Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance

ANONYMOUS

Increasingly, organizations are pairing humans with AI systems to improve decision-making and reducing costs. Proponents of human-centered AI argue that team performance can even further improve when the AI model *explains* its recommendations. However, a careful analysis of existing literature reveals that prior studies observed improvements due to explanations only when the AI, alone, outperformed both the human and the best human-AI team. This raises an important question: can explanations lead to *complementary performance*, *i.e.*, with accuracy higher than both the human and the AI working alone?

We address this question by devising comprehensive studies on human-AI teaming, where participants solve a task with help from an AI system without explanations and from one with varying types of AI explanation support. We carefully controlled to ensure comparable human and AI accuracy across experiments on three NLP datasets (two for sentiment analysis and one for question answering). While we found complementary improvements from AI augmentation, they were *not* increased by state-of-the-art explanations compared to simpler strategies, such as displaying the AI's confidence. We show that explanations increase the chance that humans will accept the AI's recommendation *regardless* of whether the AI is correct. While this clarifies the gains in team performance from explanations in prior work, it poses new challenges for human-centered AI: how can we best design systems to produce complementary performance? Can we develop explanatory approaches that help humans decide whether and when to trust AI input?

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Interactive systems and tools*; • **Computing methodologies** → Machine learning.

Additional Key Words and Phrases: explainable AI, human-AI teams, augmented intelligence, human in-the-loop systems

1 INTRODUCTION

Although the accuracy of Artificial Intelligence (AI) systems is rapidly improving, in many cases it remains risky for an AI to operate autonomously, *e.g.*, in high-stakes domains or when legal and ethical matters prohibit full autonomy. A viable strategy for these scenarios is to form *Human-AI teams*, in which the AI system augments one or more humans by recommending its predictions, but the people retains agency and have accountability on the final decisions. Examples include AI systems that predict likely hospital-readmission to assist doctors with correlated care decisions [5, 11, 65] and AIs that estimate recidivism to help judges decide whether to grant bail to defendants [1, 25]. In such scenarios, the human-AI team is expected to perform better than either would alone.
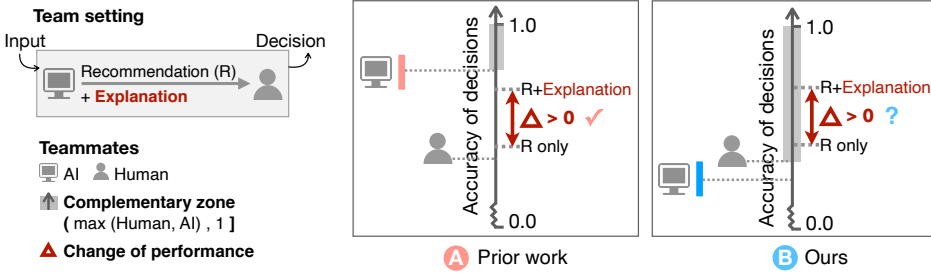
39

Fig. 1. (Best viewed in color) Do AI explanations lead to complementary team performance? In a team setting, when given an input, the human uses (usually imperfect) recommendations from an AI model to make the final decision. We seek to understand if automatically generated explanations of the AI's recommendation improve team performance compared to baselines, such as simply providing the AI's recommendation, R, and confidence. (A) Most previous work concludes that explanations improve team performance (*i.e.*, $\Delta_A > 0$); however, it usually considers settings where AI systems are much more accurate than people and even the human-AI team. (B) Our study considers settings where human and AI performance is comparable to allow room for complementary improvement. We ask, "Do explanations help in this context, and how do they compare to simple confidence-based strategies?" (Is $\Delta_B > 0$?).

Many researchers have argued that such human-AI teams would be improved if the AI systems could *explain their reasoning*. In addition to increasing trust between human and machine, and between humans across the organization [27], one hopes that an explanation should help the responsible human know when to trust the AI's suggestion and when to be skeptical, *e.g.*, when the explanation doesn't "make sense." Such *appropriate reliance* [39] is crucial for users to leverage AI assistance and improve task performance. Indeed, at first glance, it appears that researchers have already confirmed the utility of explanations on tasks ranging from medical diagnosis [10, 47] and data annotation [57] to deception detection [36]. In each case, the papers show that, when the AI provides explanations, team accuracy reaches a level higher than human-alone.

However, a careful reading of these papers shows another commonality: in every situation, adding humans to the team *lowers* performance compared to the AI working independently (Figure 1A & Table 1). Explanations are shown to help raise team performance *closer* to that of the AI, but if pure accuracy were the sole objective, one would achieve an even better result by stripping humans from the loop and letting the AI operate autonomously. Thus, existing work suggests several important open questions for the AI and HCI community: Do explanations provide *complementary performance* (*i.e.*, better than either the AI or the human alone) by enabling humans to anticipate when the AI system is potentially wrong? Furthermore, do explanations provide significant value over simpler strategies such as displaying AI's confidence in these situations? In the quest to build the best human-machine organizations, such questions deserve critical attention.

To explore these questions, we conduct new studies where we control the study design such that the AI's accuracy is *comparable* to the human's (Figure 1B). This decision aims at simulating a setting where (i) there is a stronger incentive to deploy hybrid human-AI teams, and (ii) there exists more potential for complementary performance.

We are especially interested in high-stakes tasks, such as medical diagnosis and recidivism prediction. However, since conducting controlled studies on these domains is challenging and ethically fraught, we started by selecting three common-sense tasks which can be tackled by crowd-workers with little training: a set of *LSAT* questions that require logical reasoning, and sentiment analysis of book and beer reviews.) We measure human skill on these problems and then control AI accuracy by purposely selecting study samples where AI has comparable accuracy. Among

Domain	Task	Performance				
		Metric	Human alone	AI alone	Team	Complementary?
Classification	Deceptive review [36]	Accuracy ↑	51.1%	**87.0%**	74.6%	✗
	Deceptive review [35]	Accuracy ↑	60.4%	**86.3%**	74.0%	✗
	Income category [70]	Accuracy ↑	65%	**75%**	73%	✗
	Loan defaults [21]	Norm. Brier ↑	0	**1**	0.682	✗
	Hypoxemia risk [47]	AUC ↑	0.66	**0.81**	0.78	✗
QA	Quiz bowl [17]	*"AI outperforms top trivia players."*				✗
Regression	House price [55]	Avg. prediction error ↓	$331k	**$200k**	$232k	✗

Table 1. Categorization of recent studies that evaluate effect of automatically generated explanations on human-AI team performance. While explanations did improve team accuracy, performance was not complementary — acting autonomously, the AI would have performed even better. For papers with multiple domains or experiments, we took one sample with the most comparable human and AI performance.

AI-assisted conditions that displayed explanations, we varied the representation of explanations by explaining just the predicted class versus explaining the top-two classes. We also introduced a novel *Adaptive Explanation* mechanism, that switches between strategies based on the AI's confidence, to discourage blind agreement and achieve appropriate reliance. Since explanation *quality* is also a likely factor in its effectiveness, and AI technology keeps improving, we also varied the source of the explanations, considering those generated by state-of-the-art AI methods as well as by skilled humans. In summary, our paper makes the following contributions:

- We highlight an important limitation of previous work on human-centered AI. While several researchers have shown that automatically-generated explanations of AI predictions can increase team performance (Table 1), these results are only shown for settings where the AI system is significantly more accurate than the human partner and the human-AI team. To our knowledge, no previous work has evaluated the effectiveness of AI explanations at producing *complementary performance*, where the human-AI team outperforms both solo human and AI.

- We develop an experimental setting to study human-AI complementary performance, and conducted studies with 1626 users on three tasks, using a variety of explanation sources and representations. We observed complementary performance on every task, but surprisingly, explanations did *not* yield improved team performance, compared to simply showing the AI's confidence. Explanations increased team performance when the AI system was correct, but *decreased* it when the AI erred — so the net value was minimal.

- Our novel *Adaptive Explanations* method, which attempts to solve the problem of blind agreement by reducing human trust when the AI has low confidence, failed to produce significant improvement in final team performance over other explanation types, but there is suggestive evidence that the adaptive approach is promising for high-quality explanations. Through extensive qualitative analysis, we extract potential causes for the behavioral differences among datasets, and summarize them as design implications.

2 BACKGROUND AND RELATED WORK

AI explanations are useful in many scenarios: communicating model predictions [17, 30, 33, 56], teaching humans tasks like translation [48] or content moderation [28], augmenting human analysis procedure [28] or creativity [14], legal imperatives [51, 64], and others. An assumption underlying all these case is that a human will be in the loop, and thus some communication (*e.g.*, an explanation)

is likely to benefit the overall process. As a result, various studies have evaluated the effect of explanations from different dimensions, including whether the explanation improves users' trust on the AI [34, 68] or enables users to simulate the model predictions [13, 55], or assisting developers to debug and train a better ML model [7, 30].

In this paper, we focus explicitly on *AI-assisted decision making* scenarios [3], where an AI assistant (*e.g.*, a classification model) makes recommendations to a human (*e.g.*, a judge), who is responsible for making final decisions (*e.g.*, whether or not to grant bail). In particular, we assess performance in terms of the *accuracy* of the human-AI team. Intuitively, deploying such a human-AI team is ideal if it achieves *complementary performance*, *i.e.*, if it outperforms both the AI and the human acting alone. AI-assisted decision making is often listed as a major motivation for AI explanations. In recent years numerous papers have employed user studies to show that human accuracy increases if the AI system explains its reasoning for tasks as diverse as medical diagnosis, predicting loan defaults, and answering trivia questions [17, 21, 35, 36, 47, 55, 70]. However, as summarized in Table 1, complementary performance was not observed in any of these studies – in each case, adding the human to the loop *decreased* performance.

