theorem 1 *Under Assumptions 1 and 2, f in Algorithm 1 satisfies both Definitions 1 and 2 if and only if $\epsilon = \epsilon_{\min}$.*

4.3 No other compatible mechanism

Definition 1 is a must to design a payment mechanism. However, under our hybrid-setting here, there exists another popular “harsh no-free-lunch” axiom (Definition 3), which is adapted from Definition 2 in Shah and Zhou (2016).

Definition 3 (*Harsh no-free-lunch axiom*) If all answers attempted by the worker in “gold standard” questions are either wrong or based on hints, then the payment for the worker should be zero. More formally, $f(\mathbf{a}) = 0$, $\mathbf{a} \in \{\mathbb{D}_-, \mathbb{H}_+, \mathbb{H}_-\}^G$.

Compared to the “mild no-free-lunch” axiom, Definition 3 encourages the worker to answer without hints no matter whether he/she is unsure. Thus, it is stronger than the “mild no-free-lunch” axiom and can be used to replace Definition 2. We wonder whether we can find another payment function which satisfies this more restrictive condition. However, below Theorem 2 shows a contradiction to Definition 3. Its proof is in Appendix A.5.

Theorem 2 *Under Assumptions 1 and 2, there is no mechanism that satisfies both Definitions 1 and 3.*

Therefore, the “harsh no-free-lunch” axiom is too strong for the existence of any incentive-compatible payment mechanism here. This further illustrates the uniqueness of the proposed payment mechanism.

5 Numerical experiments

We conduct real-world experiments on Amazon Mechanical Turk,² which is the leading platform to collect crowdsourced labels. We compare our hint-guided approach with: (1) Baseline approach : a single-stage setting with an additive payment mechanism (details are in “Appendix B”). (2) Skip-based approach (Shah and Zhou 2015; Ding and Zhou 2017): a

² <https://www.mturk.com/mturk/welcome>.

skip-stage setting with a skip-based payment mechanism. Note that the skip-based payment mechanism is multiplicative. For the self-corrected approach (Shah and Zhou 2016), it has not been verified on AMT tasks, since there is no criteria how to set references. Therefore, we do not include it in our comparison. Note that additive and multiplicative payment mechanisms are respectively denoted as “+” and “×” for subsequent use in Sect. 5.2.4.

5.1 Experimental setup

All these datasets are collected by us on Amazon MTurk, where hints are easily designed according to the criteria in Sect. 3.2. We conducted three real tasks as follows.

- *Sydney Bridge* (binary-choice questions): we collect 30 images of various bridges. Each image contains one bridge. The task is to identify whether the bridge in each image is the Sydney Bridge. The content of hints includes discriminative phrases, such as “concrete pylons” and “around Sydney Opera House”.
- *Stanford Dogs* (multiple-choice questions): we collect 100 images of four breeds of dogs. The task is to identify the breed of dogs in each image. We build a lookup table as hints, which includes the characteristics of four breeds of dogs, such as “prick ears” for Norwich Terrier.
- *Speech Clips* (subjective questions): we collect 10 speech clips. Each speech clip consists of 1 or 2 short sentences (15 words). The task is to recognize each speech clip and write down the corresponding sentence. We leverage the open tool³ to roughly recognize each speech clip and save the key words (≤ 4) as the hints.

We verify the effectiveness of our hint-guided approach from three perspectives (Table 1), and each perspective includes one to two metrics in brackets: requester (“label quantity” and “label quality”), worker (“worker quality detection” and “spammer prevention”) and platform (“money cost”). Except “worker quality detection”, other metrics have been popularly used by previous works (Shah and Zhou 2015, 2016). They are detailed as follows.

- Label quantity: we evaluate the label quantity by the percentage of the completion of three tasks. In the skip-stage setting, worker yields unlabeled (uncompleted) data by skipping unsure questions. In the single-stage and the hybrid-stage settings, for objective questions, worker yields (few) unlabeled data because he/she forgets or ignores few questions. For subjective questions, worker yields (more) unlabeled data by inputting invalid answers. For example, they write sentences, such as “I do not know” in the answer box.
- Label quality: we evaluate the label quality from two aspects: (i) the percentage of correct answers and incorrect answers on three tasks; and (ii) the error of aggregated labels (Shah and Zhou 2015). For the i th question where $i \in \{1, \dots, n\}$, if there are m_i options after majority voting (the tie situation), and the ground-truth label is one of m_i options, then we consider that $\frac{1}{m_i}$ of the i th question is correct. Therefore, the error of aggregated labels is $1 - (\sum_{i=1}^n 1/m_i)/n$. Since text answers cannot be majority voted on *Speech Clips*, we do not report the error of aggregated labels on *Speech Clips*.
- Worker quality detection: we evaluate the worker quality detection of the hint-guided approach implicitly, by the error rate (in %) of aggregating original and rescaled crowd-sourced labels. For example, *Sydney Bridge* (origin) means the original labels collected by our approach. For *Sydney Bridge* (rescale), we rank the worker quality from high to low by the usage frequency of the hints in the collection of original labels. Then, we

³ <https://speech-to-text-demo.mybluemix.net/>.

Table 2 Evaluation of the label quantity. We provide the percentage of the completion on three tasks

Data set	Baseline (%)	Skip-based (%)	Hint-guided (%)
<i>Sydney Bridge</i>	100.00	74.00	99.11
<i>Stanford Dogs</i>	99.72	58.18	99.91
<i>Speech Clips</i>	58.33	30.00	75.00

The best results are highlighted in bold

rescale original labels by adaptive weights. Labels from top 20% (bottom 20%) workers have been empirically rescaled by 1.8 (0.2). The remaining labels keep unchanged. If the error rate on rescaled dataset decreases, then we speculate that our hint-guided approach indeed detects the worker quality. Namely, the less usage of hints indicates the higher quality of the worker.

- Spammer prevention and money cost: we evaluate the spammer prevention and the money cost by the average payment to each worker. Note that the payment consists of two parts: fixed payment and reward payment. Reward payment is based on a worker’s responses to G gold standard questions. All payment parameters are in “Appendix C”.

5.2 Experimental results

We demonstrate the effectiveness of our hint-guided approach from the following five aspects. Specifically, Sect. 5.2.1 verifies whether our approach provides a sufficient number of labels. Section 5.2.2 displays whether our approach provides high-quality labels. Section 5.2.3 denotes whether our approach can detect worker quality. Section 5.2.4 indicates whether our approach prevents spammers. Section 5.2.5 demonstrates whether our approach saves money.

5.2.1 Label quantity

Table 2 denotes the percentage of the completion of three tasks. The first two tasks (*Sydney Bridge* and *Stanford Dogs*) belong to objective questions, while the last task (*Speech Clips*) belongs to subjective questions. Objective questions can be answered by the random guess. Therefore, the percentage of the completion for objective questions is much higher than that for subjective questions. In addition, the hint-guided approach has a high percentage of the completion of both objective and subjective questions. Our approach inspires workers to finish the questions, ensuring the quantity of crowdsourced labels.