For example, in Lai *et al.* [35, 36], MTurk workers classified deceptive hotel reviews with predictions from SVM and BERT-based models, as well as explanations in the form of inline-highlights. However, models outperformed every team (see Table 1 and Figure 6 in [35]). Zhang *et al.* [70] noticed the superior behavior of the models in Lai *et al.*'s work, and evaluated the accuracy and trust calibration where the gap between human and the AI performances was less severe. Still, on their task of income category prediction, their Gradient Boosted Trees model had 10% higher accuracy compared to their MTurk workers, which seemed borderline "comparable" at best. Furthermore, when run autonomously, their AI model performed just slightly better than the best team (see Section 4.2.2 and Figure 10 in [70]). Similar performance trend is observed on tasks other than classification. In Sangdeh *et al.* [55], MTurk workers predicted house price using various regression models that generated explanations in terms of most salient features. Their models' predictions resulted in lowest error (See Figure 6 in [55]). In Feng *et al.* [17], experts and novices played Quiz Bowl with recommendation from Elastic Search system. The system explained its predictions by presenting training examples that were influential, and using inline-highlights to explain the connection between question and evidence. Feng *et al.* do not report the exact performance of the AI on their study sample, but mention its superiority in Section 3.1 in [17] pointing out that it outperforms top trivia players.

While it is true that having the AI generate explanations led to a smaller drop in accuracy, removing the human from the loop would have improved performance even more.

At least two potential causes account for the absence of complementary performance in these cases. First, task design may have hindered collaboration: previous researchers considered AI systems whose accuracy was substantially higher than the human's, leading to a small zone of potential complementary performance (see Figure 1). This large gap between human and AI performance made it more likely that human errors were a superset of the AI's, reducing the possibility of a human overseer spotting a machine mistake. Second, even when the task has the potential for complementary performance, it is unclear if the collaboration mechanisms under study supported it. For example, if the AI outputs a prediction with no measure of uncertainty, the human may find it difficult to spot and understand potential AI mistakes (even if presented with an explanation) and may therefore resort to less collaborative heuristics, such as "always trust the AI" or "never trust the AI".

To better understand the role of explanations in producing complementary performance, we conducted user studies that address both of these potential causes by (1) enlarging the zone of

potential complementarity by controlling AI accuracy to match that of an average human,[1] and (2) carefully controlling collaboration mechanisms, including explanations and measures of AI confidence.

3 DESIGN OF USER STUDIES

3.1 Tasks and Explanation Types

Since our interest is in *AI-assisted decision making*, we studied the effect of *local* explanations on team performance – that is, explaining each individual recommendation made by a model [56] rather than trying to provide a global understanding of the full model all at once (*e.g.*, [38]).

We are motivated by high-stakes decisions, such as medical diagnosis, but were daunted by the difficulty of running a large-scale randomized control trial in such a domain. Instead we conducted crowd-sourced studies on Amazon Mechanical Turk and chose proxy domains that required no specific background knowledge. We chose text classification because it is a popular task in natural language processing (NLP) that has been used in several previous studies on human-AI teaming [17, 24, 35, 43, 53, 70], requires little domain expertise, and is thus amenable to crowdsourcing. Further, there are various explanation algorithms for classification, whose practical benefits for team performance remain unclear. Specifically, we selected two sentiment analysis datasets where specialized knowledge is not needed: beer reviews [6] and book reviews [50]. Explanations in this task are represented by *saliency* scores, where each input feature (*i.e.*, each word) is assigned a real number representing its contribution towards the prediction (positive or negative). We display these explanations with *inline-highlights*, where the interface highlights spans of text that are most salient to the model's prediction directly in the input text, so the user need not go back and forth between input and explanation [18, 36, 42, 62]. Figure 2 shows one example beer review used in our study.

We also experimented with a more challenging task: Law School Admission Test (LSAT) questions[2] where every question contains four options with a unique correct answer (Figure 3). Answering *LSAT* questions requires no specialized knowledge except common-sense reasoning skills, such as recognizing logical connections and conflicts between arguments [69]. In this case it is unclear how inline-highlights could be used to communicate logical constructs (*e.g.*, contradiction is not visible by highlighting the input alone); thus, we turned to narrative explanations which justify a candidate answer in natural language (Section 3.5).

Admittedly, explanations can also be presented in other ways. For structured data, most studies present the salient features, sometimes paired with importance weights [8, 21, 47, 52, 63]. On vision tasks, some work takes the demonstrate-by-example approach, *i.e.*, by explaining an example with its closest match in the training set [48, 67]. These two methods are typically less suitable for domains involving unstructured text, as the text features are sparser (*i.e.*, a feature list can be overwhelming) and are hard to parse (and therefore pairing with training examples is unfeasible).

3.2 Pilot Study on Sentiment Classification

To develop research strategies for the main study, we collected feedback about common explanation approaches through a pilot study. The between-subject pilot study involved 150 crowd workers,

---

[1]Of course, complementary performance may be possible even in situations when one of the team partners is significantly more accurate than the other. For example, a low-accuracy team member may be valuable if their errors are independent, because they may be able to spot mistakes made by the team majority. However, it is more difficult to observe complementary performance in such settings, so we first consider the case where humans and AI have similar accuracy. If explanations cannot provide value in such settings, it will be even more difficult to show complementary performance when teammates have disparate skills.

[2]https://en.wikipedia.org/wiki/LawSchoolAdmissionTest

50 per condition, each of whom judged the sentiment of 50 beer reviews using assistance from a logistic regression classifier (more details about the dataset is in Section 3.5). In all teaming conditions, the model always presented at least its prediction and confidence.

Explanations. In some conditions, AI support was extended through the following explanations, also common in previous work [21, 35]:

(1) *Explain-Top-1* explains just the predicted class by highlighting the most influential words for that class.
(2) *Explain-Top-2* explains the top two predicted classes, and unlike Explain-Top-1, it also color-codes and highlights words for the other sentiment value.

Observations. In contrast to many previous studies, we observed no significant changes or improvements in aggregated team accuracy from displaying explanations. However, *explaining just the predicted class performed better than explaining both classes.*

Additionally, when we analyzed performance, conditioning on whether the recommendation was correct, we observed that *explanations increased reliance on recommendations even when they were incorrect.* Explaining the predicted class slightly improved performance (compared to confidence only) when the recommendation was correct but decreased performance when it was incorrect. This is consistent with psychology literature [32], which has shown that human explanations cause listeners to agree even when the explanation is wrong.

This effect was less pronounced when explaining both classes with Explain-Top-2, presumably because it encouraged users to consider alternatives, deterring over-reliance. In Figure 2, for example, if counter-argument (d) was not highlighted, participants could easily stop reading at the highlighted first sentence and overlook the negative ending. These observations are consistent with previous work in psychology which showed that when a *person* presented an explanation for a proposition, listeners were more likely to think the proposition was true [32]. As a corollary, *when the model was correct, Explain-Top-1 was more effective than Explain-Top-2.*

Finally, *participants indicated that they wanted higher quality explanations.* In particular, crowd-workers were confused by situations where explanations did not seem to correlate with model behavior or did not make intuitive sense.

3.3 Additional Explanation Strategies and Sources

Adaptive Explanations. The pilot study indicated that Explain-Top-2 was more beneficial than Explain-Top-1 when the classifier made mistakes, but not otherwise. Relying on commonly seen correlations between mistakes and low-confidence [26], we developed a new, dynamic strategy, *adaptive explanations*, that switches between Explain-Top-1 and Explain-Top-2 depending on the AI's confidence. This method explains the top-two classes only when classifier confidence is below a threshold, explaining only the top prediction otherwise. Intuitively, the *Adaptive* explanation method was inspired by an efficient assistant that divulges more information (confessing doubts and arguing for alternatives) only when it is unsure about its recommendation. We test all three strategies in our final studies.

Expert-Generated Explanations. Users in our pilot study were confused when explanations did not make intuitive sense, due to either the quality of the underlying AI or the explanation proce-dure. We compared a wide variety of NLP models and automatic explainers to find the most suitable AI-generated explanations (details in Appendix A.1). However, we also added *expert-generated* explanations to serve as an upper bound on explanation quality. For sentiment classification, one author created expert explanations by selecting one short, convincing text phrase span for each class
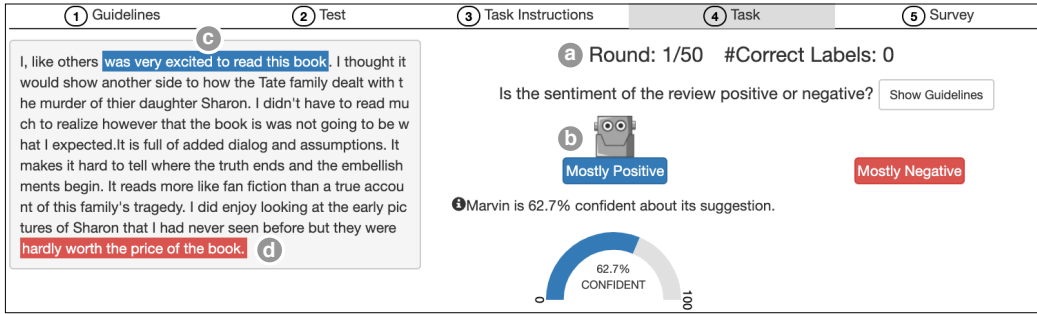
Fig. 2. A screenshot of the **Team (Adaptive, Expert)** condition for the *Amzbook* reviews dataset. Participants read the review (left pane) and used the buttons (right pane) to decide if the review was mostly *positive* or *negative*. The right pane also shows progress and accuracy (a). To make a recommendation, the AI (called "Marvin") hovers above a button (b) and displays the confidence score under the button. In this case, the AI incorrectly recommended that this review was positive, with confidence 62.7%. As part of the explanation, the AI highlighted the most positive sentence (c) in the same color as the *positive* button. Because confidence was low, the AI also highlights the most negative sentence (d) to provide a counter-argument.

(positive or negative). For *LSAT*, we found no automated method that could generate reasonable explanations (unsurprising, given that explanations rely on prior knowledge and complex reasoning); thus, we used expert explanations exclusively. The prep book from which we obtained our study sample contained expert-written explanations for the correct answer, which one author condensed to a maximum of two sentences. Additionally, since the book did not provide explanations for alternate choices, we created these by manually crafting a logical supporting argument for each choice that adhered to the tone and level of conciseness in the other explanations.

3.4 Hypotheses and Conditions

Informed by our pilot study, we formulated the following hypotheses for sentiment analysis:

H1 *Among current AI explanation representations, explaining the predicted class will perform better than explaining both classes.*

H2 *The better representation, Explain-Top-1, will still perform similarly to simply showing confidence.*

H3 *Our proposed representation with* Adaptive explanations*, which combines benefits of existing representations, will improve performance.*

H4 *Adaptive explanations would perform even better if AI could generate higher quality explanations.*

Since generating AI explanations for *LSAT* was not feasible (Section 3.3), we slightly modified the hypothesis for *LSAT*: we omitted the hypothesis on explanation quality (H4) and tested the first three hypotheses using expert- rather than AI-generated explanations.

Conditions. For both domains, we ran two baseline conditions: unassisted users (**Human**) and simple AI assistance: showing the AI's recommendation and confidence but no explanations (**Team (Conf)**). All explanation conditions showed the AI's confidence score, and we ran only explanation conditions that were essential for testing our hypothesis. In the rest of the paper, we indicate explanation conditions using the following template: **Team (Strategy, Source)**. For example, Team(Explain-Top-1, AI) indicates the condition that shows the AI's explanations for the top prediction.

Fig. 3. A screenshot of **Team (Adaptive, Expert)** for *LSAT*. Similar to Figure 2, the interface contained a progress indicator (a), AI recommendation (b), and explanations for the top-2 predictions (c and d). To discourage participants from blindly following the AI, all AI information is displayed on the right.

Interface. Figure 2 shows an example UI for sentiment classification. In all explanation conditions, explanations are displayed as inline highlights, with the background color aligned with the positive/negative label buttons. The display for Team (Adaptive, AI) is similar to Figure 2 except for that its use of AI-generated explanations, which in practice picked multiple short phrases instead of a full sentence. In other conditions, only some of these cues are shown: in Team (Explain-Top-1, AI) the counter-argument (d) is always missing, and in Team (Conf) no explanations are highlighted. Figure 3 shows a screenshot of the user interface for *LSAT* in the Team (Adaptive, Expert) condition.

3.5   AI Systems and Study Samples

Sentiment Classification. To prepare each dataset (*Beer* and *Amzbook*) for training classification models, we binarized the target labels, split the dataset into training and test sets (80/20 split), removed class imbalance from the train split by over-sampling the minority class, and further split the training set to create a validation set.

To create our AI, for each dataset we fine-tuned a RoBERTa-based text classifier [45] on the training dataset and performed hyper-parameter selection on the validation set. Explanations from a combination of RoBERTa and LIME (a popular posthoc explainer [56]) were consistently ranked the highest among the various systems we tried in an explainer comparison study (details in Appendix A.1). Despite offering accurate predictions, RoBERTa generated poorly calibrated confidence scores, a common issue with neural networks [22], which we mitigate with *posthoc calibration* (isotonic regression [4]) on the validation set.

For each domain, we selected 50 examples from the test set to create our study sample. To select these examples, we first conducted additional pilot studies to establish accuracy of unassisted users, which we observed were 87% for *Beer* and 85% for *Amzbook*[3] We then selected 50 unambiguous examples so that the AI's accuracy was 84% (i.e., comparable to human accuracy), with equal false positive and false negative rates.

For Adaptive explanations, we used the classifier's median confidence as the threshold, which was 89.2% for *Beer* and 88.9% for *Amzbook*. While one might learn a better threshold from data, we leave that to future work. For each dataset, of the 50 total examples, the Adaptive explanation

---

[3]Again, each condition contained 50 crowdworkers. Together with *LSAT*, we estimated the human accuracy on the three datasets with another 150 crowdworkers.

condition explained 18 correctly predicted and 7 incorrectly predicted examples using Explain-Top-2 and explained the rest 25 using Explain-Top-1.

LSAT. We finetuned a RoBERTa model on ReClor data [69], which contains questions from standardized exams like the *LSAT* and GMAT.[4] For our study sample, we selected 20 examples from an *LSAT* prep book [9]. We originally used questions from online practice exams, but our pilot showed the potential for subjects to cheat by searching for answers on the Web. We verified that our final questions were not easily searchable online, and were not included in the training dataset. We selected fewer questions than we did for sentiment analysis because *LSAT* questions are more time consuming to answer and could fatigue participants: *LSAT* questions took 57.4 ± 79.6 seconds to answer compared to 16.6 ± 24.7 seconds for *Beer* and 18.5 ± 98.3 seconds for *Amzbook*. The RoBERTa model achieved 65% accuracy on these examples, comparable to the 67% human accuracy that we observed in our pilot study.

3.6 Study Procedure

Sentiment Classification. For the final study, participants went through the following steps: 1) A landing page first explained the payment scheme; the classification task was presented (here, predicting the sentiment of reviews); and they were shown dataset-specific examples. 2) To familiarize them with the task and verify their understanding, a screening phase required the participant to correctly label four of six reviews [44]. Only participants who passed the gating test were allowed to proceed to the main task. 3) The main task randomly assigned participants to one of our study conditions (Section 3.3) and presented condition-specific instructions, including the meaning and positioning of AI's prediction, confidence, and explanations. Participants then labeled all 50 study samples one-by-one. For a given dataset, all conditions used the same ordering of examples. 4) A post-task survey was administered, asking whether they found the model assistance to be helpful, their rating of the usefulness of explanations in particular (if they were present), and their strategy for using model assistance.

We recruited participants from Amazon's Mechanical Turk (MTurk), limiting the pool to subjects from within the United States with a prior task approval rating of at least 97% and a minimum of 1,000 approved tasks. To ensure data quality, we removed data from participants whose median labeling time was less than 2 seconds or those who assigned the same sentiment label to all examples. We pre-registered this process based on our experiences with the pilot study. In total, we recruited 566 (*Beer*) and 552 (*Amzbook*) crowd workers, and in both datasets, 84% of participants passed the screening and post-filtering. Eventually, we collected data from around 100 participants (ranging from 93 to 101 due to filtering) per condition.

Workers received a base pay of $0.50 for participating, a performance-based bonus for the main task, and a fixed bonus of $0.25 for completing the survey. Our performance-based bonus was a combination of linear and step functions on accuracy: we gave $0.05 for every correct decision in addition to an extra $0.50 if total accuracy exceeded 90% or $1.00 if accuracy exceeded 95%; we assigned additional bonuses to motivate workers to strive for performance in the complementary zone. Since we fixed the AI performance at 84%, humans could not obtain the bonus by blindly following the AI's recommendation. Participants spent 13 minutes on average on the experiment, and received an average payment of $3.35 (equals an hourly wage of $15.77).

Modifications for LSAT. For the *LSAT* dataset, we used a very similar procedure but used two screening questions and required workers to answer both correctly. We used a stricter passing requirement to avoid low-quality workers who might cheat, which we observed more for this
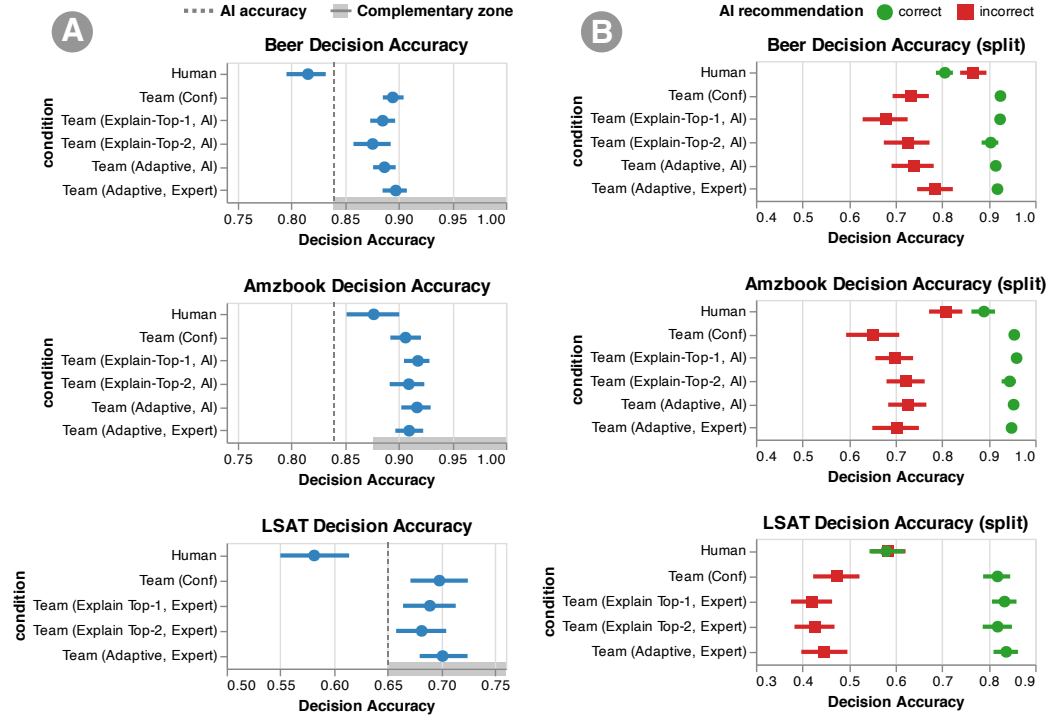

Fig. 4. Team performance (with average accuracy and 95% confidence interval) achieved by different explana-
tion conditions and baselines for three datasets, with around 100 participants per condition. (A) Across every
dataset, all team conditions achieved complementary performance. However, we did not observe significant
improvements from using explanations over simply showing confidence scores. (B) Splitting the analysis
based on the correctness of AI accuracy, we saw that for *Beer* and *LSAT*, Explain-Top-1 explanations worsened
performance when the AI was incorrect, the impact of Explain-Top-1 and Explain-Top-2 explanations were
correlated with the correctness of the AI's recommendation, and Adaptive explanations seemed to have
the potential to improve Explain-Top-1 when the AI was incorrect, and to retain the higher performance of
Explain-Top-1 when the AI was correct.

task in our pilots. We again used MTurk with the same filters as sentiment classification, and we
posthoc removed data from participants whose median response time was less than three seconds.
508 crowd workers participated in our study, 35% of whom passed the screening and completed the
main task, resulting in a total of 100 participants per condition.