5.2.2 Label quality

Figure 3 plots the percentage of correct answers and incorrect answers on three tasks. First, on all tasks, the percentage of correct answers in the hint-guided approach is higher than that in the baseline and skip-based approaches. Second, on *Speech Clips*, the percentage of incorrect answers is extremely low in the skip-based approach. The reason is that most people skip difficult speech clips, and answer several easy ones. Third, compared with other approaches, our hint-guided approach ensures a sufficient number of high-quality labels.

Figure 4a, b plot the error of aggregated labels on the *Sydney Bridge* and *Stanford Dogs* tasks. The number of workers (abbreviated as $n_{workers}$) is set to $\{5, 6, 7, 8, 9, 10\}$, since the error of aggregated labels comes from majority voting among multiple workers (Shah

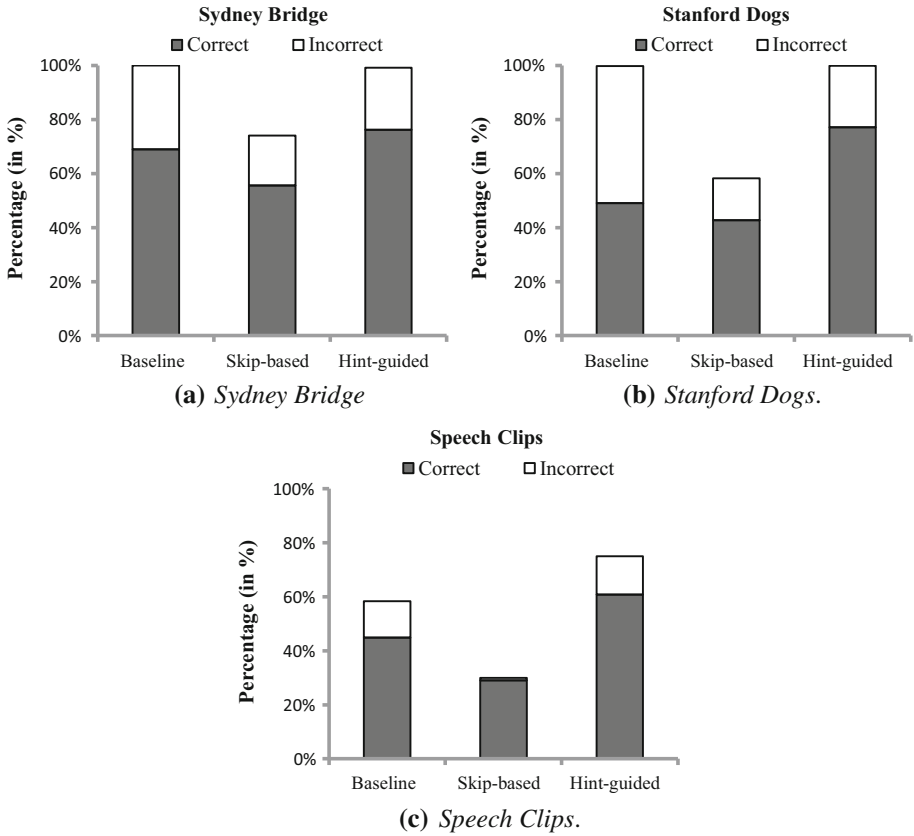


Fig. 3 Evaluation of label quality. Percentage (in %) of correct answers and incorrect answers on three tasks are provided. Note that, we do not plot the percentage of unlabeled questions

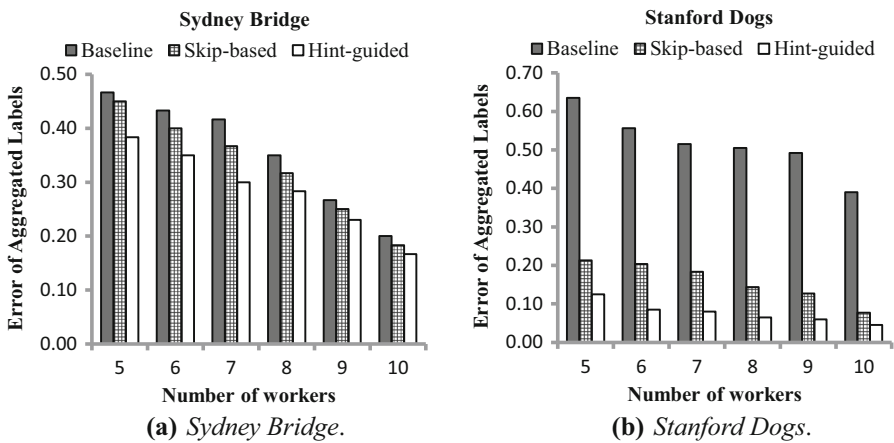


Fig. 4 Evaluation of the label quality. Results on *Speech Clips* are not reported, as text answers cannot be majority voted

Table 3 Evaluation of the worker quality detection of the hint-guided approach. Error rate (in %) is provided for aggregating original and rescaled crowdsourced labels

Number of workers		5	10
<i>Sydney Bridge</i>	Origin	38.33%	16.67%
	Rescale	30.00%	11.67%
<i>Stanford Dogs</i>	Origin	12.50%	4.50%
	Rescale	12.00%	4.00%

The best results are highlighted in bold

and Zhou 2015), and the number of multiple workers depends on varying situations. For each of combinations between tasks and $n_workers$, we perform the following actions 200 times repeatedly. In each time, for all questions, we randomly select $n_workers$ workers and perform the majority voting on their responses to yield the aggregated labels. The plotted error of aggregated labels is averaged across 200 results. We observe that the hint-guided approach consistently outperforms the baseline and the skip-based approaches, and the performance gap between the baseline and the hint-guided approaches is extremely obvious on *Stanford Dogs*.

5.2.3 Worker quality detection

Table 3 denotes the error of aggregating original and rescaled crowdsourced labels. For rescaled crowdsourced labels, labels from estimated high-quality workers are adaptively given more weights, and vice versa. From Table 3, we can see the error of aggregating rescaled labels is lower than the error of aggregating original labels. It demonstrates that our hint-guided approach can detect the high-quality workers effectively. Then, the error decreases significantly on *Sydney Bridge*, since the size of *Sydney Bridge* is relatively small (30 questions) compared to *Stanford Dogs* (100 questions). We believe that, the informative hints for *Stanford Dogs* may guide the low-quality workers to make more accurate decisions. Then, the performance gap between high-quality and low-quality workers is insignificant. Therefore, the effect of label rescaling is marginal on this dataset.

5.2.4 Spammer prevention

The baseline and hint-guided approaches are represented as Single(+) and Hybrid(\times), respectively. We provide one extra interaction: the single-stage setting with the “ \times ” mechanism (Single(\times)), and all parameters are consistent. Figure 5a explores how our approach prevents spammers. It plots the average payment to each worker under three approaches. We have one observation: the payments of Single(\times) and Hybrid(\times) are lower than that of Single(+), since an answer in G questions is incorrect, and thus the reward of the “ \times ” mechanism becomes zero. Since spammers answer each question randomly, the “ \times ” mechanism used by our approach makes the smallest payment to them. Thus, our approach prevents spammers.