Workers received a base pay of $0.50 for participating, a performance-based bonus of $0.30 for
each correct answer in the main task, and a fixed bonus of $0.25 for completing an exit survey.
They received an additional bonus of $1.00, $2.00, and $3.00 for reaching an overall accuracy of 30%,
50%, and 85% to motivate workers to answer more questions correctly and not cheat. The average
completion time for the LSAT task was 16 minutes, with an average payment of $6.30 (equals an
hourly wage of $23.34).

4 RESULTS

4.1 Effect of Explanation on Team performance

Figure 4A shows team performance (*i.e.*, accuracy of final decision) for each domain and condition.
We tested the significance of our results using Student's T-tests with Bonferroni corrections.

The baseline team condition, Team (Conf), achieved complementary performance across tasks. For example, for *Beer*, providing AI recommendations and confidence to users increased their performance to ($\mu = 0.89 \pm \sigma = 0.05$), surpassing both AI accuracy (0.84) and unassisted human accuracy (0.82 ± 0.09). Similarly, Team (Conf) achieved complementary performance for *Amzbook* and *LSAT*, with relative gains of 2.2% and 20.1% over un-assisted participants (Figure 4A).

We did not observe a significant difference between Explain-Top-1 and Explain-Top-2, or that **H1** was not supported. For example, in Figure 4A of *Beer*, explaining the top prediction performed marginally better than explaining the top-two predictions, but the difference was not significant ($z=0.85$, $p=.40$). The same was true for *Amzbook* ($z=0.81$, $p=.42$) and *LSAT* ($z=0.42$, $p=.68$).

We did not observe significant improvements over the confidence baseline by displaying explanations. For example, for *Beer*, Team (Conf) and Team (Explain-Top-1, AI) achieved similar performance, with the accuracy being 0.89 ± 0.05 vs. 0.88 ± 0.06 respectively; the difference was insignificant ($z=-1.18$, $p=.24$). We observed the same pattern for *Amzbook* ($z=1.23$, $p=.22$) and *LSAT* ($z=0.427$, $p=.64$). As a result, we could not reject our hypothesis **H2** that Explain-Top-1 performs similar to simply showing confidence . This result motivates the need to develop new AI systems and explanation methods that provide true value to team performance by supplementing the model's confidence, perhaps working in tandem with confidence scores.

Though the third explanation strategy was designed to alleviate the limitations of Explain-Top-1 and Explain-Top-2 in our experiments, **we did not observe improvements from using Adaptive explanations.** For example, we did not observe any significant differences between Team (Adaptive, AI) and Team (Conf) for *Beer* ($z=-1.02$, $p=.31$) or *Amzbook* ($z=1.08$, $p=.28$). We did not observe significant differences between Team (Adaptive, Expert) and Team (Conf) for *LSAT* ($z=0.16$, $p=.87$). More surprisingly, switching the source of Adaptive explanations to expert-generated did not significantly improve sentiment analysis results. For example, in Figure 4A, the differences in performance between Team (Adaptive, Expert) and Team (Adaptive, AI) were insignificant: *Beer* ($z=1.31$, $p=.19$) and *Amzbook* ($z=-0.78$, $p=.43$). As a result, we could not reject the null hypotheses for either **H3** or **H4**.

While Adaptive explanations did not significantly improve team performance across domains, further analysis may point a way forward by combining the strengths of Explain-Top-1 and Explain-Top-2. Considering team performance split by whether the AI made a mistake (Figure 4B), we observe that explaining the top prediction lead to better accuracy when the AI recommendation was correct but worse when the AI was incorrect, as in our pilot study. Consistent with our intuition, for *Beer*, Adaptive explanations improved performance over Explain-Top-1 when the AI was incorrect and improved performance over Explain-Top-2 when the AI was correct, although the effect was smaller on other datasets.

While Figure 4B shows team performance, the promising effects of Adaptive explanations are clearer if we study the agreement between AI predictions and human decisions (Figure 5). Adaptive explanations seem to encourage participants to consider the AI more when it is confident and solve the task themselves otherwise. Unfortunately, as our experiments show, the effect of using Adaptive did not seem sufficient to increase final team accuracy, possibly for two reasons: (1) in high confidence regions (circles in Figure 5), not only did workers have to agree more, but they also had to identify cases where the model failed with very high confidence (unknown unknowns [37]). Identifying unknown unknowns could have been a difficult and time-consuming task for workers, and they may have needed other types of support that we did not provide. (2) In low confidence regions (rectangles), not only did workers have to disagree more, but they also had to be able to solve the task correctly when they disagreed. Explain-Top-2 explanations might have enabled them
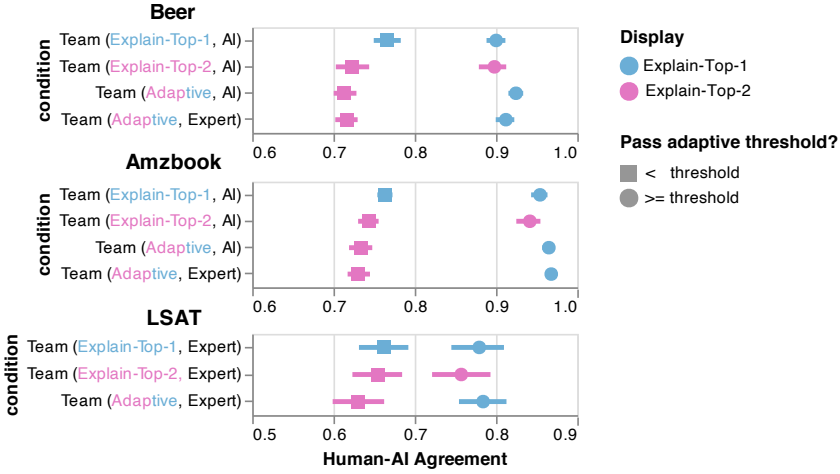
Fig. 5. Relative agreement rates between humans and AI (*i.e.*, does the final human decision match the AI's suggestion?) for various conditions, with examples split by whether AI's confidence exceeded the threshold used for Adaptive explanations. Across the three datasets, Adaptive explanations successfully reduced the human's tendency to blindly trust the AI when it was uncertain and more likely to be incorrect. For example, comparing Team (Explain-Top-1, AI) and Team (Adaptive, AI) on low confidence examples that did not pass the threshold (rectangles), participants in Explain-Top-2 (pink rectangles) were less likely to agree with the AI compared to those who saw Explain-Top-1 (blue rectangles). Note that the display (color encoded) is determined by the conditions on the y-axis.
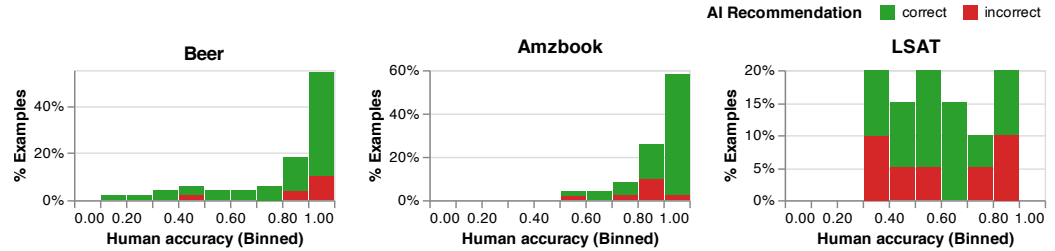
Fig. 6. The distribution of study examples as a function of average human accuracy. For each domain, examples on the right were easy for most humans working alone. Both *Beer* and *LSAT* show a distribution that shows potential for complementary team performance: humans can correct easy questions mistaken by the AI (red bars towards the right), and, conversely, the AI may add value on examples where humans frequently err (green bars towards the left). In contrast, *Amzbook* showed less potential for this kind of human-AI synergy, with less "easy for human" questions (bars towards the left).

to suspect the model more, but it is unclear if they helped participants make the right decisions. This indicates that more sophisticated strategies are needed to support humans in both situations.

Differences in expertise To understand how differences in expertise between humans and the AI impacts team performance, we computed the average unassisted-user accuracy on study examples and overlayed the AI's expertise (whether the recommendation was correct) in Figure 6. The figure helps explain why users benefited more from AI recommendations for both *Beer* and *LSAT* datasets.

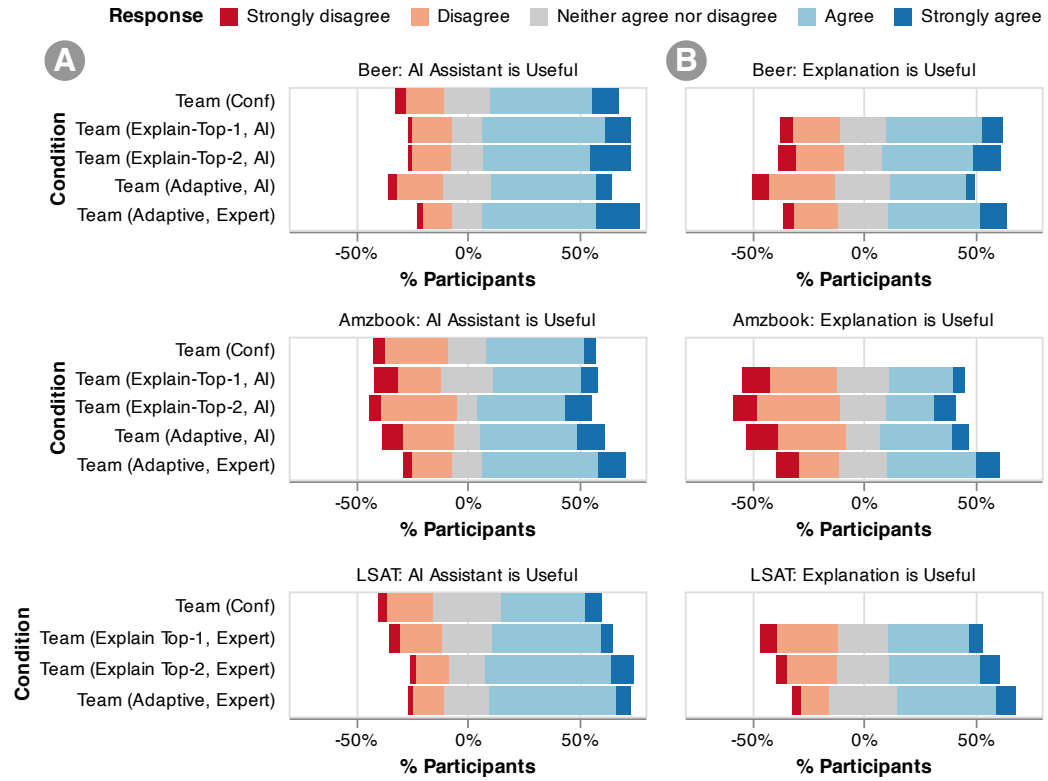Fig. 7. Analysis of participant response to two statements: (A) "AI's assistance (*e.g.*, the information it displayed) helped me solve the task", and (B) "AI's explanations in particular helped me solve the task." Across datasets, a majority of participants found AI assistant to be useful, and they rated all the conditions similarly, with a slight preference towards Team (Adaptive, Expert). In contrast to AI's overall usefulness, fewer participants rated explanations as useful, particularly Explain-Top-2 explanations across sentiment classification datasets. Participants also had a clearer preference for higher-quality (expert) Adaptive explanations.

There was a significant fraction of examples that the AI predicted correctly but humans struggled with (green bars to the left), while the same was not true for *Amzbook* (where AI recommendations did not help as much). Further, when the AI was incorrect, explaining predictions on *Amzbook* via Explain-Top-1 improved performance by 5% over showing confidence (Figure 4B), but it decreased performance for *Beer* and *LSAT*. One hypothesis that explains this observation is that most AI mistakes were predicted correctly by most humans on *Amzbook* (red bars were mostly towards the right). After observing clear model mistakes, participants may have learned to rely on them less, despite the convincing-effect of explanation. This pattern was more visible in participants' self-reported collaboration approaches – *Amzbook* participants reportedly ignored the AI's assistance the most (Section 4.3).

4.2 Survey Responses on Likert Scale Questions

Two of the questions in our post-task survey requested categorical ratings of AI and explanation usefulness. While participants generally rated AI assistance useful across datasets (Figure 7A), the improvements in ratings between most explanations and simply showing confidence were

Codes	Definitions and Examples	#Participants
Overall Collaboration Approach (codes are mutually exclusive)		
Mostly Follow AI	The participant mostly followed the AI.	23 (6%)
	"I went with Marvin most times."	
AI as Prior Guide	Used AI as a starting reference point.	190 (47%)
	"I looked at his prediction and then I read the passage."	
AI as Post Check	Double-checked after they made their own decisions.	102 (25%)
	"I ignored it until I made my decision and then verified what it said."	
Mostly Ignore AI	Mostly made their own decisions without the AI.	90 (22%)
	"I didn't. I figured out the paragraph for myself."	
The Usage of Explanation (codes can overlap)		
Used Expl.	Explicitly acknowledged they used the explanation.	138 (42%)
	"I skimmed his highlighted words."	
Speed Read	Used explanations to quickly skim through the example.	29 (9%)
	"I looked at Marvin's review initially then speed read the review. "	
Validate AI	Used the explanation to validate AI's reasoning.	17 (5%)
	"Marvin focuses on the wrong points at times. This made me	
	cautious when taking Marvin's advice."	
The Usage of Confidence (codes can overlap)		
Used Conf.	Explicitly acknowledged they used the confidence.	90 (22%)
	"I mostly relied on Marvin's confident levels to guide me."	
Conf. Threshold	Was more likely to accept AI above the threshold.	24 (6%)
	"If Marvin was above 85% confidence, I took his word for it."	
Others (codes can overlap)		
Fall Back to AI	Followed the AI's label if they failed to decide.	54 (13%)
	"i used it if i was unsure of my own decision."	
Updated Strategy	Changed their strategy as they interacted more.	12 (2%)
	"I decided myself after seeing that sometimes Marvin failed me. "	

Table 2. The codebook for participants' descriptions of how they used the AI, with the number of self-reports.

marginal. The difference was more clear for high-quality adaptive explanations; *e.g.*, for *Beer*, 70% of participants rated AI assistance useful with Team (Adaptive, Expert) in contrast to 57% with Team (Conf). We observed a similar pattern on *Amzbook* (65% vs. 49%) and *LSAT* (63% vs. 45%), though on *LSAT*, Team (Explain-Top-2, Expert) received slightly higher ratings than Team (Adaptive, Expert) (66% vs. 63%). Figure 7B shows that participants' ratings for usefulness of explanations were lower than the overall usefulness of AI's assistance (in A). Again, expert-generated Adaptive explanations received higher ratings than AI-generated ones for *Beer* (53% vs. 38%) vs. *Amzbook* (50% vs. 40%). This could indicate that showing higher quality explanations improves users' perceived helpfulness of the system.

4.3 Qualitative Coding of Collaboration Approach

To better understand how users collaborated with the AI in different tasks, we coded their response to the prompt: "Describe how you used the information Marvin (the AI) provided." Two annotators independently read a subset of the responses to identify emergent codes and, using a discussion period, created a codebook (Table 2). Using this codebook, for each team condition and dataset, they coded a total of 30 random responses: 28 were unique and 2 overlapped between annotators, allowing us to compute inter-annotator agreement. Our final analysis used 409 unique responses
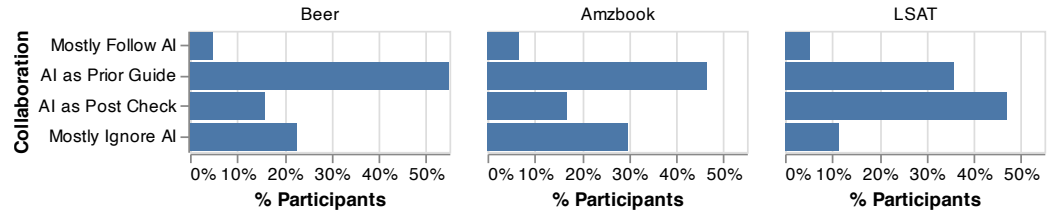
Fig. 8. Instead of ignoring or strictly following the AI, participants reported taking the AI information into consideration most of the time. They most frequently used AI as a prior guide in sentiment analysis, but used it as post-check in *LSAT*. They were also more likely to ignore the AI in sentiment analysis than in *LSAT*.

after removing 11 responses deemed to be of poor quality (Table 2). We scored the inter-annotator agreement with both the Cohen's κ and the raw overlap between the coding. We achieved reasonably high agreements, with an average $\mu(\kappa) = 0.71$, $\sigma(\kappa) = 0.18$ (the average agreement was 93% ± 6.5%). We noticed the following, which echo the performance differences observed across datasets:

Most participants used the AI's recommendation as a prior or to double-check their answers. For all datasets, more than 70% of participants mentioned they would partially take AI's recommendation into consideration rather than blindly following AI or fully ignoring it (Figure 8). Participants used the AI as a prior guide more than as a post-check for sentiment analysis, but not for *LSAT*, which aligns with our interface design: for *LSAT*, AI recommendations were on a separate pane, encouraging users to solve the task on their own before consulting the AI.

Participants ignored the AI more on domains where AI expertise did not supplement their expertise. Figure 8 shows that while only 11% of *LSAT* participants claimed that they mostly ignored the AI, the ratio doubled (*Beer*, 23%) or even tripled (*Amzbook*, 30%) for sentiment analysis. As discussed in Figure 6, this may be due to correlation differences between human and AI errors for different datasets: *Amzbook* participants were less likely to see cases where AI was more correct than they were, and therefore they may have learned to rely less on it. For example, one participant in *Amzbook* mentioned, *"I had initially tried to take Marvin's advice into account for a few rounds, and stopped after I got 2 incorrect answers. After that I read all of the reviews carefully and followed my own discretion."*

In contrast, a *Beer* participant relied more on the AI once realizing it could be correct: *"At first I tried reading the passages and making my own judgments, but then I got several items wrong. After that, I just switched to going with Marvin's recommendation every time."*

In addition to the user's collaboration behavior, these differences between domains may have affected our quantitative observations of team performance. For example, a small difference between human and AI expertise (distribution of errors) means that the improvements in performance when the AI is correct would be less substantial. In fact, in Figure 4B, if we compare the team performance when the AI is correct, the difference between team conditions and the human baseline is least substantial for *Amzbook*.