5.2.5 Money cost

Figure 5b plots the average payment to each worker under the three approaches. The higher the payment is, the worse the economy of the approach. The payment is calculated as the average of the payments across 200 random selections of G questions. This process mitigates the distortion of results caused by the randomness in the choice of G questions. We can

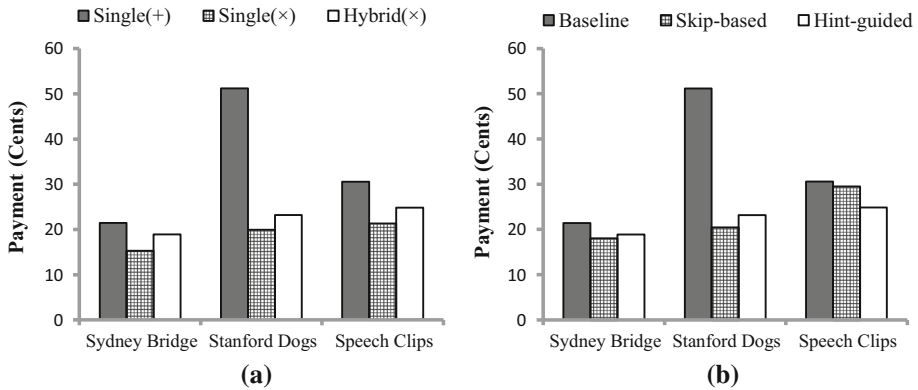


Fig. 5 Evaluation of the spammer prevention. Average payment to each worker on all three tasks are provided. Evaluation of the money cost

see that, the payments of the skip-based and hint-guided approaches are comparable but less than the payment of the baseline approach, especially in the *Stanford Dogs* task, since both the skip-based and hint-guided approaches use the multiplicative mechanism but the baseline approach use the additive mechanism. Thus, from the perspective of saving money, we should not employ the baseline approach. Note that, on the *Sydney Bridge* and *Stanford Dogs* tasks, although the payment in the skip-based approach is slightly lower than that in the hint-guided approach, the number of high-quality labels from the hint-guided approach is obviously higher than that from the skip-based approach (Fig. 4).

6 Conclusions

To improve the label quality, we proposed a hint-guided approach that encourages workers to use hints when they answer unsure questions. Our approach consists of the hybrid-stage setting and the hint-guided payment mechanism. We proved the incentive compatibility and uniqueness of our mechanism. Besides, our approach can detect the high-quality workers for more accurate result aggregation. Comprehensive experiments conducted on Amazon MTurk revealed the effectiveness of our approach and validated the simple and practical deployment of our approach. These merits are critical for the success of many machine learning applications in practice.

As for future works, first, the hint-guided approach is designed under the worker's independence. However, it would become more interesting to extend hint-guided approach under the worker's dependence, where the reward of a worker depends on the answers of the other ones. Second, we hope to extend the hybrid-stage setting from binary choice to multiple choice with the corresponding theoretical results. Third, we consider to provide hints from different levels for all questions. Specifically, we will provide the hints from coarse to fine, which corresponds the different expected payments. Finally, some workers may still be very confused even with hints, we may mix up the unsure option in the hint stage to further improve the label quality further.

Acknowledgements Ivor W. Tsang was partially supported by ARC FT130100746, LP150100671 and DP180100106. Masashi Sugiyama was supported by the International Research Center for Neurointelligence (WPI-IRCn) at The University of Tokyo Institutes for Advanced Study. Xiaokui Xiao was partially supported

by MOE, Singapore under grant MOE2015-T2-2-069, and by NUS, Singapore under an SUG. Bo Han and Ivor W. Tsang would like to thank Yao-Xiang Ding, Dengyong Zhou and Jun Zhu for helpful comments and discussions.

Appendix A: Proofs

A.1 Proposition 1

Proof *The justification of the first pricing constraint* under the hybrid-stage setting, the reasonable payment mechanism should consider two facts: (1) the payment for correct answer should be much higher than the payment for incorrect answer. Namely, $d_+ > d_-$, $h_+ > h_-$. (2) If the answer is correct, the payment to the worker who answers directly should be higher than the payment to the worker who answers using hints. Namely, $d_+ > h_+$. Why this condition penalizes workers who use the hints? The reason is that: $d_+ > h_+$ ensures that high-quality workers will answer most of questions directly for higher rewards. Under the proposed setting, whether using hints can be taken as a criterion to detect the high-quality workers. Thus, $d_+ > h_+$ assists the hybrid-stage setting to detect the high-quality workers.

The justification of the second pricing constraint for each question, $d_+ - d_-$ is the income gap for a worker who answers directly, and $h_+ - h_-$ is the income gap for a worker who answers with hints. In order to encourage the worker to use the hints properly, we consider bridge the per unit of income gap $\frac{d_+ - d_-}{1 - 2\epsilon}$ in the main stage and $\frac{h_+ - h_-}{2\epsilon}$ in the hint stage together. Namely, we impose the condition $\frac{d_+ - d_-}{1 - 2\epsilon} \geq \frac{h_+ - h_-}{2\epsilon}$ into the incentive-compatible payment mechanism, which encourages his/her to directly answer questions that he/she feels confident about. In other words, the worker should not abuse the hints.

The justification of the third pricing constraint the third condition is the complementary condition for the second one. It should incentivize the worker to leverage the hints before answering questions that he/she feels unsure of. When $P_A > \frac{1}{2} - \epsilon$ while $P_{A|H} > T$, we prefer to choose ‘‘A’’. If the worker selects ‘‘A’’ by his/her own belief, then his/her expected payment is

$$Payment(A) = \left(\frac{1}{2} - \epsilon\right) (d_+ P_A + d_- P_B) + 2\epsilon(h_+ P_{A|H} + h_- P_{B|H}).$$

When $P_A < \frac{1}{2} + \epsilon$ while $P_{B|H} > T$, we prefer to choose ‘‘B’’. If the worker selects ‘‘B’’ by his/her own belief, then his/her expected payment is

$$Payment(B) = \left(\frac{1}{2} - \epsilon\right) (d_+ P_B + d_- P_A) + 2\epsilon(h_+ P_{B|H} + h_- P_{A|H}).$$

If a payment mechanism is incentive-compatible, the worker is incentivized to choose the answer by his/her own belief while the according payment is strictly maximized. In this case, when $P_A > \frac{1}{2} - \epsilon$ while $P_{A|H} > T$, $Payment(A) > Payment(B)$. When $P_A < \frac{1}{2} + \epsilon$ while $P_{B|H} > T$, $Payment(A) < Payment(B)$. Let us infer the case of $Payment(A) > Payment(B)$, the other case gives the same result by symmetry.