Some participants developed mental models of the AI's confidence score to determine *when* to trust the AI. Among participants who mentioned they used confidence scores (90 in total), 27% reported using an explicit confidence threshold, below which they were likely to distrust the AI. The threshold mostly varied between 80 to 100 (83 ± 8 for *Beer*, 89 ± 7 for *Amzbook*, and 90 ± 0 for *LSAT*) but could go as low as 65, indicating that different users built different mental models about when they considered AI to be "trustworthy." Furthermore, **some participants consigned the task to AI when they were themselves uncertain.** For example, 13% participants mentioned
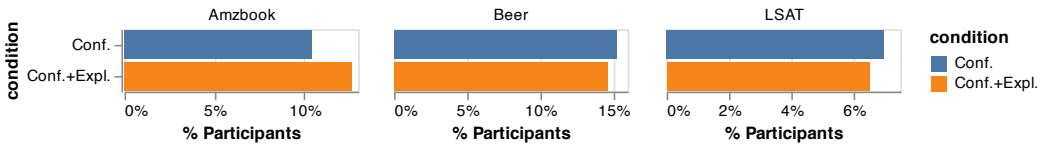
Fig. 9. On comparing the occurrence of the code "Used Conf." in just the confidence condition and in those with explanations, we saw a similar proportion of participants that explicitly acknowledged using confidence, regardless of whether they saw an explanation.

that they would go with the AI's decision if they were on the fence by themselves: *"There were some that I could go either way on, and I went with what Marvin suggested."*

These user behaviors are similar to observations in psychology literature on *Truth-Default Theory* [41], which shows that people exhibit *truth-default* behavior: by default people are biased to assume that the speaker is being truthful, especially when *triggers* that raise suspicion are absent. Furthermore, our participants' distrust in low-confidence recommendations is also consistent with examples of triggers that cause people to abandon the truth-default behavior.

Note that across all three domains, the same proportion of participants self-reported using AI's confidence scores regardless of whether they saw explanations (Figure 9). This indicates that confidence is the primary information that people use to collaborate with AI agents.

Explanations can help participants validate the AI's decisions, and the inline-highlight format helped participants speed up their decision making. Among the participants who explicitly mentioned using explanations, 27% in *Beer* and 32% in *Amzbook* reported that they used inline-highlight explanations to read review text faster. Since *LSAT* explanations required reading additional text, we did not expect *LSAT* users to find this benefit. Interestingly, for *Beer* and *Amzbook*, while a small percentage of users (17%) reported using the explanations to validate the AI's decisions (see Figure 2), only 2% did so in *LSAT*. This could be because *LSAT* is a harder task than sentiment analysis, and verifying AI's explanations is costlier. Other participants mostly mentioned that they would supplement their own reasoning with the AI's: *"I read the Marvin rationale and weighed it against my intuition and understanding."*

5 DISCUSSION & DIRECTIONS FOR FUTURE RESEARCH

Increasingly, we see organizations incorporating AI systems into the work process to automate routine tasks and advise human decision makers on higher stake decisions, such as medical diagnosis [11, 31], wildlife protection [19], improving housing for homeless youth [12, 29], and recidivism prediction [23, 25, 60]. While there are many motivations for creating these human-AI teams, a primary one is increasing the effectiveness of decision making, while reducing the cost of operations. But increased use of AI brings several well-known dangers as well, including bias and erratic behavior [64]. Many researchers have argued that if AI could explain its reasoning, these dangers would be reduced and the resulting socio-technical system might make consistent and accurate decisions.

We investigate this motivation, by evaluating the impact of AI explanations on team performance for the case when humans have comparable accuracy to their AI teammates. Though conducted in a limited scope (Section 5.1), we believe our findings provide informative insights into future work. In this section, we reflect on our study through an overview, and then provide more detailed observations that may guide the design of future explanation techniques and further experiments.

To our surprise, the simple approach of showing the AI's confidence worked as well as any of the actual explanations strategies that we considered. The use of explanations increased the

human's blind trust in the AI system, rather than appropriate reliance [39], making them more likely to adopt its recommendation both when the AI was right and when it was wrong. While this is consistent with the psychology literature [32], which has shown that human explanations cause listeners to agree even when the explanation is wrong, we believe the phenomenon has not previously been noted with AI-generated explanations.

Our observation raises the question "What is the purpose of including a human in the loop?" We have chosen to focus on complementary performance, where adding a human improves accuracy beyond what either human or AI could achieve alone; but in other contexts, a human may be required for important and essential moral or legal reasons. In these contexts, we question if the purpose of an explanation is simply to soothe the human, making them more compliant so they are more likely to blindly agree with the computer, which seems unsatisfactory. Compared to the widely adopted strategy of outputting the strongest convincing evidence for every class, we believe explanations that properly raise doubts may be even more important (Section 5.3).

Our Adaptive Explanation is aimed at enabling and improving human oversight by encouraging the human to think more carefully when the system has low confidence. While the relative agreement rates show promise (Figure 5), the method was not sufficient to significantly increase the final team accuracy (Figure 4). Further analyses on the dataset distribution differences and participants' self-reported collaboration approaches point out that (1) developers of human-AI teaming tools should assess the complementarity between the AI and the human beyond just looking at their solo performances (Section 5.2), and (2) researchers studying human-AI collaboration should explore alternative collaboration architectures (Section 5.3).

5.1 Limitations

We focused solely on the use of explanations to improve the accuracy of human-AI teams. As mentioned before, AI explanations have other motivations not addressed by this paper, such as debugging AI output [56], teaching [48], satisfying legal requirements [64], and increasing trust between sub-organizations [27]. We also did not explore benefits from explanations on metrics besides accuracy, such as increasing speed as reported by some users in Section 4.2. Further, because it is so challenging to run large-scale experiments on high-stakes domains, such as medical diagnosis, with hard to locate, expert users, we restricted ourselves to tasks amenable to crowdsourcing (text classification and question answering). While our results may not generalize to high stakes domains, we believe our large-scale study provides informative insights that will inform future work on these domains. Finally, we only explored a small fraction of possible ways to present explanations: highlighting key words [18, 36, 42, 62] and natural language arguments. It may be that alternative approaches provide more benefit to team performance.

5.2 When to collaborate?

Comparable accuracy does not guarantee complementary partners. In some of our experiment domains, AI errors correlated much more strongly with humans' than in others. For example, consider the top row of Figure 4B for *Beer*, where there is an inverse correlation — humans did worse on the problems that AI got correct than problems on which the AI was incorrect. For the *Amzbook* domain, however, the situation was reversed — both human and AI found the same problems hard or easy. It should not be surprising, then, that the benefit of AI assistance proved much greater for *Beer* than for *Amzbook*: by bringing different "perspectives" the two teammates skills were more complementary.[5] This suggests that, a comparable accuracy between the human

---

[5]Because the *LSAT* problems were so difficult, there was also more potential for complementary performance, and patterns in this domain resemble those of *Beer*.

and AI is not sufficient for creating a team with complementary performance. Rather, in an ideal team, the human and AI would have minimally overlapping mistakes so that their is a greater opportunity to correct each other's mistakes. As some recent work has suggested [2, 66], in some cases it may be useful to directly optimize for complementary behavior instead of automation accuracy. Models that are trained to make errors when the human is correct are more likely to benefit from human interventions on those errors, and therefore are likely to achieve a more complementary team performance.

The definition of "complementarity" can go beyond accuracy. Our work, as well as those in Table 1, evaluated team performance along one dimension: accuracy of decisions. In general, one may wish to achieve complementary performance on a multi-dimensional metric. For example, when grading programming exercises with an AI [20], the team may wish to balance several metrics, such as speed of grading, precision of error discovery, and quality of feedback. To achieve this, the human and AI could maximize their talents in more than one dimensions: AI could use its computation power to quickly gather statistics and highlight commonly missed corner cases, whereas the human teacher could focus on ranking the intelligence of the student's proposed algorithms. In fact, research shows that large collaborative communities such as Wikipedia require AI systems that balance multiple aspects such as reducing human effort, improving trust and positive engagement [58]. We encourage future researchers to extend the definition of complementarity, and to evaluate the impact of explanations on those dimensions accordingly.

5.3 How to collaborate?

Complementarity requires more than a shift in agreement. Our Adaptive explanation reduced user tendency to blindly trust the AI, yet its impact on the team accuracy was less observable. This highlights that complementary performance requires more than simple changes in agreement between the human and AI. As mentioned in Section 4.1, we note: (1) when the model is confident, not only should workers agree more, but they also have to identify cases where the model fails with high confidence (unknown unknowns [37]). (2) When the model is not confident, not only should workers disagree more, but they also have to solve the task correctly once they discern that AI is wrong. The Explain-Top-2 explanations might cue them to suspect the model's veracity, but it is unclear if those explanations can help participants make the correct final decision.

Explanations should be informative, instead of just convincing. A followup question is, then, what kinds of interactions would help the humans identify unknown-unknowns, or perform correctly when the AI is incorrect?

Strategy-wise, while our current Adaptive explanations change based on the AI's confidence, it may be possible to design different explanation strategies that adapt based on the *frequency* of agreement between the human and AI. For example, instead of explaining why it believes an answer to be true, the AI might play a devil's advocate role, explaining its doubts — even when it agrees with the human.