$$\begin{aligned} & \left(\frac{1}{2} - \epsilon\right) (d_+ P_A + d_- P_B) + 2\epsilon(h_+ P_{A|H} + h_- P_{B|H}) \\ & > \left(\frac{1}{2} - \epsilon\right) (d_+ P_B + d_- P_A) + 2\epsilon(h_+ P_{B|H} + h_- P_{A|H}). \end{aligned}$$

Due to the facts that $P_A = 1 - P_B$ and $P_{A|H} = 1 - P_{B|H}$,

$$\begin{aligned} & \left(\frac{1}{2} - \epsilon\right) (d_+(2P_A - 1) + d_-(1 - 2P_A)) + 2\epsilon(h_+(2P_{A|H} - 1) + h_-(1 - 2P_{A|H})) > 0. \\ & \left(\frac{1}{2} - \epsilon\right) (2P_A - 1)(d_+ - d_-) + 2\epsilon(2P_{A|H} - 1)(h_+ - h_-) > 0. \end{aligned} \tag{2}$$

Due to $P_A > \frac{1}{2} - \epsilon$ and $P_{A|H} > T$, then we have $2P_A - 1 > -2\epsilon$ and $2P_{A|H} - 1 > 2T - 1$. According to Eq. (2), we have,

$$\left(\frac{1}{2} - \epsilon\right) (2P_A - 1)(d_+ - d_-) > -2\epsilon(2P_{A|H} - 1)(h_+ - h_-). \tag{3}$$

For Eq. (3), the term in the left hand side should always be larger than the term in the right hand side for the same ϵ . Hence, the ‘‘Infimum’’ value of the left hand side should be always larger than the ‘‘Supremum’’ value of the right hand side.

$$\inf_{P_A} \left\{ \left(\frac{1}{2} - \epsilon\right) (2P_A - 1)(d_+ - d_-) \right\} \geq \sup_{P_{A|H}} \{ -2\epsilon(2P_{A|H} - 1)(h_+ - h_-) \}.$$

Therefore, we have

$$\left(\frac{1}{2} - \epsilon\right) (-2\epsilon)(d_+ - d_-) \geq -2\epsilon(2T - 1)(h_+ - h_-). \tag{4}$$

Due to $-2\epsilon < 0$, therefore, we eliminate the same negative parameter in both ends of the Eq. (4), then we have,

$$\left(\frac{1}{2} - \epsilon\right) (d_+ - d_-) \leq (2T - 1)(h_+ - h_-).$$

Finally, due to $\epsilon \in [0, \frac{1}{2})$, we have the condition as follows,

$$(d_+ - d_-) \leq \frac{2T - 1}{1/2 - \epsilon} (h_+ - h_-).$$

□

A.2 Proposition 2

Proof When we consider the last two pricing constraints in Proposition 1 under the ‘‘mild no-free-lunch Axiom’’ in Definition 2, we can see that $N = G = 1$ and $d_- = h_- = 0$. Therefore, we have,

$$\frac{1 - 2\epsilon}{2\epsilon} \leq \frac{d_+}{h_+} \leq \frac{2T - 1}{\frac{1}{2} - \epsilon}.$$

as $\epsilon \in [0, \frac{1}{2})$, then we have

$$\begin{aligned} & 2 \left(\frac{1}{2} - \epsilon\right) \leq (2T - 1)2\epsilon. \\ & \epsilon^2 - 2T\epsilon + \frac{1}{4} \leq 0. \\ & \left(\epsilon - \frac{2T + \sqrt{4T^2 - 1}}{2}\right) \left(\epsilon - \frac{2T - \sqrt{4T^2 - 1}}{2}\right) \leq 0. \end{aligned}$$

Thus the feasible region for ϵ is

$$\left[0, \frac{1}{2}\right) \cap \left[T - \sqrt{T^2 - \frac{1}{4}}, T + \sqrt{T^2 - \frac{1}{4}}\right].$$

In addition, due to $T \in (\frac{1}{2}, 1)$ and $T + \sqrt{T^2 - \frac{1}{4}}$ increases monotonically with T , then

$$\min_T \left(T + \sqrt{T^2 - \frac{1}{4}}\right) > \left(T + \sqrt{T^2 - \frac{1}{4}}\right) |_{T=\frac{1}{2}} = \frac{1}{2}.$$

Similarly, $T - \sqrt{T^2 - \frac{1}{4}} = \frac{\frac{1}{4}}{T + \sqrt{T^2 - \frac{1}{4}}}$ decreases monotonically with T , then

$$\max_T \left(T - \sqrt{T^2 - \frac{1}{4}}\right) < \left(T - \sqrt{T^2 - \frac{1}{4}}\right) |_{T=\frac{1}{2}} = \frac{1}{2}.$$

and

$$\min_T \left(T - \sqrt{T^2 - \frac{1}{4}}\right) > \left(T - \sqrt{T^2 - \frac{1}{4}}\right) |_{T=1} = 1 - \frac{\sqrt{3}}{2} > 0.$$

To sum up, ϵ satisfies

$$\epsilon \in \left[T - \sqrt{T^2 - \frac{1}{4}}, \frac{1}{2}\right).$$

Therefore, for a fixed $T \in (\frac{1}{2}, 1)$, the minimum ϵ for a incentive-compatible payment mechanism should be $\epsilon_{\min} = T - \sqrt{T^2 - \frac{1}{4}}$. For the case of $1 \leq G \leq N$, we have the same result because, for a random question to be a gold standard question, we get the minimum value ϵ_{\min} , which is the lower bound of ϵ . Due to previous assumptions that each question is independent and all questions share the same ϵ in system design, ϵ_{\min} will be the most suitable value to cover all cases. This completes the proof. \square

A.3 Corollary 1

Proof Under Assumption 2, we have $T \in (5/8, 1)$. Therefore, we can build up the above inequality,

$$(8T - 5)(T - 2/4) > 0.$$

Namely,

$$8T^2 - 9T + 10/4 > 0.$$

Then we have,

$$(T - 1/2)(T + 1/2) < (3T - 3/2)^2,$$

which equals to

$$3/2 + \sqrt{T^2 - 1/4} < 3T.$$

Therefore, we have

$$1/2 - T + \sqrt{T^2 - 1/4} < 2T - 1.$$

According to Proposition 2 ($\epsilon_{\min} = T - \sqrt{T^2 - 1/4}$), we have $(1/2 - \epsilon_{\min}) < (2T - 1)$. \square

A.4 Theorem 1

Proof To prove “if and only if”, the standard way is to prove the existence first, and then prove the uniqueness.

Existence To consider the case of $N = G = 1$, when $\epsilon = \epsilon_{\min}$, the proposed payment mechanism meets three pricing constraints, therefore, the proposed payment mechanism is incentive-compatible. For the case of $1 \leq G \leq N$, $a_1, \dots, a_G \in \{\mathbb{D}_+, \mathbb{D}_-, \mathbb{H}_+, \mathbb{H}_-\}$ are the evaluations of the answers to the G gold standard questions. The final payment is $\beta \prod_{i=1}^G f(a_i) + \mu_{\min}$. Due to the assumption that each $a_i, i \in \{1, \dots, G\}$ is independent, the overall expected payment $Payment(a_1, \dots, a_G)$ equals to the product (scaled by β ; shifted by μ_{\min}) of the expected values $f(a_i)$ for each $a_i, i \in \{1, \dots, G\}$. As $f(a_i)$ is maximized when the worker answers the question by his/her own belief, the overall expected payment is then maximized when all workers answer the questions by their own beliefs. This proves the existence.

Uniqueness To consider the case of $N = G = 1$, according to the “mild no-free-lunch” axiom, both d_- and h_- are equal to 0. Furthermore, according to three pricing constraints, when $\epsilon = \epsilon_{\min}$, we get an equality case of 1, namely, $\frac{1-2\epsilon}{2\epsilon} = \frac{d_+}{h_+} = \frac{2T-1}{\frac{1}{2}-\epsilon}$, hence, the relation between d_+ and h_+ is fixed. The derived mechanism is identical to the hint-guided payment mechanism. Therefore, we further consider whether the derived mechanism is identical to the hint-guided payment mechanism for the general case of $1 \leq G \leq N$.

The base case of the induction hypothesis is that the derived mechanism is identical to the hint-guided payment mechanism whenever $\mathbf{y} \in \{\mathbb{H}_+, \mathbb{H}_-\}^G \setminus \{\mathbb{H}_+\}^G$. For G gold standard questions, we assume that the worker answers at least $G - r - 1$ questions with hints, namely, $\sum_{i=1}^G \mathbf{1}\{y_i \in \{\mathbb{H}_+, \mathbb{H}_-\}\} \geq G - r - 1$.⁴ We suppose that the induction hypothesis is true whenever $\mathbf{y} \in \{\mathbb{D}_+, \mathbb{D}_-, \mathbb{H}_+, \mathbb{H}_-\}^G \setminus \{\mathbb{D}_+, \mathbb{H}_+\}^G$. Now we prove the induction hypothesis keeps true whenever $\mathbf{y} \in \{\mathbb{D}_+, \mathbb{D}_-, \mathbb{H}_+, \mathbb{H}_-\}^G \setminus \{\mathbb{D}_+, \mathbb{H}_+\}^G$ and $\sum_{i=1}^G \mathbf{1}\{y_i \in \{\mathbb{H}_+, \mathbb{H}_-\}\} = G - r$.

Let y_i denote the state evaluation of his/her answer to the question $i \in \{1, \dots, G\}$. Assume that $y_1, \dots, y_{r-1} \in \{\mathbb{D}_+, \mathbb{D}_-\}$ and $y_{r+1}, \dots, y_G \in \{\mathbb{H}_+, \mathbb{H}_-\}$. For N questions, suppose for $i \in \{1, \dots, r-1\}$, we have $P_A > \frac{1}{2} + \epsilon$; for $i \in \{r+1, \dots, N\}$, we have $\frac{1}{2} - \epsilon < P_A < \frac{1}{2} + \epsilon$ while $P_{A|H} > T$. Thus, he/she will select “A” for all questions $\{1, \dots, N\} \setminus \{r\}$. Moreover, the worker holds a belief that questions $1, \dots, r-1$ can be selected directly; while questions $r+1, \dots, G$ will be answered with hints.

Assume that $q : \{\mathbb{D}_+, \mathbb{D}_-, \mathbb{H}_+, \mathbb{H}_-\} \rightarrow [0, \mu_{\max}]$ be a function defined as follows. $q(y_r)$ is an expected payment conditioned on the r th question, which is composed of a convex combination of two parts. The first part is the payment that r th question is in the gold standard question; the second part is the payment that r th question is not in the gold standard question. Hence, $q(y_r) = \phi q^*(y_r) + (1 - \phi)c$, where $\phi \in (0, 1), c \geq 0$. q^* denote the first part payment, which is dependent on $q(y_r)$.

Assume that the function q^* is a convex combination of the payment function f evaluated at various points. Due to “mild no-free-lunch” axiom, we have $q^*(y_r) = 0$ when $y_r \in \{\mathbb{H}_-\}$. Let $P_B = 1 - P_A$; $P_{B|H} = 1 - P_{A|H}$ be the worker’s confidence for the question r , if the payment mechanism incentivizes the worker to select the answer for question r properly, then the result should be:

⁴ $\mathbf{1}\{x\}$ is an indicator function, and $\mathbf{1}\{x = true\} = 1$ while $\mathbf{1}\{x = false\} = 0$.

If $P_A > \frac{1}{2} - \epsilon$ and $P_{A|H} > T$, then the answer to question r is A, hence, the expected payment by A should be larger than the expected payment by B.

$$\begin{aligned} & \left(\frac{1}{2} - \epsilon\right) (P_A q^*(\mathbb{D}_+) + P_B q^*(\mathbb{D}_-)) + 2\epsilon(P_{A|H} q^*(\mathbb{H}_+) + P_{B|H} q^*(\mathbb{H}_-)) \\ & > \left(\frac{1}{2} - \epsilon\right) (P_B q^*(\mathbb{D}_+) + P_A q^*(\mathbb{D}_-)) + 2\epsilon(P_{B|H} q^*(\mathbb{H}_+) + P_{A|H} q^*(\mathbb{H}_-)). \end{aligned}$$

In turn, if $P_A < \frac{1}{2} + \epsilon$ and $P_{B|H} > T$, then the answer to question r is B, hence, the expected payment by B should be larger than the expected payment by A.

$$\begin{aligned} & \left(\frac{1}{2} - \epsilon\right) (P_A q^*(\mathbb{D}_+) + P_B q^*(\mathbb{D}_-)) + 2\epsilon(P_{A|H} q^*(\mathbb{H}_+) + P_{B|H} q^*(\mathbb{H}_-)) \\ & < \left(\frac{1}{2} - \epsilon\right) (P_B q^*(\mathbb{D}_+) + P_A q^*(\mathbb{D}_-)) + 2\epsilon(P_{B|H} q^*(\mathbb{H}_+) + P_{A|H} q^*(\mathbb{H}_-)). \end{aligned}$$

Let $P_A = \frac{1}{2} + \epsilon$ and $P_{B|H} = T$, then we have,

$$\begin{aligned} & \left(\frac{1}{2} - \epsilon\right) \left(\frac{1}{2} + \epsilon\right) q^*(\mathbb{D}_+) + \left(\frac{1}{2} - \epsilon\right)^2 q^*(\mathbb{D}_-) + 2\epsilon(1 - T)q^*(\mathbb{H}_+) + 2\epsilon T q^*(\mathbb{H}_-) \\ & \leq \left(\frac{1}{2} - \epsilon\right)^2 q^*(\mathbb{D}_+) + \left(\frac{1}{2} - \epsilon\right) \left(\frac{1}{2} + \epsilon\right) q^*(\mathbb{D}_-) + 2\epsilon T q^*(\mathbb{H}_+) + 2\epsilon(1 - T)q^*(\mathbb{H}_-). \end{aligned}$$