Source-wise, our current expert explanations did not help more than the AI explanations, which may indicate that maximally convincing explanations — a common objective shared in the design of many AI explanation algorithms — might be a poor choice for complementary performance. Indeed, while it is known that human explainers and listeners often have different objectives [46], it seems quite problematic to develop AI systems that actively seek to minimize human oversight. A more ideal condition might be to reflect the AI's correctness directly through the explanation, such that the explanation makes sense when the recommendation is correct, but is irrelevant, nonsensical and raises suspicion when it is incorrect. We thought that this might happen with our RoBERTa-based explanations for sentiment analysis, but informal analysis showed that the

AI usually found some plausible justification when it made mistakes. We hope future research can produce explanation algorithms that are more capable of generating informative explanations, which better enable the human to effectively check an AI's reasoning and catch its mistakes.

The timing of AI recommendations is important. Besides the types of explanations, it is also important to carefully design *when* the AI provides its viewpoint. All of our methods used a workflow that showed the AI's prediction (and its explanation) to the human, before they attempted to solve the problem on their own. However, by presenting an answer and accompanying justification upfront, and perhaps overlaid right onto the instance, our design makes it almost impossible for the human to reason independently, ignoring the AI's opinion while considering the task. This approach risks invoking the anchor effect, studied in Psychology [16] — people rely heavily on the first information that is presented by others when making decisions. This effect was reflected in an increase in the use of the "AI as Prior Guide" collaboration approach in the sentiment analysis domains, compared to *LSAT* (Figure 8).

Alternate approaches that present AI recommendations in an asynchronous fashion might increase independence and improve accuracy. For example, pairing the humans with slower AIs (that take more time to make recommendation) may provide the humans with a better chance to reflect on their own decisions [54]. Methods that embody management-science recommendations for avoiding group-think [49] might also be effective, *e.g.*, showing the AI's prediction after the human's initial answer or only having the AI present an explanation if it disagreed with the human's choice. We note that these approaches correspond to the Update and Feedback methods of Green & Chen [21], which *were* effective, albeit not in the complementary zone. However, by delaying display of the AI's input until after the human has solved the task independently, one may preclude improvement to the *speed* of problem-solving, which often correlates to the cost of performing the task (participants self-reported that explanations, in particular the inline-highlights, helped them speed up their reading). There is a strong tension between competing objectives of speed, accuracy, and independence; We see our paper as a call to action, and encourage the field to design and conduct experiments in similar complementary settings, and explore different architectures for balancing these factors.

6 CONCLUSIONS

Previous work has shown that the accuracy of a human-AI team can be improved when the AI explains its suggestions, but these results are only obtained in situations where the AI, operating independently, is better than either the human or the best human-AI team. We ask if AI explanations help achieve *complementary* team performance, i.e. whether the team is more accurate than either the AI or human acting independently. We conducted large-scale experiments on three datasets with more than 1500 participants. Importantly, we selected our study questions to ensure that our AI systems had accuracy comparable to humans and increased the opportunity for seeing complementary performance. We considered multiple explanation strategies: explaining the top prediction, the top two predictions, and a novel adaptive explanation that explains alternative answers if AI's confidence is low. Furthermore, we tested explanations from two sources: those generated by LIME on a state-of-the-art RoBERTa model and those produced by a human expert.

While all human-AI teams showed complementary performance, none of the explanation conditions produced accuracy significantly higher than the simple baseline of showing the AI's confidence. Explanations increased team performance when the system was correct, but they decreased accuracy on examples when the system was wrong. In short, the use of explanations increased the human's blind trust in the AI system, rather than appropriate reliance [39], making them more likely to adopt its recommendation both when the AI was right and when it was wrong. This

is consistent with psychology literature [32], which has shown that human explanations cause listeners to agree even when the explanation is wrong, but to our knowledge this hasn't been shown for AI explanations. While several prior studies' results showed net team improvement from explanations, their AI was so much more accurate than the humans that indiscriminate trust in AI appeared to be desirable. The result also shows the importance of investigating complementary performance.

Our work highlights critical challenges for the HCI and AI communities: characterize when human-AI collaboration can be beneficial (*i.e.*, when both parties complement each other), develop explanation approaches and coordination strategies that result in a complementary team performance that exceeds what can be produced by simply showing AI's confidence, and communicate explanations to increase understanding rather than just to persuade. As a first step, we analyzed the weaknesses of existing explanation representations and developed *adaptive explanations*, which change the representation based on the AI's confidence. While we did not observe significant improvements with these conditions, we saw suggestive evidence, in several datasets, that Adaptive explanation approach has the potential to achieve appropriate reliance [39] — increase trust for high confidence predictions yet encourage autonomy when the AI might be wrong.

Through extensive qualitative analysis, we characterized the connection between improved team performance and the correlation of errors between the human and AI system. For example, in the *Beer* and *LSAT* domains, humans did poorly on problems that the AI answered correctly while many problems that stumped the AI were easy for most humans (Figure 6). In these domains, there was both increased complementarity and more benefit from AI recommendations than *Amzbook*, where human and AI expertise were more correlated. This suggests that the benefit of AI assistance may be much more complex than mere accuracy, and that it may be useful to directly optimize for complementary behavior as suggested by recent work [2, 66]

We hope this paper will serve as a "Call to action" for the HCI and AI communities, as there is much future work remaining to be done. At the highest level, we hope researchers can develop new interaction methods that increase complementary performance beyond having an AI telegraph its confidence. We considered accuracy as the primary performance metric, but we hope to investigate other measures as well, such as time required to complete a task and human satisfaction. Even in the context of improving accuracy, it seems as if there are additional factors that should be investigated, such as the correlation between human and machine errors. Different ways of presenting explanations, *e.g.*, besides our highlights and narrative explanations, may also yield different results. Additionally, one should consider different workflows — our design essentially has humans double-checking the AI's recommendation, the opposite flow might work better. For example, Green and Chen showed benefit to a Feedback condition, where the AI only volunteered its opinion after the human completed the task independently [21]; perhaps this approach offers benefit in the setting of complementary performance? We also hope to develop improved versions of our Adaptive explanation, which may improve accuracy by fostering appropriate reliance.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software across the country to predict future criminals and it's biased against blacks. *ProPublica* (2016).

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. 2020. Optimizing AI for Teamwork. *ArXiv* abs/2004.13102 (2020).

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In *AAAI*.

[4] Richard E Barlow and Hugh D Brunk. 1972. The isotonic regression problem and its dual. *J. Amer. Statist. Assoc.* 67, 337 (1972), 140–147.

[5]   Mohsen Bayati, Mark Braverman, Michael Gillam, Karen M Mack, George Ruiz, Mark S Smith, and Eric Horvitz. 2014.
      Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PloS one* 9, 10
      (2014).
[6]   R. Beer. 1995. A Dynamical Systems Perspective on Agent-Environment interaction. *Artificial Intelligence* 72, 1–2
      (1995), 173–215.
[7]   Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri,
      José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020
      Conference on Fairness, Accountability, and Transparency.* 648–657.
[8]   Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human
      Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. *CoRR* abs/1801.10408 (2018). arXiv:1801.10408
      http://arxiv.org/abs/1801.10408
[9]   Test Prep Books. 2016. *LSAT prep book: study guide & practice test questions for the Law School Admission Councils
      (LSAC) Law School Admission test.* Test Prep Books.
[10] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": Uncovering the
      Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on
      Human-Computer Interaction* 3, CSCW (2019), 1–24.
[11] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. 2015. Intelligible models for healthcare: Predicting
      pneumonia risk and hospital 30-day readmission. In *KDD*.
[12] Hau Chan, Long Tran-Thanh, Bryan Wilder, Eric Rice, Phebe Vayanos, and Milind Tambe. 2018. Utilizing Housing
      Resources for Homeless Youth Through the Lens of Multiple Multi-Dimensional Knapsacks. *Proceedings of the 2018
      AAAI/ACM Conference on AI, Ethics, and Society* (2018).
[13] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations
      make VQA models more predictable to a human? *EMNLP* (2018).
[14] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a
      machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces.*
      329–340.
[15] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace.
      2019. ERASER: A Benchmark to Evaluate Rationalized NLP Models. *arXiv preprint arXiv:1911.03429* (2019).
[16] Birte Englich, Thomas Mussweiler, and Fritz Strack. 2006. Playing dice with criminal sentences: The influence of
      irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin* 32, 2 (2006), 188–200.
[17] Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me: Evaluating Machine Learning Interpretations in
      Cooperative Play. *IUI* (2019).
[18] Reza Ghaeini, Xiaoli Z. Fern, and Prasad Tadepalli. 2018. Interpreting Recurrent and Attention-Based Neural Models: a
      Case Study on Natural Language Inference. *CoRR* abs/1808.03894 (2018). arXiv:1808.03894 http://arxiv.org/abs/1808.
      03894
[19] Shahrzad Gholami, Lily Xu, Sara Mc Carthy, Bistra N. Dilkina, Andrew J. Plumptre, Milind Tambe, Rohit Singh,
      Mustapha Nsubaga, Joshua Mabonga, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Tom Okello, and Eric Enyel.
      2020. Stay Ahead of Poachers: Illegal Wildlife Poaching Prediction and Patrol Planning Under Uncertainty with Field
      Test Evaluations (Short Version). *2020 IEEE 36th International Conference on Data Engineering (ICDE)* (2020), 1898–1901.
[20] Elena L Glassman, Jeremy Scott, Rishabh Singh, Philip J Guo, and Robert C Miller. 2015. OverCode: Visualizing
      variation in student solutions to programming problems at scale. *ACM Transactions on Computer-Human Interaction
      (TOCHI)* 22, 2 (2015), 1–35.
[21] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of
      the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[22] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In
      Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 1321–1330.
[23] M. Hardt, E. Price, and N. Srebro. 2016. Equality of opportunity in supervised learning. In *NIPS*.
[24] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict
      Model Behavior?. In *the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*.
[25] Yugo Hayashi and Kosuke Wakabayashi. 2017. Can AI become reliable source to support human decision making in a
      court scene?. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.*
      195–198.
[26] Dan Hendrycks and Kevin Gimpel. 2016. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in
      Neural Networks. *arXiv preprint arXiv:1610.02136* (2016).
[27] Sungsoo Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices,
      Challenges, and Needs. *CSCW* (2020).

[28] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[29] Naveena Karusala, Jennifer Wilson, Phebe Vayanos, and Eric Rice. 2019. Street-Level Realities of Data Practices in Homeless Services Provision. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 184 (Nov. 2019), 23 pages. https://doi.org/10.1145/3359286

[30] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2019. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *CHI*.

[31] Jackson A. Killian, Bryan Wilder, Amit Sharma, Vinod Choudhary, Bistra N. Dilkina, and Milind Tambe. 2019. Learning to Prescribe Interventions for Tuberculosis Patients Using Digital Adherence Data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019).

[32] Derek J Koehler. 1991. Explanation, imagination, and confidence in judgment. *Psychological bulletin* 110, 3 (1991), 499.

[33] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*.

[34] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[35] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is Chicago deceptive?" Towards Building Model-Driven Tutorials for Humans. In *CHI*.

[36] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. *FAT** (2019).

[37] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[38] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & Explorable Approximations of Black Box Models. *FATML Workshop, KDD* (2017).

[39] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[40] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155* (2016).

[41] Timothy R Levine. 2014. Truth-default theory (TDT) a theory of human deception and deception detection. *Journal of Language and Social Psychology* 33, 4 (2014), 378–392.

[42] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-attentive Sentence Embedding. *ArXiv* abs/1703.03130 (2017).

[43] Zachary C Lipton. 2016. The mythos of model interpretability. In *Machine Learning: Workshop on Human Interpretability in Machine Learning*.

[44] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H. Lin, Xiao Ling, and Daniel S. Weld. 2016. Effective Crowd Annotation for Relation Extraction. In *Proceedings of NAACL and HLT 2016*.

[45] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv* abs/1907.11692 (2019).

[46] T. Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive psychology* 55, 3 (2007), 232–257.

[47] Scott M. Lundberg, Bala G. Nair, Monica S Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor W. Adams, David E. Liston, Daniel King-Wai Low, S L Newman, Jerry J. Kim, and Su-In Lee. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. In *Nature biomedical engineering*. 749–760.

[48] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. 2018. Teaching categories to human learners with visual explanations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3820–3828.

[49] Les Macleod. 2011. Avoiding "groupthink" A manager's challenge. *Nursing management* 42, 10 (2011), 44–48.

[50] Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 1020–1025.

[51] T. Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (February 2018), 1–38.

[52] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *ArXiv* (2018). arXiv:1802.00682.

[53] Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long Papers). 1069–1078.

[54] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–15.

[55] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[56] M. Ribeiro, S. Singh, and C. Guestrin. 2016. Why Should I Trust You?: Explaining the Predictions of any Classifier. In *KDD*.

[57] Philipp Schmidt and Felix Biessmann. 2019. Quantifying Interpretability and Trust in Machine Learning Systems. *AAAI Workshop on Network Interpretability for Deep Learning* (2019).

[58] C Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping Community in the Loop: Understanding Wikipedia Stakeholder Values for Machine Learning-Based Systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[59] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 3319–3328.

[60] Niels van Berkel, Jorge Gonçalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M. Kelly, and Vassilis Kostakos. 2019. Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proceedings of the ACM on Human-Computer Interaction* 3 (2019), 1 – 21.

[61] Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models. *arXiv preprint arXiv:1909.09251* (2019).

[62] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 606–615. https://doi.org/10.18653/v1/D16-1058

[63] Hilde J. P. Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A Human-Grounded Evaluation of SHAP for Alert Processing. *CoRR* abs/1907.03324 (2019). arXiv:1907.03324 http://arxiv.org/abs/1907.03324

[64] Daniel S. Weld and Gagan Bansal. 2019. The Challenge of Crafting Intelligible Intelligence. *Commun. ACM* 62, 6 (May 2019), 70–79. https://doi.org/10.1145/3282486

[65] Jenna Wiens, John Guttag, and Eric Horvitz. 2016. Patient risk stratification with time-varying parameters: a multitask learning approach. *JMLR* 17, 1 (2016), 2797–2819.

[66] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to Complement Humans. (2020).

[67] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[68] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I trust my machine teammate? an investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 460–468.

[69] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. *arXiv preprint arXiv:2002.04326* (2020).

[70] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. *arXiv preprint arXiv:2001.02114* (2020).
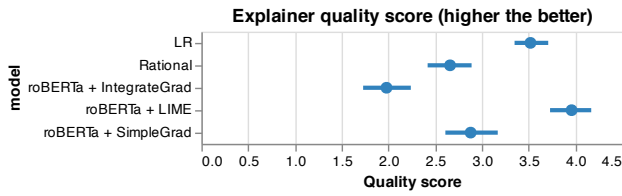
Fig. 10. Among the many approaches we tried, roBERTA+LIME was consistently ranked higher than others. **roBERTa+LIME** appeared to be most highly scored (3.96 ± 1.2).

A APPENDIX

A.1 Choosing a Classification Model & Explanation Method

While various studies have been conducted to compare explanations from different aspects (*e.g.*. [24, 53]), there has yet to be a clear conclusion on which models/explainers perform the best. For our studies, we wished to select a model that offered both credible predictions and explanations. So, we compared three approaches, which broadly covered existing NLP models for generating explanations along with predictions [43, 53]: an intelligible model (*Logistic Regression* classifier), a encoder-decoder *rationale generator* (Lei *et al.* [40]), and a recent contextual embedding-based model (*RoBERTa* with linear classification layer). While the first two models inherently generate explanations by providing access to weight vectors which represent salience of input words for predictions, the third method is a black-box and requires using a *post-hoc explainer* to generate explanations. We tried three existing post-hoc explainers to explain the RoBERTa model: *LIME* trains a new intelligible model to provide a local approximation of a model's behavior[56], *SimpleGrad* computes gradients with respect to input, and *IntergratedGrad* uses an approach that aggregates multiple gradients [59].[6]

We conducted the following small study on *Beer* to select a model for both datasets: We selected ten beer reviews (five positive and negative) on which all three models were correct and analyzed explanations for the predicted class (highlighted similar to our main task). Since we experimented with three explainers for RoBERTa, we analyzed five explanations per example. Then, we asked ten graduate students in computer science to rank the explainers from the best quality (score 5) to the worst (score 1). Specifically, we asked them to *"rank the explainers by quality, i.e., which explainer picks the most reasonable words associated with predicted positive/negative sentiment."* All explanations were anonymized and randomly ordered.

For our main task, we selected the setting that received the highest average score across participants and reviews, which was *roBERTa+LIME* (Figure 10).

A.2 Stratified Sampling Using Classifier's Confidence

Despite performing calibration, our *roBERTa* model produced a highly skewed distribution of confidence scores (Figure 10A). For example, a large number of test examples had a confidence score greater than 0.99, which is problematic for our studies. If all (or most) samples have a confidence score greater than 0.99, then showing confidence scores to users is essentially uninformative. Since a uniformly randomly drawn study sample would inherit a similar skew, we performed stratified sampling.

---

[6]Implementation: For logistic regression we used scikit-learn, for the rational generator we used re-implementaion by [15], and, we used AllenNLP's demo code for classification using RoBERTa https://demo.allennlp.org/sentiment-analysis. For the post-hoc explainers, we used LIME's original implementation [56] and AllenNLP Interpret [61] for gradient-based explainers.
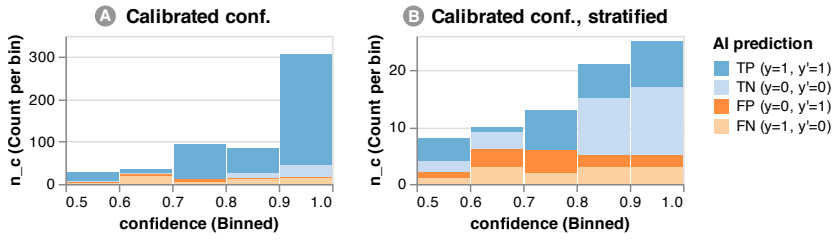
Fig. 11. A comparison of the confidence distribution, before and after stratified sampling. The sampling helped select examples to achieve a controlled confidence distribution needed for our analysis.

To select examples with a wide range of confidence scores, we first sampled examples such that $0.5 \leq$ confidence < 0.99, and binned them with a step size of 0.1 (Figure 11A). We then performed stratified sampling to draw examples from sub-populations defined by a confidence bin and category of AI prediction: True Positive, True Negative, False Positive, and False Negative. Since drawing equal number of samples for each sub-population may be too unnatural, for each confidence bin in a prediction category, we drew samples proportional to the logarithmic of number of samples in that bin. Let $n_{c,p}$ denote number of samples in confidence bin c with prediction category p. We drew $m_{c,p}$ samples such that:

$$m_{c,p} = \frac{\ln(n_{c,p})}{\sum_{c'}(\ln(n_{c',p}))} \cdot n_{*,p}$$

As shown in Figure 11A, our sampling helped to construct a subset with a reasonable number of examples per bin, with different AI predictions.