Due to $q^*(\mathbb{H}_-) = 0$, we have,

$$\begin{aligned} & \left(\frac{1}{2} - \epsilon\right) \left(\frac{1}{2} + \epsilon\right) q^*(\mathbb{D}_+) + \left(\frac{1}{2} - \epsilon\right)^2 q^*(\mathbb{D}_-) + 2\epsilon(1 - T)q^*(\mathbb{H}_+) \\ & \leq \left(\frac{1}{2} - \epsilon\right)^2 q^*(\mathbb{D}_+) + \left(\frac{1}{2} - \epsilon\right) \left(\frac{1}{2} + \epsilon\right) q^*(\mathbb{D}_-) + 2\epsilon T q^*(\mathbb{H}_+). \end{aligned}$$

After simplifying, we have,

$$\begin{aligned} & \left(\frac{1}{2} - \epsilon\right) 2\epsilon q^*(\mathbb{D}_+) - \left(\frac{1}{2} - \epsilon\right) 2\epsilon q^*(\mathbb{D}_-) \leq 2\epsilon(2T - 1)q^*(\mathbb{H}_+). \\ & q^*(\mathbb{D}_+) - q^*(\mathbb{D}_-) \leq \frac{2T - 1}{\frac{1}{2} - \epsilon} q^*(\mathbb{H}_+). \end{aligned}$$

When designing a payment mechanism, we assume that $q^*(\mathbb{D}_+)$ is fixed, and $q^*(\mathbb{H}_+)$ can be derived by the relation with $q^*(\mathbb{D}_+)$. We hope to set $q^*(\mathbb{H}_+)$ to the boundary value. Specifically, as we want to penalize the use of hints, we set $q^*(\mathbb{H}_+)$ as small as possible while meeting the inequality. Therefore, we set $q^*(\mathbb{D}_+)$ to the lower bound. Namely, $q^*(\mathbb{H}_+) = \frac{\frac{1}{2} - \epsilon}{2T - 1} (q^*(\mathbb{D}_+) - q^*(\mathbb{D}_-))$. Due to ‘‘mild no-free-lunch’’ axiom, $q^*(\mathbb{H}_+) = \frac{\frac{1}{2} - \epsilon}{2T - 1} q^*(\mathbb{D}_+)$.

Since q^* is a convex combination of the payment function f evaluated at various points. Therefore, we have,

$$(2T - 1)f(\mathbb{H}_+, y_2, \dots, y_G) = \left(\frac{1}{2} - \epsilon\right) f(\mathbb{D}_+, y_2, \dots, y_G),$$

where the augments above hold for any permutation of the G gold standard questions.

$$(2T - 1)f(y_1, y_2, \dots, \mathbb{H}_+^i, \dots, y_G) = \left(\frac{1}{2} - \epsilon\right) f(y_1, y_2, \dots, \mathbb{D}_+^i, \dots, y_G),$$

where the notation \mathbb{H}_+^i denotes the state evaluation of y_i is \mathbb{H}_+ . The same is true for \mathbb{D}_+^i . Then, according to the recursive induction, we have

$$f(\overbrace{\mathbb{D}_+, \dots, \mathbb{D}_+}^G) = \left(\frac{2T-1}{\frac{1}{2}-\epsilon}\right)^G f(\overbrace{\mathbb{H}_+, \dots, \mathbb{H}_+}^G). \tag{5}$$

Based on Eq. (5) and using the fact that all arguments apply to any permutation of the G gold standard questions, we can see that f should be identical to the hint-guided payment mechanism. This proves the uniqueness. \square

A.5 Theorem 2

Proof Under the hybrid-stage setting, the incentive-compatible mechanism encourages the worker to use hints when he/she is unsure of the questions. Since Definition 1 is contradictory to Definition 3, the ‘‘Harsh No-Free-Lunch Axiom’’ is too strong for the existence of any incentive-compatible payment mechanism. \square

Appendix B: Baseline approach in AMT

Baseline approach consists of single-stage setting and additive payment mechanism. The single-stage setting is a special case of the hybrid-stage setting, where ϵ is set to 0. For every question $i \in \{1, \dots, N\}$, the worker should be incentivized to choose answers matching his/her own belief.

- The single stage: he/she should be incentivized to select the option that he/she feels more confident, namely,

$$\text{select } \begin{cases} \text{‘‘A’’ } P_{A,i} \in [\frac{1}{2}, 1) \\ \text{‘‘B’’ } P_{A,i} \in (0, \frac{1}{2}). \end{cases}$$

Under the single-stage setting, the current state space is $\{\mathbb{D}_+, \mathbb{D}_-\}$. ‘‘ \mathbb{D}_+ ’’ and ‘‘ \mathbb{D}_- ’’ denote correct answer and incorrect answer, respectively. The state evaluations of his/her responses to G questions are denoted by $a_1, \dots, a_G \in \{\mathbb{D}_+, \mathbb{D}_-\}$. Assume that any values d_- and d_+ such that $0 \leq d_- \leq d_+$, a function $f_a : \{\mathbb{D}_+, \mathbb{D}_-\} \rightarrow \mathbb{R}_+$, where $f_a(\mathbb{D}_+) = d_+$, $f_a(\mathbb{D}_-) = d_-$. The additive payment mechanism f is:

$$f([a_1, \dots, a_G]) = \sum_{i=1}^G f_a(a_i). \tag{6}$$

Remark 3 Additive payment mechanism, i.e., Eq. (6) is not only additive but also incentive-compatible.⁵ However, due to additive form, if half attempts in G questions are correct, the workers still acquire $\frac{(d_-+d_+)G}{2}$ payments. This payment mechanism may not effectively prevent spammers who select options randomly. Here, we prove the additive payment mechanism is incentive-compatible as follows.

In the single-stage setting, for each question, the expected payment is ‘‘ $d_+P(\mathbb{D}_+) + d_-P(\mathbb{D}_-)$ ’’, where $P(\mathbb{D}_+)$ and $P(\mathbb{D}_-)$ are the probability of correct answer and incorrect

⁵ A payment mechanism is incentive-compatible if it incentivizes the worker to choose the answers to all questions by his/her own belief and his/her expected payment is strictly maximized.

answer, respectively. For binary-value questions, $P_B = 1 - P_A$, if the worker chooses “A”, which means that $P_A > \frac{1}{2} > P_B$ in his/her belief, then the expected payment for this question is equal to “Payment(A)”. To verify the incentive compatibility, we compare “Payment(A)” with “Payment(B)”. Due to “ $d_+ \geq d_-$ ”, we have $\text{Payment(A)} = d_+ P_A + d_- P_B > d_+ P_B + d_- P_A = \text{Payment(B)}$. Therefore, if the worker chooses “A” by his/her belief, the “Payment(A)” is larger than “Payment(B)”, and vice versa.

Appendix C: Payment parameters in Sect. 5.1

Here, we provide the details of parameter setting for different payment mechanisms. The payment is often composed of two parts: a fixed (minimum) payment and a reward (bonus) payment. The fixed payment is paid for each worker who undertakes all tasks, which avoids all multiplicative payment mechanisms too harsh for workers. The reward payment is based on his/her responses to G gold standard questions. For each task, the fixed payment and reward payment are denoted as FP and RP , respectively.

For additive mechanism, k_1 denotes the unit reward for each correct answer. For skip-based mechanism, RP starts from k_2 , increasing by P_s for each correct answer. Note that the RP remains the same for skip option, but becomes zero for any incorrect answer. For hint-guided mechanism, RP starts from k_3 , increasing by P_d and P_h for each correct answer in the main stage and the hint stage, respectively. However, RP will become zero for any incorrect answer. All payment parameters of this paper are in Table 4. Here, we will explain how we set these parameters as follows.

According to the suggestion on Amazon MTurk, the reward per question (denotes as r_a) should be set according to the minimum wage, e.g., a 30s question that pays 5 cents is a 6 dollars hourly wage. Therefore, r_a are set to 1 cent for each image annotation question and 4.5 cents for each speech recognition question. Moreover, the fixed payment FP are set to 5 cents, 7 cents, 5 cents for Sydney Bridge, Stanford Dogs, and Speech Clips respectively. For different payment mechanisms, the other parameters can be decided as follows:

- Additive payment mechanism: the payment for a strong worker who answers all the G gold questions correctly should be the same as the total payment, thus k_1 satisfies the constrain $FP + G * k_1 = N * r_a$.
- Skip-based payment mechanism (multiplicative): To incentive the worker to provide high-quality labels, the reward per question with the multiplicative payment mechanism (denotes as r_m) should be higher than r_a . Therefore, in our experiments, r_m are set to 1.2 cents for each image annotation question and 5.4 cents for each speech recognition

Table 4 The payment parameters. The total payment is composed of FP and RP

Data set	FP	G	Baseline	Skip-based		Hint-guided		
			k_1	k_2	P_s (%)	k_3	P_d (%)	P_h (%)
<i>Sydney Bridge</i>	5	3	8.5	9.15	50	9.15	50	30
<i>Stanford Dogs</i>	7	10	9	1.02	60	1.02	60	37
<i>Speech Clips</i>	5	2	20	12.25	100	12.25	100	62

RP is based on worker’s responses to G questions. $P_h = \frac{\frac{1}{2} - \epsilon}{2T - 1} P_d$ according to Algorithm 1, where we set $T = 0.75$ and $\epsilon = 0.191$ due to the Proposition 2

question. Furthermore, the amount payment for a strong worker who answers all the G gold questions correctly under the Skip-based payment mechanism should be the same as the total payment, thus k_2 satisfies the constrain $FP + k_2 * (1 + P_s)^G = N * p_m$, where P_s are set to 50%, 60% and 100% for Sydney Bridge, Stanford Dogs, and Speech Clips respectively.

- Hints-guided payment mechanism (multiplicative): For a strong worker who answers all the G gold questions correctly in the main stage, the reward should be the same as that under the Skip-based payment mechanism, thus parameters k_3 and P_d satisfy $k_3 = k_2$ and $P_d = P_s$ for each task specifically. Moreover, for the reward in the hint stage, we decide P_h according to $P_h = \frac{\frac{1}{2} - \epsilon}{2T - 1} P_d$ in Hint-guided payment mechanism, where we set $T = 0.75$ and $\epsilon = 0.191$ according to the equation $\epsilon_{\min} = T - \sqrt{T^2 - \frac{1}{4}}$.

References

- Bi, W., Wang, L., Kwok, J., & Tu, Z. (2014). Learning to predict from crowdsourced data. In *Conference on uncertainty in artificial intelligence* (pp. 82–91).
- Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., et al. (2016). End to end learning for self-driving cars. arXiv preprint [arXiv:1604.07316](https://arxiv.org/abs/1604.07316).
- Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1), 3–5.
- Chen, X., Lin, Q., & Zhou, D. (2013). Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *International conference on machine learning* (pp. 64–72).
- Chen, Y., Chong, S., Kash, I., Moran, T., & Vadhan, S. (2016). Truthful mechanisms for agents that value privacy. *ACM Transactions on Economics and Computation*, 4(3), 13.
- Difallah, D., Demartini, G., & Cudré-Mauroux, P. (2012). Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch* (pp. 26–30).
- Ding, Y. X., & Zhou, Z. H. (2017). Crowdsourcing with unsure option. *Machine Learning*, 107, 749–766.
- Fan, J., Li, G., Ooi, B., Tan, K., & Feng, J. (2015). iCrowd: An adaptive crowdsourcing framework. In *ACM special interest group on management of data* (pp. 1015–1030).
- Goel, G., Nikzad, A., & Singla, A. (2014). Mechanism design for crowdsourcing markets with heterogeneous tasks. In *AAAI conference on human computation and crowdsourcing*.
- Han, B., Pan, Y., & Tsang, I. (2017). Robust Plackett–Luce model for k-ary crowdsourced preferences. *Machine Learning*, 107, 675–702.
- Han, B., Tsang, I., & Chen, L. (2016). On the convergence of a family of robust losses for stochastic gradient descent. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 665–680). Springer.
- Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6), 82–97.
- Ho, C., Slivkins, A., Suri, S., & Vaughan, J. (2015). Incentivizing high quality crowdwork. In *World Wide Web* (pp. 419–429).
- Hu, H., Zheng, Y., Bao, Z., Li, G., Feng, J., & Cheng, R. (2016). Crowdsourced POI labelling: Location-aware result inference and task assignment. In *International conference on data engineering* (pp. 61–72).
- Ipeirotis, P., Provost, F., & Wang, J. (2010). Quality management on Amazon Mechanical Turk. In *ACM SIGKDD conference on knowledge discovery and data mining workshop* (pp. 64–67).
- Joglekar, M., Garcia-Molina, H., & Parameswaran, A. (2013). Evaluating the crowd with confidence. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 686–694).
- Kajino, H., Tsuboi, Y., & Kashima, H. (2012). A convex formulation for learning from crowds. *Transactions of the Japanese Society for Artificial Intelligence*, 27(3), 133–142.
- Karger, D., Oh, S., & Shah, D. (2011). Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems* (pp. 1953–1961).
- Koedinger, K., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19(3), 239–264.

- Lambert, N., Langford, J., Vaughan, J., Chen, Y., Reeves, D., Shoham, Y., et al. (2015). An axiomatic characterization of wagering mechanisms. *Journal of Economic Theory*, 156, 389–416.
- Li, G., Chai, C., Fan, J., Weng, X., Li, J., Zheng, Y., et al. (2017a). CDB: Optimizing queries with crowd-based selections and joins. In *ACM SIGMOD international conference on management of data* (pp. 1463–1478).
- Li, G., Wang, J., Zheng, Y., & Franklin, M. J. (2016). Crowdsourced data management: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 28(9), 2296–2319.
- Li, G., Zheng, Y., Fan, J., Wang, J., & Cheng, R. (2017b). Crowdsourced data management: Overview and challenges. In *ACM SIGMOD international conference on management of data* (pp. 1711–1716).
- Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior Research Methods*, 47(2), 519–528.
- Liu, Q., Peng, J., & Ihler, A. (2012). Variational inference for crowdsourcing. In *Advances in neural information processing systems* (pp. 692–700).
- Natarajan, N., Dhillon, I., Ravikumar, P., & Tewari, A. (2013). Learning with noisy labels. In *Advances in neural information processing systems* (pp. 1196–1204).
- Nisan, N., Roughgarden, T., Tardos, E., & Vazirani, V. (2007). *Algorithmic game theory* (Vol. 1). Cambridge: Cambridge University Press.
- Patrini, G., Rozza, A., Menon, A., Nock, R., & Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *IEEE conference on computer vision and pattern recognition*.
- Pennock, D., Syrgkanis, V., & Vaughan, J. (2016). Bounded rationality in wagering mechanisms. In *Conference on uncertainty in artificial intelligence*.
- Raykar, V., & Yu, S. (2012). Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research*, 13, 491–518.
- Raykar, V., Yu, S., Zhao, L., Valadez, G., Florin, C., Bogoni, L., et al. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11, 1297–1322.
- Rodrigues, F., Pereira, F., & Ribeiro, B. (2014). Sequence labeling with multiple annotators. *Machine Learning*, 95(2), 165–181.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.
- Shah, N., & Zhou, D. (2015). Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. In *Advances in neural information processing systems* (pp. 1–9).
- Shah, N., & Zhou, D. (2016). No Oops, You Wont Do it again: Mechanisms for self-correction in crowdsourcing. In *International conference on machine learning*.
- Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484–489.
- Singla, A., & Krause, A. (2013). Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *World Wide Web* (pp. 1167–1178).
- Singla, A., Santoni, M., Bartók, G., Mukerjji, P., Meenen, M., & Krause, A. (2015). Incentivizing users for balancing bike sharing systems. In *Association for the advancement of artificial intelligence* (pp. 723–729).
- Smith, J. (2007). *Qualitative psychology: A practical guide to research methods*. Thousand Oaks: Sage.
- Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., et al. (2015). A neural network approach to context-sensitive generation of conversational responses. In *Conference of the association for computational linguistics* (pp. 196–205).
- Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., & Fergus, R. (2015). Training convolutional networks with noisy labels. In *International conference on learning representations workshop*.
- Tian, T., & Zhu, J. (2015). Max-margin majority voting for learning from crowds. In *Advances in neural information processing systems* (pp. 1621–1629).
- Vuurens, J., Vries, A., & Eickhoff, C. (2011). How much spam can you take? An analysis of crowdsourcing results to increase accuracy. In *ACM special interest group on information retrieval workshop* (pp. 21–26).
- Wais, P., Lingamneni, S., Cook, D., Fennell, J., Goldenberg, B., Lubarov, D., Marin, D., & Simons, H. (2010). Towards building a high-quality workforce with Mechanical Turk. In *Advances in neural information processing systems workshop*.
- Wang, L., & Zhou, ZH. (2016). Cost-saving effect of crowdsourcing learning. In *International joint conference on artificial intelligence* (pp. 2111–2117).
- Wang, W., Guo, X. Y., Li, S. Y., Jiang, Y., & Zhou, ZH. (2017). Obtaining high-quality label by distinguishing between easy and hard items in crowdsourcing. In *International joint conference on artificial intelligence* (pp. 2964–2970).
- Yan, Y., Rosales, R., Fung, G., & Dy, J. (2011). Active learning from crowds. *International Conference on Machine Learning*, 11, 1161–1168.

- Yan, Y., Rosales, R., Fung, G., Schmidt, M., Hermosillo, G., Moy, L., & Dy, J. (2010). Modeling annotator expertise: Learning when everybody knows a bit of something. In *International conference on artificial intelligence and statistics* (pp. 932–939).
- Yan, Y., Rosales, R., Fung, G., Subramanian, R., & Dy, J. (2014). Learning from multiple annotators with varying expertise. *Machine Learning*, 95(3), 291–327.
- Yu, X., Liu, T., Gong, M., & Tao, D. (2017a). Learning with biased complementary labels. arXiv preprint [arXiv:1711.09535](https://arxiv.org/abs/1711.09535).
- Yu, X., Liu, T., Gong, M., Zhang, K., & Tao, D. (2017b). Transfer learning with label noise. arXiv preprint [arXiv:1707.09724](https://arxiv.org/abs/1707.09724).
- Zhang, J., Wu, X. D., & Sheng, V. (2016a). Learning from crowdsourced labeled data: A survey. *Artificial Intelligence Review*, 46(4), 543–576.
- Zhang, Y., Chen, X., Zhou, D., & Jordan, M. (2016b). Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(102), 1–44.
- Zheng, Y., Cheng, R., Maniu, S., & Mo, L. (2015a). On optimality of jury selection in crowdsourcing. In *International conference on extending database technology*.
- Zheng, Y., Li, G., & Cheng, R. (2016). Docs: A domain-aware crowdsourcing system using knowledge bases. *Very Large Data Base Endowment*, 10(4), 361–372.
- Zheng, Y., Li, G., Li, Y., Shan, C., & Cheng, R. (2017). Truth inference in crowdsourcing: Is the problem solved? *Very Large Data Base Endowment*, 10(5), 541–552.
- Zheng, Y., Wang, J., Li, G., Cheng, R., & Feng, J. (2015b). QASCA: A quality-aware task assignment system for crowdsourcing applications. In *ACM special interest group on management of data* (pp. 1031–1046).
- Zhong, J. H., Tang, K., & Zhou, Z. H. (2015). Active learning from crowds with unsure option. In *International joint conference on artificial intelligence* (pp. 1061–1068).
- Zhou, D., Basu, S., Mao, Y., & Platt, J. (2012). Learning from the wisdom of crowds by minimax entropy. In *Advances in neural information processing systems* (pp. 2195–2203).
- Zhou, D., Liu, Q., Platt, J., & Meek, C. (2014). Aggregating ordinal labels from crowds by minimax conditional entropy. In *International conference on machine learning* (pp. 262–270).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Bo Han^{1,2}  · Quanming Yao³ · Yuangang Pan¹ · Ivor W. Tsang¹ · Xiaokui Xiao⁴ · Qiang Yang⁵ · Masashi Sugiyama^{2,6}

✉ Ivor W. Tsang
ivor.tsang@uts.edu.au

Bo Han
bo.han@student.uts.edu.au

Quanming Yao
yaoquanming@4paradigm.com

Yuangang Pan
yuangang.pan@student.uts.edu.au

Xiaokui Xiao
xkxiao@nus.edu.sg

Qiang Yang
qyang@cse.ust.hk

Masashi Sugiyama
sugi@k.u-tokyo.ac.jp

¹ Centre for Artificial Intelligence (CAI), University of Technology Sydney, Sydney, Australia

² Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan

³ 4Paradigm Inc., Beijing, China

⁴ Department of Computer Science, National University of Singapore, Singapore, Singapore

⁵ Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

⁶ Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan