Selective Explanations: Leveraging Human Input to Align Explainable AI

VIVIAN LAI[*], University of Colorado Boulder, USA
YIMING ZHANG[*], University of Chicago, USA
CHACHA CHEN, University of Chicago, USA
Q. VERA LIAO, Microsoft Research, Canada
CHENHAO TAN, University of Chicago, USA

While a vast collection of explainable AI (XAI) algorithms has been developed in recent years, they have been criticized for significant gaps with how humans produce and consume explanations. As a result, current XAI techniques are often found to be hard to use and lack effectiveness. In this work, we attempt to close these gaps by making AI explanations *selective*—a fundamental property of human explanations—by selectively presenting a subset of model reasoning based on what aligns with the recipient's preferences. We propose a general framework for generating selective explanations by leveraging human input on a small dataset. This framework opens up a rich design space that accounts for different selectivity goals, types of input, and more. As a showcase, we use a decision-support task to explore selective explanations based on what the decision-maker would consider relevant to the decision task. We conducted two experimental studies to examine three paradigms based on our proposed framework: in Study 1, we ask the participants to provide critique-based or open-ended input to generate selective explanations (self-input). In Study 2, we show the participants selective explanations based on input from a panel of similar users (annotator input). Our experiments demonstrate the promise of selective explanations in reducing over-reliance on AI and improving collaborative decision making and subjective perceptions of the AI system, but also paint a nuanced picture that attributes some of these positive effects to the opportunity to provide one's own input to augment AI explanations. Overall, our work proposes a novel XAI framework inspired by human communication behaviors and demonstrates its potential to encourage future work to make AI explanations more human-compatible.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; • **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → **Law, social and behavioral sciences**.

ACM Reference Format:
Vivian Lai, Yiming Zhang*, Chacha Chen, Q. Vera Liao, and Chenhao Tan. 2023. Selective Explanations: Leveraging Human Input to Align Explainable AI. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 357 (October 2023), 35 pages. https://doi.org/10.1145/3610206

1 INTRODUCTION

With advances and widespread adoption of artificial intelligence (AI) systems, the need for people to understand AI in order to appropriately trust and effectively interact with AI has spurred great interest in the emergent field of explainable AI (XAI) [27, 38, 43]. The technical field of XAI has

[*] denote equal contribution

Authors' addresses: Vivian Lai, vivian.lai@colorado.edu, University of Colorado Boulder, Boulder, CO, USA; Yiming Zhang*, yimingz0@uchicago.edu, University of Chicago, Chicago, IL, USA; Chacha Chen, chacha@uchicago.edu, University of Chicago, Chicago, IL, USA; Q. Vera Liao, veraliao@microsoft.com, Microsoft Research, Montreal, Canada; Chenhao Tan, chenhao@uchicago.edu, University of Chicago, Chicago, IL, USA.

This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2023 Copyright held by the owner/author(s).
2573-0142/2023/10-ART357
https://doi.org/10.1145/3610206

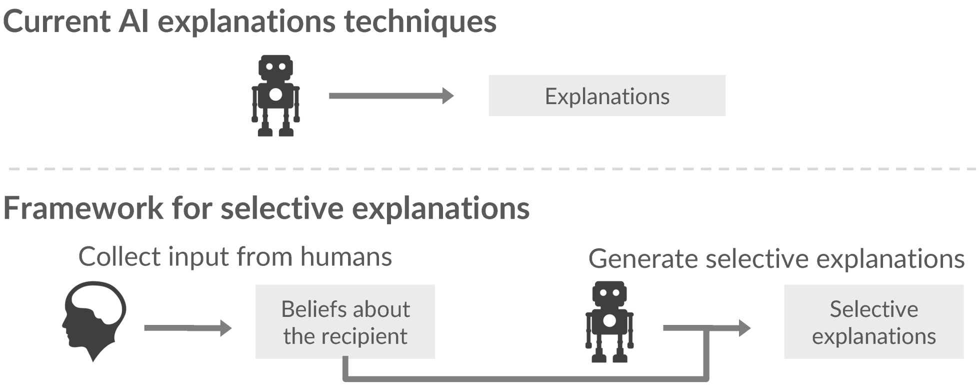

Current AI explanations techniques
Explanations
Framework for selective explanations
Collect input from humans
Beliefs about the recipient
Generate selective explanations
Selective explanations

Fig. 1. A high-level overview of the framework. See the full framework in Figure 2 and a detailed discussion in Section 3.

made remarkable progress in recent years, producing a large collection of algorithms that aims to reveal the decision processes of machine learning (ML) models [5, 42]. However, empirical human-subject studies that examine how people interact with state-of-the-art XAI techniques have not found conclusive evidence that these techniques help end-users better complete AI-assisted tasks [6, 11, 13, 41, 56, 57]. These AI explanations are often unintuitive and demand significant effort for people to process and understand [12, 34, 55]. They have been found to risk impairing task performance [6, 40, 41, 57, 81], efficiency [1, 17, 34, 55, 99], and user satisfaction [24, 55, 68], ultimately preventing users from harnessing reals benefits due to explanations [12, 35, 93].

This difficulty to use current XAI techniques can be attributed to their lack of compatibility with how humans produce and consume explanations, as pointed out by social sciences literature [62, 74, 96]. For example, Malle's theory of explanation [71] describes that a human explainer must engage in two fundamental processes to produce explanations—an process to *gather* all reasons that can explain, and an impression management process to *communicate* the explanation in social interactions. Arguably, by solely focusing on revealing the model decision processes, current XAI paradigms deal only with the *reasoning* process and concern little with the *communication* process.

How do people engage in explanation communication? Among other characteristics, people rarely present all explanatory causes but *select* what they believe as serving the recipient's interest for achieving their goal, such as finding common grounds and providing new knowledge (more on selectivity goals to be discussed in Section 2.2). That is, at the core of effective communication of explanation is *explanation selection* based on the explainer's goal and their beliefs about the recipient.

Inspired by this *selective* property of human explanation, we introduce a novel framework to selectively present AI explanations based on beliefs about the recipient's preferences. This framework can be used to augment any existing feature-based local explanations—XAI techniques that explain a particular model prediction by how the model weighs different features of the instance, with potential of extending to other XAI techniques discussed in Section 7. Figure 1 gives a high-level illustration of our framework (see the full version of our framework in Figure 2 and a detailed discussion in Section 3). On a high level, our framework consists of two steps: (1) collecting input from humans on a small sample and (2) generating selective explanations according to beliefs about the recipient's preferences for explanation, as inferred from the human input collected in step 1.

In Section 4, we present a way to instantiate this framework in a text classification task. Using this instantiation, we empirically explore the effects of selective explanations in an AI-assisted review sentiment judgment task as a testbed. Similar to people achieving better impression management with explanation selection, we expect selective explanations to improve users' perception of AI explanations. Meanwhile, a curious question is whether selective explanations can help people make better decisions. Rather than "enhancing trust", the field of XAI is shifting its attention to "calibrating trust" [108], arguing that a more desirable goal of AI explanation should be to help people discern correct and incorrect model predictions to have more appropriate reliance on the model and thus more accurate human-AI joint decisions. We note that such a goal—facilitating the recipient to detect flaws of the explainer—is rarely the focus of human explanations, suggesting a possible tension with providing "human-like explanations" by AI.

In Section 5 and 6, we explore these questions through two controlled human-subjects experiments (N=118, N=161) where we test three paradigms (from a broader space) of selective explanations based on our proposed framework. In Study 1, we ask participants to provide their own input to generate selective explanations (self-input), either with *Open-ended* (selecting any features as aligning with their preferences) or *Critique-based* (critiquing AI's explanation) feedback, and compare their effects to a baseline condition with non-selective explanations. In Study 2, we show participants selective explanations that are generated based on input *from a panel of similar users* (annotator input).

Results from these experiments demonstrate the promise of selective explanations. We found evidence that selective explanations were better aligned with the decision ground truth, improved decision outcomes, and decreased over-reliance when the AI predictions were wrong. They also consistently improved people's perceived understanding of the model over unselected explanations. Interestingly, in self-input paradigms, the opportunity to provide one's own input and have control over AI explanation also improved the perceived usefulness of AI, albeit at the cost of the increased overall workload.

In summary, our main contributions can be summarized as follows:

- We propose a novel conceptual framework for generating selective explanations by leveraging human input and laying out the rich design space. Our work aligns with human-centered XAI efforts [32, 62, 101] by providing a concrete way to operationalize human-like explanation communication behaviors that can be broadly applied to augment existing XAI techniques.
- We instantiate the framework in text classification and develop the corresponding algorithms and interface.
- We conduct two controlled experiments and demonstrate the promise of selective explanations in improving decision outcomes and subjective perceptions of AI.

In the rest of the paper, we first review related work that informed our research, then provide an overview of our framework. Then we instantiate the framework and present the two experiments exploring the effects of selective explanations. In Section 7, we reflect on the results to discuss lessons learned, generalizability, and future directions, as well as open questions for our framework.

2 RELATED WORK

2.1 Explainable AI and Its Pitfalls

Recent years have seen a booming interest in explainable AI (XAI) [43], thanks to the unprecedented popularity of "black-box" AI models that are built on complex algorithms and architectures such as deep neural networks. Among the growing collection of XAI techniques (as surveyed in [2, 5, 18, 38, 42]), we focus on those explaining deep machine learning classifiers (as opposed to other types of AI systems such as planning or multi-agent systems). "Local" XAI techniques that explain a model prediction (as opposed to "global" explanations to describe the entire model) can be roughly categorized into feature-based, example-based, and counterfactual explanations [42], with feature-based explanations being the most popular approach and the focus of this work. In short, feature-based explanations describe how the model weighs different features of the input instance to arrive at its prediction, often by highlighting the most salient features. As a general form of explanation, feature-based explanations can be generated by many different algorithms that vary in computational properties, such as LIME [83] and SHAP [69].

Because explainable AI is fundamentally about supporting human understanding of models, the broad XAI community has been pushing for human-centered approaches [29, 32, 62, 101] that consider people's needs and preferences, as well as study how people actually interact with AI explanations. One line of such work focuses on summarizing common use cases of or objectives people have with AI explanations [5, 20, 63, 91], including supporting verifying and debugging models, assisting decision-making, auditing model (e.g., on bias, privacy and security issues), and knowledge discovery. Meanwhile, many HCI and CSCW researchers have explored developing XAI applications in various domains (e.g. [49, 50, 102]), and conducting empirical studies to investigate the effects of explanations on people's task performance [6, 57, 108], efficiency [1, 17, 21, 34, 36, 39, 54, 55, 59, 64, 88, 99, 103], cognitive load [1, 37], understanding [4, 8, 11, 14, 21, 68, 88, 98, 103], subjective perceptions of AI [9, 15, 25, 37, 54, 55, 68, 75, 92] among others.

Unfortunately, results from these recent empirical studies of AI explanations are mixed at best. On the one hand, many studies found positive evidence that explanations improve people's understanding of the model [21, 58, 64, 84], enhance people's subjective perception of and tendency to follow AI [78], help data scientists debug the model [49, 76], and auditors detect model biases [26]. On the other hand, multiple studies reported that end-users found the explanations generated from popular technical approaches hard to use, distracting, time-consuming, and cognitively demanding [50, 60, 85, 89, 102]. Due to the added cognitive load, studies also found that showing explanations reduce task satisfaction for people with a low "need for cognition" trait [12, 37] (not enjoying cognitively demanding activities). These surprisingly negative effects of explanations from empirical studies have been referred to as XAI pitfalls [30, 62]

In particular, recent studies begin to call out a prominent XAI pitfall—increasing people's over-reliance when the AI is wrong, which is especially problematic in the common use case of XAI for decision support. While the expectation is that explanations can help people detect flawed model reasoning and make better decisions, empirical studies either failed to observe this effect [98] or even found the opposite that explanations make people more likely to blindly follow the model when it is wrong compared to showing only AI predictions [6, 82, 98, 108]. Research has attributed this phenomenon to a lack of cognitive engagement with AI explanations [11, 35, 52, 62]: when people lack either the motivation or ability to carefully analyze and reason about explanations, they make a heuristic judgment, which tends to superficially associate being explainable to being trustworthy [28, 61]. A recent CSCW work by Vasconcelos et al. [93] further calls out that this lack of cognitive engagement will persist if XAI techniques remain hard to use, as people strategically choose between engaging with explanations and simply deferring to AI after weighing the cognitive costs.

Motivated by these prior works, we aim to make AI explanations more human-compatible by making them easier to use and thereby tackling these XAI pitfalls. To explore the benefits of the proposed approach, as informed by prior empirical studies of XAI for decision-support, we will measure participants' decision outcomes, reliance on AI, efficiency, subjective cognitive load, understanding and perceived usefulness of the AI.

2.2 Making Explainable AI Human-Compatible: The Case for Selectivity

While HCI researchers have taken various efforts to design XAI systems that are more user-friendly [50, 102], less cognitively demanding [1], or nudge people to better engage with explanations [12], current XAI techniques' difficulty to use can be fundamentally attributed to their disconnect with how humans produce and consume explanations [62, 74, 97]. Such criticism is best reflected in Miller's work [74] that brings insights about human explanations from social sciences literature and argues that XAI should be built with human explanation properties in mind. Miller summarizes three fundamental properties of human explanations: contrastive (against counterfactual scenarios), selective, and social (as part of social interactions). This work has since inspired many new techniques aiming to make AI explanations more human-compatible, such as counterfactual [94, 95] and weight-of-evidence explanations [3] that cater to the contrastive property, and various kinds of interactive explanations [87, 107] inspired by the social property.

Our work is directly inspired by the selective property, which Miller points out as missing from current XAI techniques. As discussed in Section 1, human explanations are often selected for social and cognitive reasons [46, 71], as the complete reasoning or causal chains are often too large to comprehend (e.g. the causes of a fatal car accident can be explained by a chain of a few dozen of events). There has been a line of psychology work arguing explanation selection is not arbitrary but follows common criteria [45]. For example, Hilton and Slugoski [48] demonstrate that abnormal (unusual or rare events) factors or events are more often presented in explanations while commonplace knowledge is often omitted (e.g., an unexpected lane change versus driving at 75 mph on a highway). Hilton and John [47] show that intentional actions that are deliberately changed (e.g., the driver is drunk), and relatedly, controllable events [72] that can be changed with intentions, tend to take priority in explanations. Many also suggest that people prioritize the most important or relevant reasons, which can be matters of necessity, sufficiency, or robustness in causal reasoning [65, 100]. Our proposed framework is directly inspired by this body of literature on how humans selectively present explanations, with some of of the most common criteria being relevance [65, 100], abnormality [48], and changeability [47, 72].

2.3 Learning from Human Input on Explanations

Our work is also informed by a small but growing set of works on eliciting human input relevant to AI explanations. Prior work explored eliciting human feedback on model explanations as additional supervision signals for model training [16, 37, 90]. For example, Ghai et al. [37] proposed explainable active learning where labelers are asked to not only provide labels to train the model but also critique feature-based explanations produced by the learning model. Other works proposed new XAI techniques by eliciting human's own rationales (e.g. which keywords are important or what rules to follow to reach decisions) [31] or domain concepts [53] to help generate or improve AI explanations. For example, Ehsan et al. [31] propose to train an explanation generation model directly from elicited human rationale data to help lay users make sense of model actions. Another relevant work by Feng and Boyd-Graber [33] trains a model to select different combination of explantions to accommodate different users' needs and preferences. However, we note that "selectivity" in this paper is about selection from multiple explanation sources using user feedback, rather than the selectivity demonstrated in human explanation communication, which is the focus of our work.

Instead of proposing a new XAI algorithm, we propose a novel framework that can be broadly applied to *augment* the outputs generated by any existing feature-based XAI algorithm. Closest to our work is a recent study by Boggust et al. [10]. To help people better and more efficiently analyze model behaviors, they propose a set of metrics to contrast model reasoning via the saliency method (a feature-based explanation for image data) and human reasoning gathered from annotations.

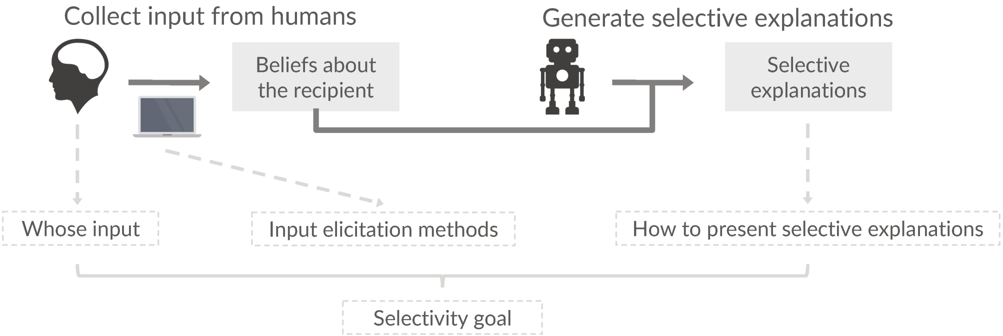

Collect input from humans
Beliefs about the recipient
Generate selective explanations
Selective explanations
Whose input
Input elicitation methods
How to present selective explanations
Selectivity goal

Fig. 2. Illustration of the design space in our framework. The dashed-line boxes highlight dimensions of design considerations.

To name a few, the "human aligned" metric measures how often human and model reasoning are consistent, and the "sufficient subset" metric measures the degree to which model rationale contains human rationale. These metrics can then be used to rank and sort a large number of data instances, helping people identify and analyze different patterns of model behavior. Our work is inspired by this general approach of contrasting raw outputs generated by XAI algorithms and human rationale, but we leverage the latter to augment the former and lay out the design space by considering different goals, as well as ways to elicit human rationale, and present selective explanations. To measure the effectiveness of our proposed framework, we investigate the effect of selective explanations through controlled human-subject experiments.

3 FRAMEWORK: SELECTIVE EXPLANATION WITH HUMAN INPUT

Inspired by the selective property of human explanation, we propose a general framework for generating selective AI explanations by leveraging human input. Our framework consists of two steps: the input step and the selection step.

In the **input step**, the goal is to elicit human input that can be used to infer beliefs about the user, such as which features the user would consider relevant to the decision task. Ideally, this step should be efficient and requires only a small sample set to elicit human input.

In the **selection step**, a separate prediction model, which we call *belief prediction model* hereafter, is used to generalize from input gathered in the first step to predict the recipient's preferences regarding the given instance of explanation, then selectively augment the original explanation by prioritizing features that align with the predicted preference. At the heart of our framework is an algorithm of this belief prediction model, for which we propose a simple yet effective approach in Section 4.

We start by discussing the design choices under our framework (see Figure 2 for an illustration), and create an instantiation in Section 4 using a subset of this design space to conduct empirical studies. While we limit our user studies to local feature-based explanations, we will discuss algorithmic considerations for generalizing beyond this particular instantiation and feature-based explanations in Section 7.

Our entire framework is contingent on the *selectivity goal* of the model, which may vary across XAI use cases. In the input step, the input can be obtained from *different kinds of stakeholder group* with different *elicitation methods*. In the selection step, once the selective explanations are generated, the key question is *how to present* them visually. Below we elaborate on these dimensions of design choices. Our goal is to explore this design space and layout possibilities for future work to utilize this framework, rather than to make conclusive recommendations. In Table 1, we list examples of how to make choices in these design dimensions (excluding the *how to present* dimension) for popular XAI use cases discussed in the literature [5, 20, 63, 91]. We will refer to this table throughout the discussions below.

• **Selectivity goal**: Drawing from social science literature on common criteria based on which humans produce selected explanations (reviewed in Section 2.2), we suggest three general selectivity goals that can appear in different XAI applications: relevance, abnormality, and changeability—future work can further expand these goals. In Table 1, we list examples of XAI use cases where each selectivity goal is appropriate.

- *Relevance* prioritizes reasons that the recipient would deem relevant or important to the task. In different XAI use cases, relevance may have specific meanings. For example, for a decision-support AI that helps people detect review sentiment, the relevance goal would prioritize presenting features (i.e., words) relevant for judging the sentiment. When applying XAI for auditing model biases, relevance would focus on features related to protected attributes such as race and gender (including potentially correlated features, e.g., zip code). By producing explanations that are more concise and relevant to what the recipient is looking for, we hypothesize that the relevance goal can help the recipient discover useful information more easily, improve the intuitiveness, ease of use, and overall perception of explanation.

- *Abnormality* prioritizes reasons that the recipient would find abnormal or surprising. For example, when applying explanations to debug why the model makes certain mistakes, the abnormality goal could highlight features that the model unexpectedly (for the person doing debugging) picked up in its decision process to help people detect model abnormality more accurately and efficiently. In XAI use cases for knowledge discovery (e.g., supporting data analysts), if the model process is verifiably correct, this abnormality goal could be used to help people learn new knowledge such as identifying contributing factors that are unknown to the user. Note that in some use cases (e.g., learning new knowledge about judging review sentiment), the abnormality goal can be seen as the reverse of the relevance goal—while the latter selectively prioritizes reasons that align with the recipient's rationale, the former prioritizes reasons that do not align with human intuition but are nevertheless useful.

- *Changeablity* prioritizes reasons that can be changed or are more easily changeable. This goal is especially helpful for XAI use cases where explanations are sought for recourse [51]—taking actions that can result in a different, often more desirable prediction in the future. For example, if an applicant's loan application is rejected due to an algorithmic risk assessment tool, an explanation should prioritize features that they can take action to change (e.g., reducing frequency of credit inquiry) and de-emphasize what they cannot easily change (e.g., significantly increasing income). While counterfactual explanations [94, 95], which automatically search for features that with minimum change can alter the prediction, are often proposed to support recourse, existing techniques do not consider the changeability of the features shown and thus have been criticized for lacking actionability [7, 94].

• **Whose input:** Another key design dimension under this framework is from whom to elicit the input in the first step. In a most straightforward form, the input can come from the *individual recipient* who will receive the explanations (Table 1 gives specific examples of who the individual recipient is according to the XAI use case). However, this approach creates additional workload and requires time and resources that not every individual can afford. Alternatively, one may assume there is shared preferences for a task, and collect human input via *a panel of annotators similar to the target users* and apply their input to generative selective explanations for all. In some situations, individuals may lack the domain knowledge to effectively articulate what is relevant or abnormal. One may choose to gather input from an "ideal user archetype", such as *domain experts*, and use the input to improve the experience for all. Importantly, different choices of “whose input” can introduce different effects and even biases, and must be carefully tested and justified for a specific XAI use case. We will empirically explore the differences between eliciting input from individual recipients versus a panel of annotators, and further reflect on this design dimension in Section 7.

• **Input elicitation method:** Once the selectivity goal is determined for a specific XAI use case, the elicitation asks the human input provider "which features should be considered as relevant/abnormal/changeable for this use case? " While it is possible to ask such a question in the absence of context, the knowledge elicitation literature [22] suggests that people are often better at articulating their knowledge or opinions with examples. Therefore, we suggest eliciting human input based on a small sample of examples. Example-based input can be *open-ended*—asking directly to pick features from the example, or *critique-based*—asking for agreement or disagreement with AI explanations for the given example. In Table 1, we list example questions to ask for specific XAI use cases and selectivity goals, focusing on the open-ended feedback (critique-based feedback would simply require pointing to the model explanation, e.g., "which features *in the model explanation* are relevant"). We generally recommend lower-precision input as natural human rationales are often qualitative [74]. That is, the elicitation could ask the person to select relevant features or rank feature by their relevance, as opposed to specifying precisely how relevant each feature is.

A contingent design decision here is the *sampling strategy* to select examples to get the input. While a simple strategy could be random sampling or sampling examples with high-coverage features (i.e., shared by many instances), there exist more sophisticated strategies that depend on the selectivity goal (e.g., searching for examples with possible abnormalities). Furthermore, many design decisions can be made at the elicitation interface level, such as the modality (e.g., graphic vs. conversational interface) and language styles.

• **How to present selective explanation:** In the last step, once the selective explanation can be generated, one needs to decide how to present it visually. This decision depends on UI characteristics, user preferences, and the holistic system user experience. We can only propose a few possibilities. We start by considering popular UI designs for non-selected feature-based explanations: for text or image data, saliency map is often used to visually highlight important keywords or superpixels (perceptual grouping of pixels). For tabular data, a horizontal bar chart is often used to visualize the importance of different features in the given instance. One possibility is to only present features that align with the predicted recipient preferences and hide presenting misaligned features. While this approach can produce explanations that are the most lightweight visually, it comes with a *tradeoff of faithfulness*—losing information about how the model actually works. To mitigate this tradeoff, an alternative is to still maintain the presence of misaligned features, but augment them with different visual cues, such as by graying out or using an underlying waveline. For cases where faithfulness is critical (e.g., debugging [86]), one can preserve the original explanations and add additional highlights to the aligned parts.

4 INSTANTIATING THE FRAMEWORK: PREDICTING AND EXPLAINING MOVIE REVIEW SENTIMENT

Building on our proposed framework, we develop a testbed in the context of AI-supported sentiment judgment. We first discuss the task, model, and base explanations, then the decision choices we made for each design dimension of our proposed framework and how selective explanations are generated. We will use this instantiation to conduct two empirical studies described in Section 5 and 6.

Task, model, and explanations. We choose a sentiment analysis model as our testbed because it is one of the most studied problems in classification [77]. In addition, prior work has shown that

Example XAI Use Case	Selectivity Goal and Benefit	Whose Input (individual recipient)	Example Questions for Input Elicitation (open-ended)
AI assisting consumers to detect review sentiment	Relevance: improve ease of use	Consumers	Which words are relevant for judging the example review's sentiment?
AI assisting loan officers to assess loan application risk	Relevance: improve ease of use	Loan officers	Rank the features by their importance for assessing the example applicant's risk.
Audit model biases in recidivism prediction	Relevance: improve ease of use	Auditors	What features are relevant for making unfair predictions (e.g. protected attributes)?
Debug classification models	Abnormality: help detect model errors accurately and efficiently	Machine learning engineers	Which features should the model NOT base its decisions on?
Assist knowledge discovery for sales analysts	Abnormality: help detect unknown patterns efficiently	Analysts	Which features are less familiar for you to know how they may predict the sales outcomes?
Support recourse for loan applicants	Changeability: facilitate actionable changes	Loan applicants	Which features are possible/require less effort for you to make changes on?

Table 1. Illustration of design choices made with our framework for common XAI use cases. All the columns should be taken as examples instead of best practices.

explanations can increase over-reliance on AI when it is wrong even in this relatively simple task to humans [6].

We train a movie sentiment prediction model using a dataset of IMDb movie reviews (**IMDb**) [70]. Maas et al. [70] collected a balanced set of 50,000 reviews, where negative reviews have scores ≤ 4 and positive reviews have scores ≥ 6. We randomly sampled without replacement to obtain three subsets: a training set of 200 examples, a development set of 500 examples, and a test set of 500 examples. Because sentiment analysis is a relatively easy classification task (state-of-the-art models can achieve an accuracy of almost 95% [105]), we intentionally used a small training set so that the model would perform less than perfectly. This set-up would require people to make more careful judgments with each AI prediction and allow us to study human-AI collaborative decision-making. Specifically, we use a BERT [23] model (**bert-base-uncased**) as the backbone architecture. Following the standard practice, we fine-tune a linear layer on top of the language model with a learning rate of 5×10^{-5} and a batch size of 128 for 200 steps. The fine-tuned model achieves an accuracy of 85.2% on the IMDb test set.

We use feature-based explanations by highlighting important words contributing to the model prediction for text classification. We apply LIME [83], a popular post-hoc XAI algorithm, to generate the importance scores for each word—measuring the degree to which it contributes (positively or negatively) to the model's prediction. LIME estimates this importance score by fitting a sparse linear bag-of-words model to locally approximate the BERT model. Then, we take the unique words with top-10 importance scores as the keywords explanation set for *why* the instance gets a particular prediction (every occurrence of a word is highlighted). Note this explanation set could include both positive and negative keywords. For example, a movie review may have 8 keywords with positive weights for positive sentiment, and 2 keywords with negative weights. The fact that the majority of keywords are positive explains why the review is predicted to be positive. We visually present the explanations with saliency highlights: as shown in Figure 3, we highlight the keywords, with colors indicating the direction of the weights (blue for positive sentiment, red for negative sentiment), and shades indicating the importance of the features (e.g., dark red means the word strongly contribute to a prediction of negative sentiment).

Design choices in instantiating the framework of selective explanations. We make the following design choices to study in our empirical studies out of a larger possible set of choices based on our proposed framework. We will reflect on these choices and discuss alternatives in Section 7.

• **Selectivity goal:** We focus on the goal of *relevance* for an AI-assisted decision-making task since the explanations are expected to help the decision-makers discover relevant information and should be intuitive to use. That is, the selective explanation should prioritize presenting words (from its original explanation) that the recipient would consider relevant for judging movie sentiment.

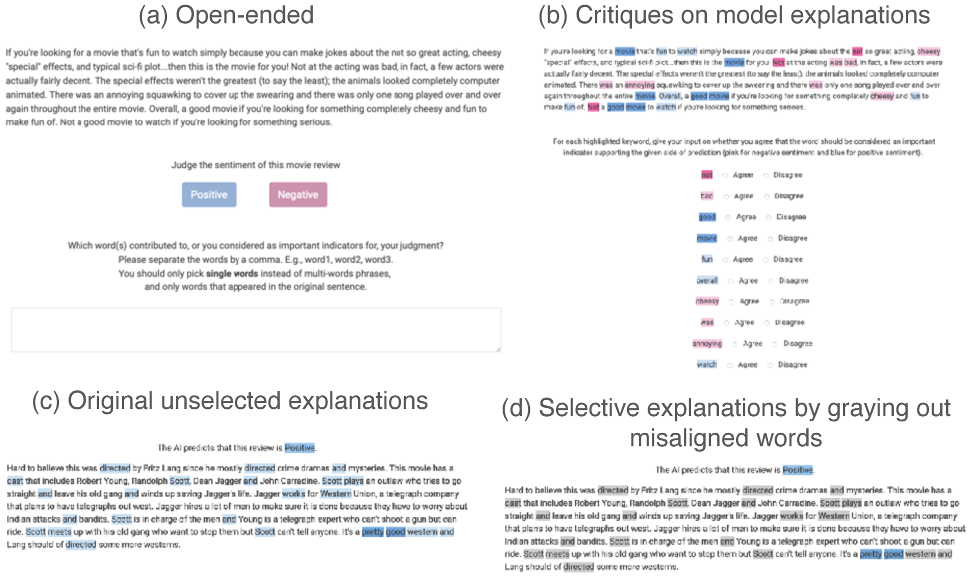
• **Whose input:** We choose to study two possible scenarios to empirically explore the effects of different design choices in this dimension. In Study 1, we ask each *individual user* to provide input and the selective explanations are thus personalized. In Study 2, we obtain input from a *panel of similar users* so the selective explanation is fixed for all participants for a given input.

• **Input elicitation methods:** We choose two elicitation methods to be compared in Study 1: *open-ended* and *critique-based* input. Specifically, for the input phase, we present a sample of movie reviews to people and ask them to provide input for each review. For open-ended input, as shown in Figure 3a, we show people the sample and ask them to pick words that they find as important indicators for them to judge the review sentiment. For critique-based input, as shown in Figure 3b, we show the model's explanation (highlighted keywords) and ask people to critique each word's importance (agree/disagree). While the first approach can be more effortful, it can possibly obtain input for a broader set of words not limited to what is highlighted in model explanations.

For the sampling method, we aim to select reviews where their important words show up in many other instances, which would allow good coverage of user preference information. Proposed by Ribeiro et al. [83] as an application of LIME, SP-LIME is an example selection algorithm that selects representative instances of a data distribution. SP-LIME greedily selects examples that maximize the weight of features they contain, after omitting duplicate features. The weight of each word is defined as the square root of the total sum of its importance across the training dataset. Using SP-LIME, we select the top 10 examples in the development set as the sample for the input step.

• **How to present:** as illustrated in Figure 3d, once the selective explanation is generated for an instance—which features in the raw explanations would be considered relevant or not—we *gray out* irrelevant keywords. This presentation allows de-emphasizing irrelevant keywords but still maintains information about which features carried weight in the AI's prediction.

Belief prediction model: How to generalize from the input to generate selective explanations. Going from the input step to the selection step requires developing computational algorithms to predict what features would be considered relevant by the user in unseen instances. To develop a model to predict such user beliefs, we use elicited user feedback on which words are relevant or not for sentiment judgment from the input stage as labels to train a word-level logistic regression model [79], which we refer to as the "*belief prediction model*". In the task phase, this belief prediction



(a) Open-ended

If you're looking for a movie that's fun to watch simply because you can make jokes about the not so great acting, cheesy "special" effects, and typical sci-fi plot...then this is the movie for you! Not at the acting was bad, in fact, a few actors were actually fairly decent. The special effects weren't the greatest (to say the least); the animals looked completely computer animated. There was an annoying squawking to cover up the swearing and there was only one song played over and over again throughout the entire movie. Overall, a good movie if you're looking for something completely cheesy and fun to make fun of. Not a good movie to watch if you're looking for something serious.

Judge the sentiment of this movie review

Which word(s) contributed to, or you considered as important indicators for, your judgment?
Please separate the words by a comma. E.g., word1, word2, word3.
You should only pick **single words** instead of multi-words phrases,
and only words that appeared in the original sentence.

(b) Critiques on model explanations

If you're looking for a movie that's fun to watch simply because you can make jokes about the not so great acting, cheesy "special" effects, and typical sci-fi plot...then this is the movie for you! Not at the acting was bad, in fact, a few actors were actually fairly decent. The special effects weren't the greatest (to say the least); the animals looked completely computer animated. There was an annoying squawking to cover up the swearing and there was only one song played over and over again throughout the entire movie. Overall, a good movie if you're looking for something completely cheesy and fun to make fun of. Not a good movie to watch if you're looking for something serious.

For each highlighted keyword, give your input on whether you agree that the word should be considered an important indicator supporting the given side of prediction (pink for negative sentiment and blue for positive sentiment).

not ○ Agree ○ Disagree
bad ○ Agree ○ Disagree
good ○ Agree ○ Disagree
movie ○ Agree ○ Disagree
fun ○ Agree ○ Disagree
overall ○ Agree ○ Disagree
cheesy ○ Agree ○ Disagree
was ○ Agree ○ Disagree
annoying ○ Agree ○ Disagree
watch ○ Agree ○ Disagree

(c) Original unselected explanations

The AI predicts that this review is Positive.

Hard to believe this was directed by Fritz Lang since he mostly directed crime dramas and mysteries. This movie has a cast that includes Robert Young, Randolph Scott, Dean Jagger and John Carradine. Scott plays an outlaw who tries to go straight and leave his old gang and winds up saving Jagger's life. Jagger works for Western Union, a telegraph company that plans to have telegraphs out west. Jagger hires a lot of men to make sure it is done because they have to worry about Indian attacks and bandits. Scott is in charge of the men and Young is a telegraph expert who can't shoot a gun but can ride. Scott meets up with his old gang who want to stop them but Scott can't tell anyone. It's a pretty good western and Lang should of directed some more westerns.

(d) Selective explanations by graying out misaligned words

The AI predicts that this review is Positive.

Hard to believe this was directed by Fritz Lang since he mostly directed crime dramas and mysteries. This movie has a cast that includes Robert Young, Randolph Scott, Dean Jagger and John Carradine. Scott plays an outlaw who tries to go straight and leave his old gang and winds up saving Jagger's life. Jagger works for Western Union, a telegraph company that plans to have telegraphs out west. Jagger hires a lot of men to make sure it is done because they have to worry about Indian attacks and bandits. Scott is in charge of the men and Young is a telegraph expert who can't shoot a gun but can ride. Scott meets up with his old gang who want to stop them but Scott can't tell anyone. It's a pretty good western and Lang should of directed some more westerns.

Fig. 3. Instantiation of the framework in this work. We consider two approaches to collecting human inputs: (a) open-ended and (b) critique-based, and present (d) selective explanations by graying out features that are predicted to be misaligned with what the user would consider as relevant for judging review sentiment. (c) represents the original explanations generated by LIME.

model predicts whether each token in the unseen instance would be considered relevant by the user, and augment the explanation accordingly—in the instantiation, we chose to grey out tokens that are in the explanations but predicted as not relevant to sentiment judgment based on the user's belief. Since we can only gather a small amount of feedback data (≈ 100 labels) from each user, we intentionally choose logistic regression, which is sample-efficient due to its simplicity. The model uses GloVe embeddings (glove-100d) [80] as features, and out-of-vocabulary words are ignored.

With open-ended input, people would only provide positive signals (which words are relevant to sentiment judgment, both positive and negative sentiment). For critique-based input, although people would explicitly provide negative signals (which words are irrelevant to sentiment judgment), empirically we find these signals are not always reliable. One reason is that people tend to disagree with the importance of words pointing in the opposite direction of the review sentiment (even if they are highlighted in the negative direction's color). Therefore, for both types of input, we used a strategy known as negative sampling in the literature [73], i.e., randomly sampling the same number of *unselected* tokens from annotation instances as negative examples, to obtain class-balanced negative signals to train the belief prediction model.

5 STUDY 1: SELECTIVE EXPLANATIONS WITH SELF-INPUT

In the first experiment, we explore generating selective explanations with the instantiation described in Section 4, focusing on getting input from the *individual user* with two input elicitation methods: open-ended and critique-based. The main task is to judge the sentiment of 20 movie reviews with the help of a sentiment analysis AI system, which provides its prediction for the review sentiment and explanation for its prediction. We compare participants' experience with the two paradigms to that of a control condition with the original, unselected explanations.

5.1 Procedure and Participants

User study task flow. Participants went through four phases depending on their condition during the study: (1) consent and attention-check; (2) input phase (omitted for the Control condition with unselected explanations); (3) task phase; (4) exit survey. Participants' answers in the input phase were used to train a belief prediction model as described in Section 4 to generate selective explanations shown in the task phase. That is, with self-input in Study 1, each participant had a personalized belief prediction model and therefore selective explanations that varied accordingly. Instructions to provide input and complete the movie sentiment judgment task were given before phases 2 and 3 separately. We added simple multiple-choice questions about the purpose of the study and what kind input they need to provide if applicable as attention-check questions. We disqualified participants who answered these questions incorrectly. In the exit survey, we collected basic demographic information and answers to the subjective measures described in Section 5.3. The study is approved by the IRB at the University. Refer to the Appendix for specific details of the user study task flow.

Participant information. Since sentiment analysis is relatively straightforward for fluent English speakers, we recruited about 40 participants for each condition from Prolific,[1] a popular crowdsourcing platform. To ensure high-quality responses, all participants satisfy the following three criteria: (1) residing in the United States; (2) English is their first language; (3) minimal approval rate of 95%. We did not allow repeated participants as the experiment follows a between-subjects design.

There were 69 male, 42 female, 5 non binary, and 2 preferred not to answer. 16 participants are aged 18-25, 55 aged 26-40, 33 aged 41-60, 12 aged over 61 and above, and 2 preferred not to answer. Participants had diverse education background. 5 have no diploma, 19 have a diploma or an equivalent, 24 have some college credit without a degree, 8 have technical/vocational training, 59 have a Bachelor's degree or above, and 3 preferred not to answer. Participants were paid an average wage of $10 per hour.

5.2 Experimental Conditions

To generate selective explanations, following our framework, participants are asked to provide input based on a sample of 10 reviews (Input Phase). Participants then perform the AI-assisted decision task by judging the sentiment of 20 new movie reviews (Task Phase).

We conduct a between-subjects experiment with the following three conditions:

- **Original explanations (*Control*).** Participants are not asked to provide any input. In the task phase, the original explanations generated by LIME are shown together with the model prediction, as illustrated in Figure 3c.
- **Selective explanations with open-ended input (*Open-ended*).** In the input phase, for each review, participants are asked to write down words that they consider important indicators for their judgment of the review sentiment (Figure 3a). In the task phase, participants are provided with the same AI assistance as in the *Control* condition but with selective explanations instead of the original explanations.
- **Selective explanations with model explanation critiques (*Critique-based*).** In the input phase, with the same sample as in the *Open-ended* condition, participants are given the AI's explanations and asked to provide input on whether they agree that each of the highlighted keywords should be considered important for the given sentiment (see Figure 3b). The task phase then shows selective explanations generated based on their critique-based input.

[1] https://www.prolific.co/.

	Correct AI predictions	**Wrong AI predictions**
Humans agree with	Appropriate agreement	**Over-reliance**
Humans disagree with	Under-reliance	Appropriate disagreement

Table 2. Definition of different human reliance situations based on whether the human agrees with the AI prediction and whether the AI prediction is correct. In an ideal scenario, humans will have *appropriate agreement* and *appropriate disagreement* with the model. Though in reality, prior work found that explanations tend to increase *over-reliance*. Therefore, in this work, we focus on the measurement of *over-reliance* and explore whether selective explanations can reduce it.

Review selection strategy. The sampling strategy for the input phase (N=10) is explained in Section 3 under "input elicitation method". Out of 10 reviews, 8 reviews are predicted correctly by the model, a close approximation of the model's accuracy.

For the task phase, we randomly sampled 20 movie reviews from the test set balanced for sentiment classes and model prediction correctness. We over-sampled cases where the model predictions are incorrect to better explore whether appropriate reliance happens. For Study 1, following Yin et al. [106], we opted for a fixed-seeding approach (i.e., all participants saw the same 20 reviews) to reduce variance, which turned out to be a limitation of this study, as we will discuss in the results and address in Study 2.

5.3 Evaluation Measures

As discussed in Section 2, informed by prior work conducting empirical studies of human-AI decision-making with explanations, we measure participants' decision accuracy (performance), reliance on AI, efficiency, and subjective perceptions about task workload, usefulness of AI assistance, and understanding of AI.

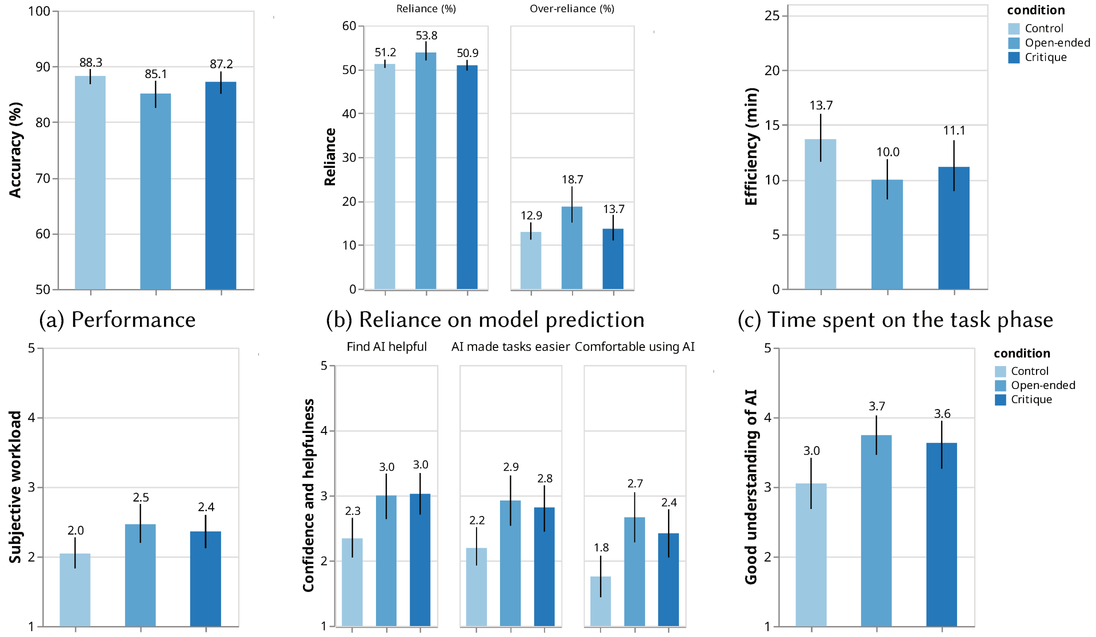
Accuracy. Human decision performance with AI assistance is measured by accuracy—percentage of reviews a participant judged correctly according to the groundtruth.

Reliance. We are interested in investigating the effect of selective explanations on people's reliance on AI assistance, defined as the percentage of cases where people's final decision is consistent with AI prediction. Informed by prior work, we are particularly interested in whether selective explanations can reduce *over-reliance*, as often found to be a pitfall of XAI for decision support. Over-reliance is defined as the percentage of cases people's decision is consistent with AI prediction when *AI is incorrect.* Table 2 illustrates when over-reliance happens.

Efficiency. We measure the total elapsed time in the task phase. Elapsed time starts from the moment participants enter the evaluation phase until they complete the last review.

Subjective measures. Our hypothesis is that selective explanations that prioritize features the recipient would consider relevant could make the explanations easier to use and more positively perceived. Meanwhile, it is also important to evaluate user experience with regard to the whole paradigm. For example, it is an open question of how providing input would impact the overall workload. We measure subjective perception with an exit survey, focusing on three categories: subjective workload, perceived usefulness of AI (with sub-measures of helpfulness, ease of task, and confidence), and understanding of AI. We list the self-rated items below, all based on a five-point Likert scale (Strongly Disagree to Strongly Agree):

- **Subjective workload**. We measure it by the average rating for three applicable items selected from NASA-TLX [44]:
  - Mental demand: I felt that the task was mentally demanding.
  - Feelings of success (reverse item): I felt successful accomplishing what I was asked to do.



(a) Performance (b) Reliance on model prediction (c) Time spent on the task phase

(d) Subjective workload (e) Confidence and helpfulness (f) Perceived understanding

Fig. 4. Results for Study 1. Error bars represent 95% conference intervals.

  - Negative emotions: I was stressed, insecure, discouraged, irritated, and annoyed during the task.
- **Perceived usefulness of AI**, with sub-measures below. We report these sub-measures separately as these items are not as established as subjective workload.
  - Helpfulness: I find the information provided by the AI helpful for making movie sentiment judgments.
  - Ease of task: Overall, the AI's assistance made the tasks easier.
  - Confidence: If I want to make movie choices, I would feel comfortable using this AI to help me find and read positive/negative reviews.
- **Perceived understanding of AI**, with one item: I feel I had a good understanding of how the AI makes predictions.

5.4 Results

We first present results on the evaluation measures introduced in Section 5.3, then dive into relevant model and user behaviors to further interpret the results. For all evaluation measures, we plot the descriptive statistics and run one-way ANOVA with the condition as the independent variable, and when significant, we conduct post-hoc Tukey's HSD test for pairwise comparisons.

Effect of selective explanations on accuracy and reliance (see Figure 4a and Figure 4b).) Using one-way ANOVA, we did not find a statistically significant effect of selective explanations on decision accuracy ($p = 0.09$), reliance on AI (percentage agreeing with AI, $p = 0.06$), or over-reliance on AI (percentage agreeing with AI among cases where the AI is wrong, $p = 0.06$). Figure 4a and Figure 4b suggest that there is a slight, albeit non-significant, trend of the *Open-ended* condition resulting in the lowest accuracy and highest over-reliance.

We note that the results of these performance-related measures should be interpreted with caution. As we will discuss later, behavior analysis showed that all participants tended to make mistakes over a small set of reviews with our fixed-seeding sampling.

Effect of selective explanations on efficiency (see Figure 4c). Selective explanations led to moderately better task efficiency among participants, especially with open-ended input, as shown in Figure 4c. Although this impact is not statistically significant based on one-way ANOVA ($p = 0.06$).

Effect of selective explanations on subjective measures. Subjective workload (see Figure 4d) is measured by average ratings of the three items from NASA-TLX. Participants who experienced the two selective explanation paradigms found the task to be more mentally demanding than those in the *Control* condition. One-way ANOVA showed that selective explanations had a significant impact on mental demand ($p < 0.05$) and post-hoc Tukey's HSD showed statistical differences in both *Open-ended* and *Critique-based* conditions with the *Control* condition ($p < 0.05$). Looking at the sub-items, this difference is mainly due to the sub-items of being "mentally demanding", rather than feeling unsuccessful with the task or having negative emotions. This increased subjective workload is likely due to the extra work required in providing input.

For *perceived usefulness (see Figure 4e)*, participants who were in the two conditions with selective explanations reported higher perceived usefulness of the AI tool than the *Control* condition across the board: higher perceived helpfulness, improved ease of the task with the AI, and higher confidence in the AI. In all three measures, one-way ANOVA revealed a significant effect ($p < 0.05$) and post-hoc Tukey's HSD suggested a statistical difference ($p < 0.05$) in both treatment conditions versus the *Control* condition.

Finally, for *perceived understanding (see Figure 4f)* participants who were in the two conditions with selective explanations reported a better understanding of the model than those in the *Control* condition. Using one-way ANOVA, selective explanations had a significant impact ($p < 0.05$), and post-hoc Tukey's HSD found the difference between *Control* versus *Critique-based* is statistically significant ($p < 0.05$), and *Control* versus *Open-ended* is marginally significant ($p = 0.05$).

In summary, while we did not find improvement in decision performance over the fixed set of reviews, we found that selective explanations with both types of self-input paradigms have a positive effect in improving perceived usefulness and understanding of the AI, a moderate effect in increasing efficiency, but also increased the overall subjective workload by requiring the additional effort of providing input.

5.5 Model and User Behavioral Analysis to Further Understand the Lack of Improvement on Performance

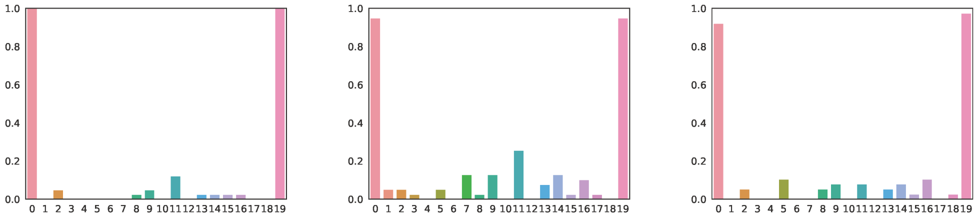
We conducted further analyses to unpack why we observed a slight decrease in accuracy with selective explanations in the *Open-ended* condition.

Accuracy results are biased because of two reviews. First, we looked at the distribution of error rates among the 20 reviews all participants saw. As shown in Figure 6, participants' mistakes are highly concentrated in review 0 and 19. While selective explanations based on open-ended and critique-based input reduce errors for those two reviews, they increased the errors on other reviews. The increase in errors on other reviews in open-ended input is especially prominent, which we will further address in the next paragraph. This observation suggests that our fixed-seeding sampling might have limited the generalizability of our results regarding the decision performance.

Input was comparatively worse in the Open-ended *condition.* We compared the quantity and quality of participants' input in the two conditions. In the *Critique-based* condition, participants gave relevance-positive feedback on an average of 51.9 unique words, while in *Open-ended* condition,

Top 10 selected words		Top 10 misaligned words	
Open-ended	Critique	Open-ended	Critique
masterpiece	good	this	this
good	annoying	not	is
amazing	fun	is	movie
beautiful	excellent	movie	no
happy	best	no	cast
believable	masterpiece	best	and
enjoyable	waste	like	story
heart	worst	and	her
great	amazing	cast	not
real	beautiful	story	bad

Table 3. Top 10 selected and misaligned words in the *Open-ended* and *Critique-based* condition.



(a) *Control*

(b) *Open-ended*

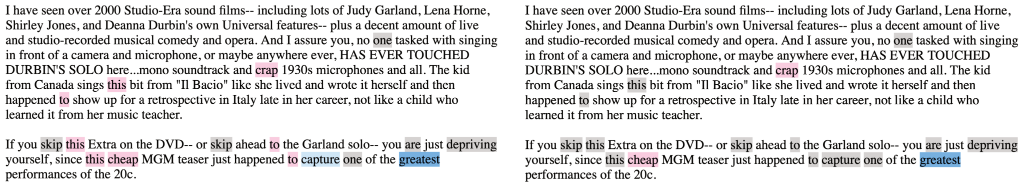
(c) *Critique-based*

Fig. 5. Error rates grouped by review id in the three conditions.

the average number of unique words given was only 11.1. Table 3 further shows the top relevant words chosen by participants in the input phase and the top misaligned words in the task phase. Participants in the *Open-ended* condition overwhelmingly focus on positive-sentiment words in their feedback. In fact, none of the top 10 relevant words is relevant for negative sentiment, while the *Critique-based* condition identified words like “annoying”, ”waste”, and ”worst”. This bias and lack of diversity also led to lower-quality misaligned words identified. For instance, “best” and “like” are among the top 10 misaligned words in the *Open-ended* condition.

This lower quantity and quality of input in the open-ended may have contributed to the slight decrease in accuracy in the *Open-ended* condition. As shown in Figure 5b and Figure 5c, in some cases (e.g., review 11 and review 7), participants made more mistakes in the *Open-ended* condition. Figure 6 shows examples for review 11 from two participants assigned to the two conditions. While the groundtruth label is positive, the AI explanation in the *Open-ended* condition included many irrelevant words that are highlighted as “negative” such as “this” and “to”. In comparison, the *Critique-based* condition managed to gray these words out. As a result, participants were more likely to over-rely on the incorrect model prediction in the *Open-ended* condition and judged the review to be negative.

Do selective explanations increase the percentage of highlighted features supporting the groundtruth label? We ask this question to probe on whether there is a theoretical possibility or limit for selective explanations to improve performance and decrease over-reliance. In general, if selective explanations can increase this percentage, it is more likely to nudge people toward making correct predictions consistent with the groundtruth. For example, considering when a model prediction is incorrect, the original explanation would highlight fewer features supporting the groundtruth



I have seen over 2000 Studio-Era sound films-- including lots of Judy Garland, Lena Horne, Shirley Jones, and Deanna Durbin's own Universal features-- plus a decent amount of live and studio-recorded musical comedy and opera. And I assure you, no one tasked with singing in front of a camera and microphone, or maybe anywhere ever, HAS EVER TOUCHED DURBIN'S SOLO here...mono soundtrack and crap 1930s microphones and all. The kid from Canada sings this bit from "Il Bacio" like she lived and wrote it herself and then happened to show up for a retrospective in Italy late in her career, not like a child who learned it from her music teacher.

If you skip this Extra on the DVD-- or skip ahead to the Garland solo-- you are just depriving yourself, since this cheap MGM teaser just happened to capture one of the greatest performances of the 20c.

(a) *Open-ended*

I have seen over 2000 Studio-Era sound films-- including lots of Judy Garland, Lena Horne, Shirley Jones, and Deanna Durbin's own Universal features-- plus a decent amount of live and studio-recorded musical comedy and opera. And I assure you, no one tasked with singing in front of a camera and microphone, or maybe anywhere ever, HAS EVER TOUCHED DURBIN'S SOLO here...mono soundtrack and crap 1930s microphones and all. The kid from Canada sings this bit from "Il Bacio" like she lived and wrote it herself and then happened to show up for a retrospective in Italy late in her career, not like a child who learned it from her music teacher.

If you skip this Extra on the DVD-- or skip ahead to the Garland solo-- you are just depriving yourself, since this cheap MGM teaser just happened to capture one of the greatest performances of the 20c.

(b) *Critique-based*

Fig. 6. Example review with selective explanations in the *Open-ended* condition and the *Critique-based* condition. For this positive review, more irrelevant words are highlighted in pink, indicating negative label, in the *Open-ended* conditions, while they are grayed out in the *Critique-based* condition.

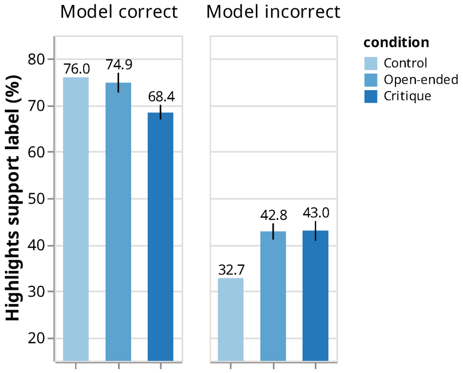


Fig. 7. Percentage of highlighted features supporting the groundtruth grouped by reviews where the model made correct and incorrect predictions.

(contradicting the prediction) than features contradicting the groundtruth (supporting the prediction). If the selective explanation can increase the percentage of the former, or even make it the majority, people may be more likely to notice keywords that support the groundtruth instead of following AI's incorrect prediction.

In Figure 7, we plot the percentage of highlighted words supporting the groundtruth over all highlighted words,[2] separating reviews where the model made correct and incorrect predictions. While this percentage slightly decreased when model predictions were correct in the *Critique-based* condition, it consistently increased when the model predictions were incorrect with selective explanations over the control conditions. These observations suggest a theoretical possibility for selective explanation to reduce over-reliance by following the visual highlighting patterns. However, the review and keywords content may still play a greater role in people's judgment.

In short, the additional behavioral analyses suggest that the changes to the keywords highlighting patterns through selective explanations point to a positive direction of reducing over-reliance. However, it did not translate to actual improvement, which can possibly be attributed to our sampling effect. Meanwhile, our analysis suggests that critique-based input may result in higher quantity and quality input that are more effective in identifying misaligned words than open-ended input. In Study 2, we remove the limitation of fixed sampling by using a random sampling strategy.

6 STUDY 2: SELECTIVE EXPLANATIONS WITH ANNOTATOR INPUT

In Study 2, our main goal is to explore a different input strategy by eliciting input from a panel of similar users as annotators. As discussed in Section 3, the design choice of *whose input* is an

[2] We count all occurrences of a unique word to account for the visual effect that each of them is highlighted.

important one in practice as individuals may not afford the time and effort to provide input, as further highlighted by the increased cognitive load in Study 1. However, it is an open question whether individual users would find the input from others useful. Furthermore, in light of the limitation of fixed sampling in Study 1, we also introduce a random sampling strategy in the task phase to improve the generalizability of our results. To allow comparisons of results across the two studies and explore the robustness of Study 1 results with the fixed sample, besides an experimental condition and a control condition with random samples, we introduced another pair of them with the same fixed sample used in Study 1.

6.1 Study Design

We conduct a between-subjects experiment with the following four conditions. To generate selective explanations with annotator input, we take all the keywords shown to participants in the *Critique-based* condition in Study 1 (who are "similar" participants recruited from the same platform with the same criteria), and take the majority vote among all previous participants in this condition as the input data, then train the belief prediction model as described in Section 4. That is, different from Study 1, where each participant had a personalized model to predict their beliefs about feature relevance based on their own input data, in Study 2, there is a fixed model for all participants. We chose to include all participants in Study 1 as the panel of annotators to avoid making arbitrary filtering decisions. The required number of annotators in practice is likely much lower. We encourage future work to explore other, more efficient, approaches to obtain annotator input.

- **Random sample control (with original explanations).** This condition is similar to the control condition in Study 1, except that the reviews shown to the participants are randomly sampled while maintaining the balance of sentiment class and prediction correctness.
- **Random sample with selective explanations.** This condition shows selective explanations generated with annotator input but with random sampling as the condition above.
- **Fixed sample control (with original explanation).** This condition is identical to the control condition in Study 1.
- **Fixed sample with selective explanations.** This condition shows the same fixed sample in Study 1 and selective explanations with annotator input.

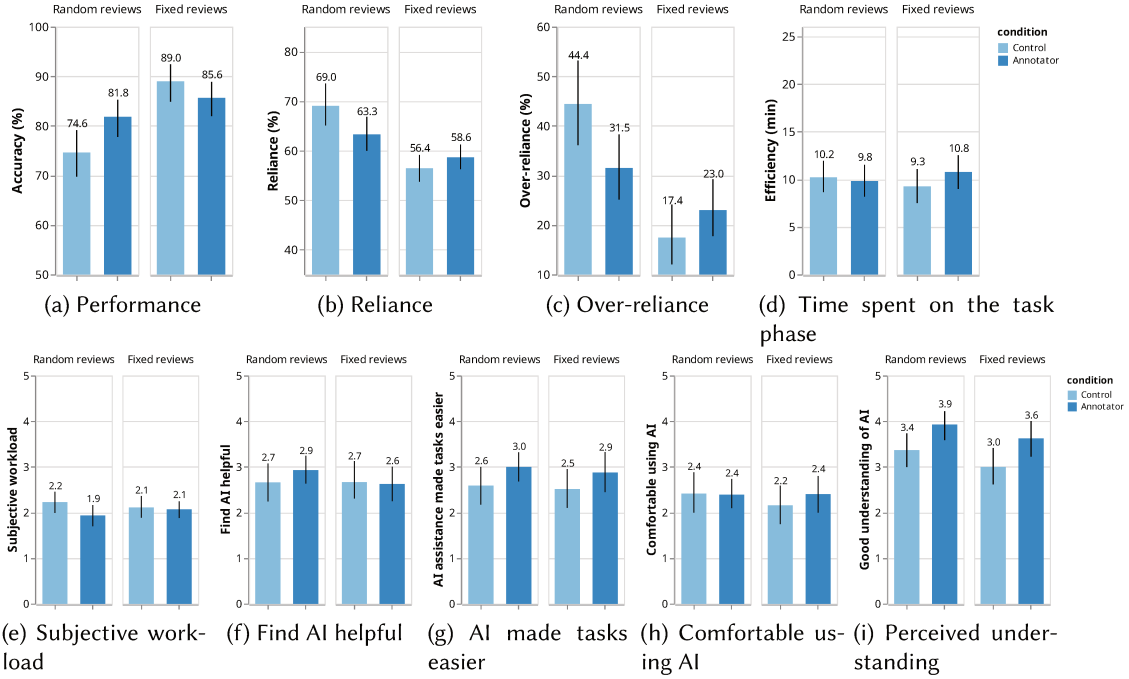
The evaluation measures and procedures are similar to Study 1, except that the input phase is removed for all conditions.

Participant information. Similar to Study 1, for each condition, we recruited about 40 participants from Prolific. There were 75 male, 83 female, and 3 non-binary. 31 participants are aged 18-25, 71 aged 26-40, 46 aged 41-60, 12 aged over 61 and above, and 1 preferred not to answer. In addition, participants had diverse education background: 19 are high school graduates or equivalent, 42 have some college credit without a degree, 12 have technical/vocational training, 77 have a Bachelor's degree or above, and 2 preferred not to answer. Participants were paid an average wage of $12 per hour. Refer to Section 5 for details on user study task flow (the only difference lies in that there is no input phase for all conditions in Study 2).

6.2 Results

We used the same evaluation measures in Study 2 as in Study 1. We start by comparing the results from the two random-sample conditions to understand the effect of selective explanations with annotator input. Then we conduct the same analyses for the two fixed-sample conditions to allow cross-study comparisons. We conduct t-tests for all measures.

6.2.1 *Effect of Selective Explanations on Random Samples.*



(a) Performance (b) Reliance (c) Over-reliance (d) Time spent on the task phase

(e) Subjective workload (f) Find AI helpful (g) AI made tasks easier (h) Comfortable using AI (i) Perceived understanding

Fig. 8. Results for Study 2. Error bars represent 95% confidence interval.

Effect of selective explanations on accuracy and reliance (see Figure 8a). We find a sizable improvement in accuracy from selective explanations (74.6% vs. 81.8%), and this difference is statistically significant ($p < 0.05$). Furthermore, we observe a substantial drop in reliance (69.0% to 63.3%), which can be mainly attributed to the greater drop in over-reliance from 44.4% to 31.5% (recall that our test samples are balanced across prediction correctness). Both differences are statistically significant ($p < 0.05$).

These results suggest that, with random sampling, selective explanations reduced over-reliance and improved overall decision performance. Following the model behavior analysis in Section 5.5, to understand the reasons for the reduced over-reliance, we examine the percentage of highlighted words that support the groundtruth label, grouped by prediction correctness. Figure 9 shows that, with random samples, selective explanations based on annotator input substantially increases the percentage of highlighted words that support the correct label in incorrectly predicted instances from 35.9% to 43.5%, which could have contributed to the reduced over-reliance. Moreover, we also observe a slight increase of this percentage in correctly predicted instances, from 77.7% to 79.2%.

Effect of Selective Explanations on Efficiency (see Figure 8d). Different from Study 1, selective explanations with annotator input virtually has no impact on efficiency ($p = 0.76$). This suggests that the moderate benefit in efficiency observed in Study 1 should be attributed to an improved familiarity with the task by completing the input phase, rather than seeing selective explanations.

Effect of Selective Explanations on subjective measures (see Figure 8e-8i). Similar to Study 1, we found that selective explanations with annotator input led to a significant improvement in the perceived understanding of AI over the control condition, which is arguably the most important subjective measure of explanations [56]. However, different from Study 1, we did not observe significant differences in subjective workload and perceived usefulness. This suggests that the positive effect on perceived AI usefulness observed in Study 1 should be attributed to the opportunity to

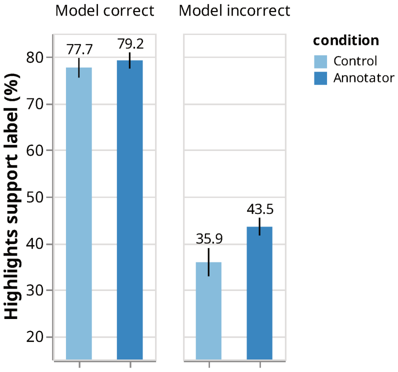


Fig. 9. Percentage of highlighted words supporting the correct label in Study 2 for the random sample. The results in the fixed sample are similar to Study 1.

provide one's own input and have control over AI outputs, rather than seeing selective explanations alone. On the other hand, removing the requirement for providing one's own input also removed the additional cognitive load observed in Study 1, suggesting a trade-off between workload and user agency to use selective explanations with others' input.

In summary, by removing the limitation of fixed sampling in Study 1, the results in Study 2 demonstrate the promise of selective explanations in improving performance and reducing over-reliance. These results are especially exciting given the growing concerns about the XAI pitfall leading to over-reliance when the AI is wrong [6, 93]. We will further reflect on this result and its implications in Section 7. Using selective explanations based on annotator input still consistently improved the perceived understanding of the model, and removed the additional subjective workload required by the input phase, but the positive effects on the perception of AI usefulness and efficiency are absent without the input phase to familiarize oneself with the task and have personal control over generating selective explanations.

6.2.2 Effect of Selective Explanations on Fixed Samples. We now briefly discuss the results based on the two conditions with the fixed sample, mainly to compare the results with Study 1. First, similar to Study 1, we did not observe a significant difference in selective explanations on decision accuracy, reliance, and over-reliance (Figure 8a), and in fact a negative trend consistent with Study 1, further confirming that the fixed sample bias results on these performance-related measures and the Study 1 results on these measures should be interpreted with caution.

Similar to the results with the random sample described above in Section 6.2.1, we found an improvement only in *perceived understanding of AI* ($p < 0.05$) (Figure 8i), but not efficiency (Figure 8d), subjective load (Figure 8e), or perceive AI usefulness (Figure 8f-8h). These results suggest that these effects (and lack thereof) of selective explanations with annotator input on subjective measures are robust, and that the sampling method might have a limited impact on our results (in both studies) on subjective measures.

7 DISCUSSION

In this work, we propose a general framework for generating selective explanations by leveraging human input. Our framework provides a recipe for closing the gaps between AI explanation algorithms and how humans consume and provide explanations. We instantiate our framework with a text classification task and use the selective explanations in a testbed of AI-assisted decision-making. Experimental results with human subjects demonstrate the promise of selective explanations and also highlight the complexity of the design space. In this section, we further interpret the results to reflect on the underlying reasons and lessons learned, then discuss the generalizability and open questions of our framework for future work.

7.1 Reflection on the Results

Effect of selective explanations. We consistently find that selective explanations improve the perceived understanding of the model, which is often considered a primary goal of providing AI explanations [56]. As shown in the example in Figure 3d, by graying out irrelevant words, selective explanations are less noisy and visually sparser, concentrating on more relevant words and enabling easier sense-making of model predictions.

We highlight the improvement in participants' decision performance and decrease in over-reliance with random sampling in Study 2. In both studies, for cases where the model is wrong, we observe an increase in the percentage of highlights supporting the groundtruth labels, thus contradicting the incorrect predictions. This suggests that, by removing irrelevant words, selective explanations are also systematically removing more "wrongly picked" features that contribute to the model's wrong predictions. This provides a possible path for better signaling groundtruth labels to help decision-makers avoid over-relying on the model predictions and make better decisions. This path resonates with a recent theoretical work by Chen et al. [19], which suggests that feature-based explanations can only reveal model decision boundaries (how the model makes decisions), and it is by their contrast with human intuitions about the task boundaries (which features *should* contribute to the outcome) can one detect model errors. We may in fact view the gray-out words as such contrasts.

We believe this systematic reduction of "wrongly picked" features and signaling of model errors should be attributed to the fact that participants in our study were able to bring in reasonable intuition about task boundaries for movie review sentiment judgment. It is unclear whether such an outcome can be observed when the input provider knows little about the task. Therefore, to harvest this benefit of reducing overreliance on AI, future work could consider eliciting input from *domain experts* of the given decision task to generate selective explanations. However, we acknowledge that the input elicitation methods used in the current instantiation may not be optimized for generating such contrasts for wrong model predictions (e.g., the input phase saw a limited number of wrong predictions). It is also possible to create a visual design that more explicitly highlights the contrast, such as the dual-color scheme used in Boggust et al. [10], which shows human rationales in a different color. We encourage future work to explore possibilities to further enhance the effect of selective explanations on reducing over-reliance.

Effect of user input. Our results identify a few intriguing effects of user input that could have broad implications for human-in-the-loop or interactive ML work. First, we may attribute the positive effects on perceived AI usefulness and task efficiency observed in Study 1 but not Study 2 to participants providing their own input. That is, not only did participants better familiarize themselves with the task and the AI by going through an input phase, but they also felt more positively about the AI knowing that they had control over its output. While at a cost of the overall workload, these benefits of providing self-input should be broadly considered for improving user experience of AI systems.

Second, our results suggest that when eliciting human rationales, whether to improve explanations [31, 33] or models [16, 37, 90], a critique-based approach by asking for feedback to model explanations may result in better quantity and efficiency of human input over an open-ended approach. That said, it is possible to design better elicitation prompts and incentives to elicit open-ended feedback if the goal is to optimize for coverage of different features.

Effect of sampling strategy. It is worth noting that our initial results on performance suffered from the choice of sampling. Inspired by Yin et al. [106], we chose to use a fixed sample of instances to reduce the variance. While the examples seemed representative to us, they introduced biases that limit the generalizability of our results. This observation highlights a critical challenge of studying interactions with AI-powered systems being the have high variance and uncertainty of the output space [104].

7.2 Generalizability, Open Questions, and Future Directions

Implications and open questions for other design choices. Our instantiation implements only a subset of design choices for the use case of AI-supported sentiment judgment (row 1 of Table 1). Table 1 lists other XAI use cases that require other selectivity goals and accordingly, modifications of the other design dimensions. Below we postulate on these design decisions and encourage future work to explore them empirically with specific XAI use cases.

To realize the abnormality goal, it is possible that providing open-ended feedback by answering "which features should the model NOT base its decisions on" will be especially challenging. Instead, critique-based feedback can be elicited with a similar interface as in Figure 3b but focusing on asking input about which parts of the model explanation is abnormal, either indicative of model errors (for debugging model) or surprising to the recipient (for knowledge discovery). For sampling strategy, it is possible that prioritizing cases where the model makes mistakes or different decisions from the human could be more effective for eliciting abnormality signals. For the visual presentation of selective explanations, if it is important to preserve the original explanations or the abnormal parts are relatively sparse, an alternative is to add highlights to parts that are potentially abnormal instead of greying out the rest. Lastly, depending on the use case, more sophisticated algorithms may need to be developed for generating selective explanations from human input. For example, to assist knowledge discovery, ideally the selected parts should be surprising to the user but also verifiability correct, which may pose additional requirements for the computational algorithms.

For the changeability goal, we believe the definition of "changeability" must be carefully operationalized according to the use case. For example, people may have different constraints on what actions they can take to improve their chance of loan approval versus their health risk. Current XAI methods, while claiming to support resource, have been criticized for false assumptions [7], including lacking mapping from changes in features to real-world actions, and negligence of interrelated changes between features or with real-life factors invisible to the model. While selective explanation offers a path for these issues by prioritizing changes that the recipient would subjectively believe to be changeable, they cannot be solved if the model features are not meaningful or lack real-world paths for change. How the changeability is operationalized should also be communicated when eliciting input, for example, by providing context such as what changes may involve and how to gauge the cost for change. Furthermore, depending on the use case, changeability could be highly personalized, and elicitation from "a group of similar users" may not yield useful results.

Extension to different models and data types. As an augmentation approach, our framework can be applied to any existing XAI techniques that output feature-importance explanations, whether through post-hoc algorithms that generate explanations for "black-box" models (e.g., LIME and SHAP), or "clear-models" that provide feature coefficients directly (e.g., linear regression model). However, it is possible that the observed effects will be diluted if the post-hoc explanation itself is highly unfaithful to how the model actually works and introduces high noises in the explanations.

How to transfer our approach to models using other types of data than texts involves non-trivial challenges. For image data, human input could focus on giving feedback on or choosing regions of the image if high-quality image segmentation is available. For tabular data, we may need to reconsider the details of the elicitation methods. While it is possible to ask people to provide feedback for a small sample or provide their own rank in an open-ended fashion, this could become hard to manage if the feature space is large. When generalizing human input to unseen cases, the main challenge could be the non-linearity of feature importance. Therefore, we speculate that more sophisticated sampling methods may be required for tabular data to elicit human input efficiently and effectively. Furthermore, for the belief prediction, we made a simplified assumption that the role of each feature is stable across instances. A natural extension is to relax this assumption and explore the different roles of features in different instances. For text data, one possible strategy is to use contextualized word embeddings.

Beyond feature-importance explanations. The extension of our framework to another category of feature-based explanation—counterfactual explanations [94, 95]—is straightforward. In fact, the changeability goal is a natural fit for counterfactual explanations, which often aim to help people identify which feature they should focus on changing in order to obtain a different, often more desirable model prediction. Current counterfactual XAI techniques simply use the theoretical "distance" to search for features that require minimum change distance to "flip" the prediction to the target prediction. They can incorporate inferred recipient beliefs about changeability in the distance measure.

Future work can also explore utilizing beliefs about user preferences to augment the selection of examples in example-based explanations [42, 67, 98]. For example, when selecting the "most similar" examples from the training data to explain or justify the current prediction, this similarity measure could account for which features would be considered most relevant by the recipient and assign higher weights to them.

Potential issues with selective explanations. We also encourage future work to critically examine potential issues in the assumptions underlying selective explanations. The first is a fundamental tension between selectivity and faithfulness of explanations. Whether model explanations should always be faithful is a debated topic. While it is argued that faithfulness is critical in high-stakes situations or serving actions on the model such as debugging [86], some also contend that explanations do not need to be perfectly faithful (e.g., using post-hoc explanations) to provide useful information to help people make better sense of and work with the model [62, 66]. We further point out that the goal of selectivity is not to deceive but to help people better process information without being overwhelmed, and people should maintain control of what they wish to see. Under our framework (Section 4), we also recommended multiple visual presentation choices that still preserve the original content with less tradeoff of faithfulness, such as adding highlights to original explanations. That being said, future work should examine in what situations selective explanation can result in missing information and what are the risks.

A second set of issues are related to who can provide input and who will be disadvantaged. As Study 1 shows, providing one's own input adds cognitive load, and in practice, not every individual can afford the time and resources to do so. Even if opportunities are given to everyone, the quality of input may vary by their expertise, time available, and other individual factors, which can result in inequality of benefits they can harness from selective explanations. Future work should explore how to narrow the gaps in input quality through more efficient and better-designed elicitation methods. Our Study 2 provides positive evidence for eliciting input from a panel of annotators, which can eliminate the burden for individual users and possibly mitigate the inequality issue. However, open questions remain on how to choose such a panel to be representative and inclusive, what are the risks of misrepresentation, and how to regulate system developers to avoid intentional misrepresentation.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments. This paper is supported in part by NSF grants, IIS-2040989 and IIS-2126602.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[3] David Alvarez-Melis, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. 2021. From human explanation to model interpretability: A framework based on weight of evidence. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

[4] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[7] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 80–89.

[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. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[9] Or Biran and Kathleen R McKeown. 2017. Human-Centric Justification of Machine Learning Predictions.. In *IJCAI*, Vol. 2017. 1461–1467.

[10] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. 2022. Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. In *CHI Conference on Human Factors in Computing Systems*. 1–17.

[11] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[12] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[13] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, 160–169.

[14] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 258–262.

[15] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 4.

[16] Samuel Carton, Surya Kanoria, and Chenhao Tan. 2022. What to Learn, and How: Toward Effective Learning from Rationales. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 1075–1088. https://doi.org/10.18653/v1/2022.findings-acl.86

[17] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.

[18] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[19] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2022. Machine Explanations and Human Understanding. *arXiv preprint arXiv:2202.04092* (2022).

[20] Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. 2022. Interpretable machine learning: Moving from mythos to diagnostics. *Queue* 19, 6 (2022), 28–56.

[21] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 559.

[22] Nancy J Cooke. 1994. Varieties of knowledge elicitation techniques. *International journal of human-computer studies* 41, 6 (1994), 801–849.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

[24] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[25] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.

[26] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 275–285.

[27] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[28] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. 2021. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509* (2021).

[29] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.

[30] Upol Ehsan and Mark O Riedl. 2021. Explainability pitfalls: Beyond dark patterns in explainable AI. *arXiv preprint arXiv:2109.12480* (2021).

[31] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.

[32] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing human-centered perspectives in explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[33] Shi Feng and Jordan Boyd-Graber. [n. d.]. Learning to Explain Selectively: A Case Study on Question Answering. In *Empirical Methods in Natural Language Processing*.

[34] Sorelle A Friedler, Chitradeep Dutta Roy, Carlos Scheidegger, and Dylan Slack. 2019. Assessing the local interpretability of machine learning models. *arXiv preprint arXiv:1902.03501* (2019).

[35] Krzysztof Z Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces*. 794–806.

[36] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[37] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2020. Explainable Active Learning (XAL): An Empirical Study of How Local Explanations Impact Annotator Experience. *arXiv preprint arXiv:2001.09219* (2020).

[38] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[39] Ana Valeria Gonzalez, Gagan Bansal, Angela Fan, Robin Jia, Yashar Mehdad, and Srinivasan Iyer. 2020. Human Evaluation of Spoken vs. Visual Explanations for Open-Domain QA. *arXiv preprint arXiv:2012.15075* (2020).

[40] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 90–99.

[41] 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), 50.

[42] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2019), 93.

[43] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web* 2, 2 (2017), 1.

[44] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[45] Germund Hesslow. 1988. The problem of causal selection. *Contemporary science and natural explanation: Commonsense conceptions of causality* (1988), 11–32.

[46] Denis Hilton. 2017. Social attribution and explanation. (2017).

[47] Denis J Hilton and L McCLURE John. 2007. The course of events: counterfactuals, causal sequences, and explanation. In *The psychology of counterfactual thinking*. Routledge, 56–72.

[48] Denis J Hilton and Ben R Slugoski. 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review* 93, 1 (1986), 75.

[49] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[50] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[51] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 353–362.

[52] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[53] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. PMLR, 2668–2677.

[54] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 411.

[55] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006* (2019).

[56] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[57] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[58] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[59] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[60] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *arXiv preprint arXiv:2001.02478* (2020).

[61] Q Vera Liao and S Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. *Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency* (2022).

[62] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).

[63] Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 147–159.

[64] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.

[65] Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements* 27 (1990), 247–266.

[66] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[67] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021),

1–45.

[68] Ana Lucic, Hinda Haned, and Maarten de Rijke. 2020. Why does my model fail? contrastive local explanations for retail forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 90–98.

[69] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[70] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 142–150.

[71] Bertram F Malle. 2006. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press.

[72] John McClure and Denis Hilton. 1997. For you can't always get what you want: When preconditions are better explanations than goals. *British Journal of Social Psychology* 36, 2 (1997), 223–240.

[73] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/arXiv.1301.3781 arXiv:1301.3781 [cs]

[74] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[75] 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 preprint arXiv:1802.00682* (2018).

[76] Shweta Narkar, Yunfeng Zhang, Q Vera Liao, Dakuo Wang, and Justin D Weisz. 2021. Model LineUpper: Supporting Interactive Model Comparison at Multiple Levels for AutoML. In *26th International Conference on Intelligent User Interfaces*. 170–174.

[77] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval* 2, 1–2 (2008), 1–135.

[78] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *CHI Conference on Human Factors in Computing Systems*. 1–9.

[79] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[80] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[81] 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).

[82] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[83] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of KDD*.

[84] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[85] Justus Robertson, Athanasios Vasileios Kokkinakis, Jonathan Hook, Ben Kirman, Florian Block, Marian F Ursu, Sagarika Patra, Simon Demediuk, Anders Drachen, and Oluseyi Olarewaju. 2021. Wait, but why?: assessing behavior explanation strategies for real-time strategy games. In *26th International Conference on Intelligent User Interfaces*. 32–42.

[86] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[87] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2022. Talktomodel: Understanding machine learning models with open ended dialogues. *arXiv preprint arXiv:2207.04154* (2022).

[88] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Dan Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of CHI*.

[89] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*. 107–120.

[90] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International*

journal of human-computer studies 67, 8 (2009), 639–662.

[91] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[92] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. 2021. Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[93] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. 2022. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *arXiv preprint arXiv:2212.06823* (2022).

[94] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).

[95] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR.

[96] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[97] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 601.

[98] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[99] Hilde JP Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A Human-Grounded Evaluation of SHAP for Alert Processing. *arXiv preprint arXiv:1907.03324* (2019).

[100] James Woodward. 2006. Sensitive and insensitive causation. *The Philosophical Review* 115, 1 (2006), 1–50.

[101] Jennifer Wortman Vaughan and Hanna Wallach. 2021. A Human-Centered Agenda for Intelligible Machine Learning. In *Machines We Trust: Perspectives on Dependable AI*, Marcello Pelillo and Teresa Scantamburlo (Eds.). MIT Press.

[102] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[103] 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.

[104] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

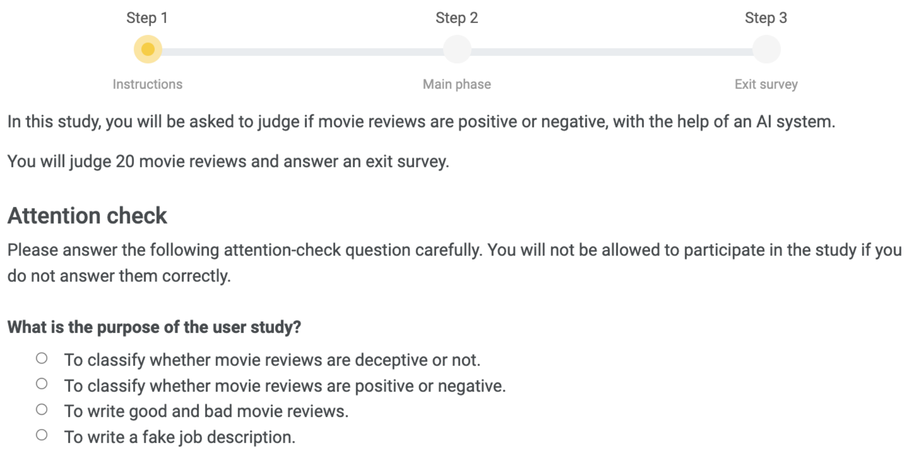
[105] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. https://doi.org/10.48550/arXiv.1906.08237 arXiv:1906.08237 [cs]

[106] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 279.

[107] Wencan Zhang and Brian Y Lim. 2022. Towards Relatable Explainable AI with the Perceptual Process. In *CHI Conference on Human Factors in Computing Systems*. 1–24.

[108] 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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

Fig. 10. The figure is omitted to preserve anonymity. A participant reads the consent form before commencing the task. The agree button is omitted in the screenshot due to space constraints. Given that the consent form includes the University, we will display the consent form when the submission does not require anonymity.



In this study, you will be asked to judge if movie reviews are positive or negative, with the help of an AI system.

You will judge 20 movie reviews and answer an exit survey.

Attention check

Please answer the following attention-check question carefully. You will not be allowed to participate in the study if you do not answer them correctly.

What is the purpose of the user study?

○ To classify whether movie reviews are deceptive or not.
○ To classify whether movie reviews are positive or negative.
○ To write good and bad movie reviews.
○ To write a fake job description.

Fig. 11. The participant is briefed on what is expected of the task. This figure shows the instruction of a participant whose task is to perform the prediction task only.

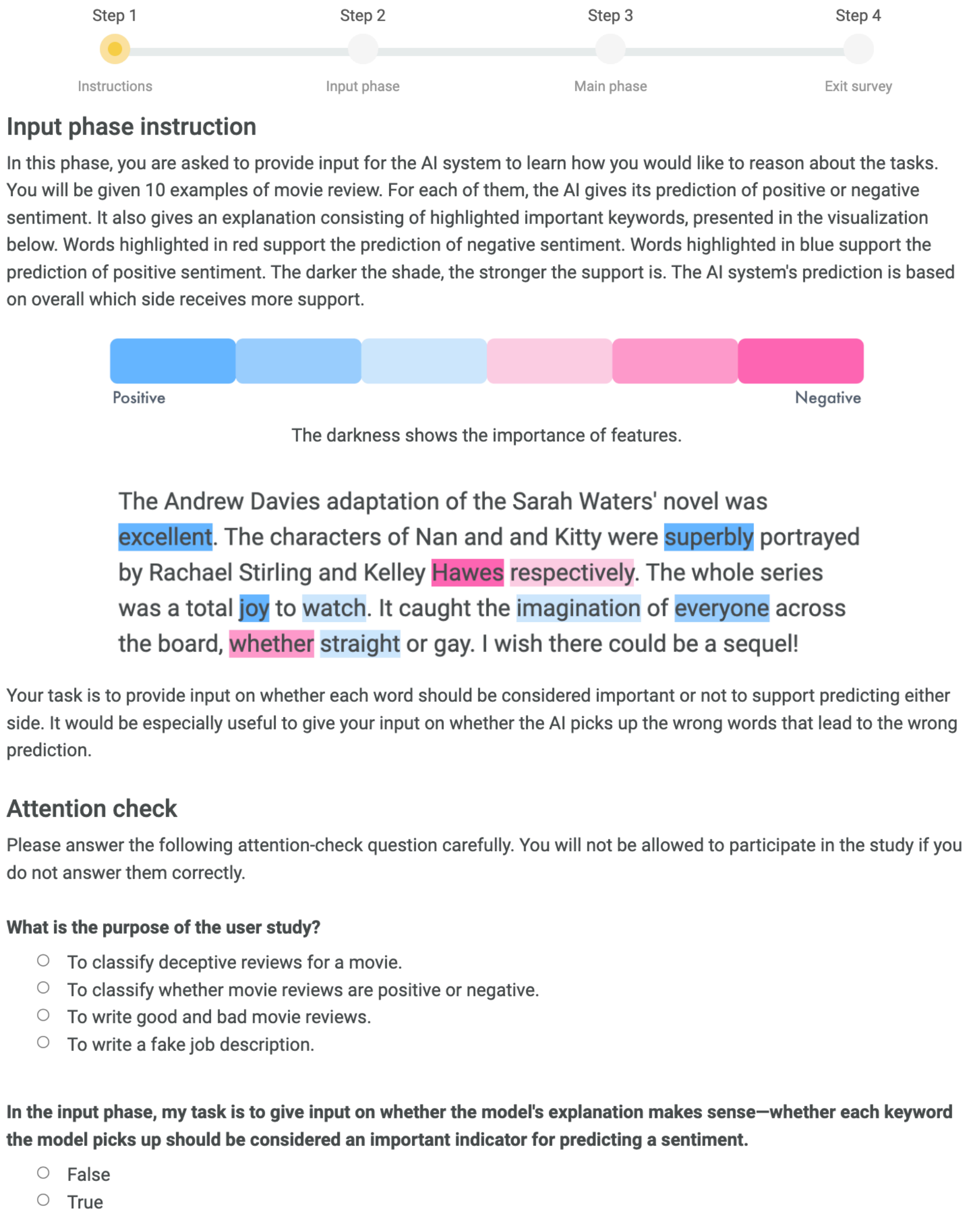
A APPENDIX

A.1 User Study Task Flow

Generally, the participants went through four phases during the study and we will describe each phase in detail. Refer to the figures for details on the interface.

(1) Read the consent form (see Figure 10), read the instructions, and answer attention check questions (see Figure 11 and Figure 12).
(2) Complete the input phase (see Figure 13 and Figure 14). Note that this phase only applies to participants tasked to provide input on their beliefs in explanations.
(3) Complete the prediction phase (see Figure 16).
(4) Complete the task by answering demographic and subjective questions (see Figure **??** and Figure 17).

Received January 2023; revised April 2023; accepted May 2023



Input phase instruction

In this phase, you are asked to provide input for the AI system to learn how you would like to reason about the tasks. You will be given 10 examples of movie review. For each of them, the AI gives its prediction of positive or negative sentiment. It also gives an explanation consisting of highlighted important keywords, presented in the visualization below. Words highlighted in red support the prediction of negative sentiment. Words highlighted in blue support the prediction of positive sentiment. The darker the shade, the stronger the support is. The AI system's prediction is based on overall which side receives more support.

The darkness shows the importance of features.

The Andrew Davies adaptation of the Sarah Waters' novel was excellent. The characters of Nan and and Kitty were superbly portrayed by Rachael Stirling and Kelley Hawes respectively. The whole series was a total joy to watch. It caught the imagination of everyone across the board, whether straight or gay. I wish there could be a sequel!

Your task is to provide input on whether each word should be considered important or not to support predicting either side. It would be especially useful to give your input on whether the AI picks up the wrong words that lead to the wrong prediction.

Attention check

Please answer the following attention-check question carefully. You will not be allowed to participate in the study if you do not answer them correctly.

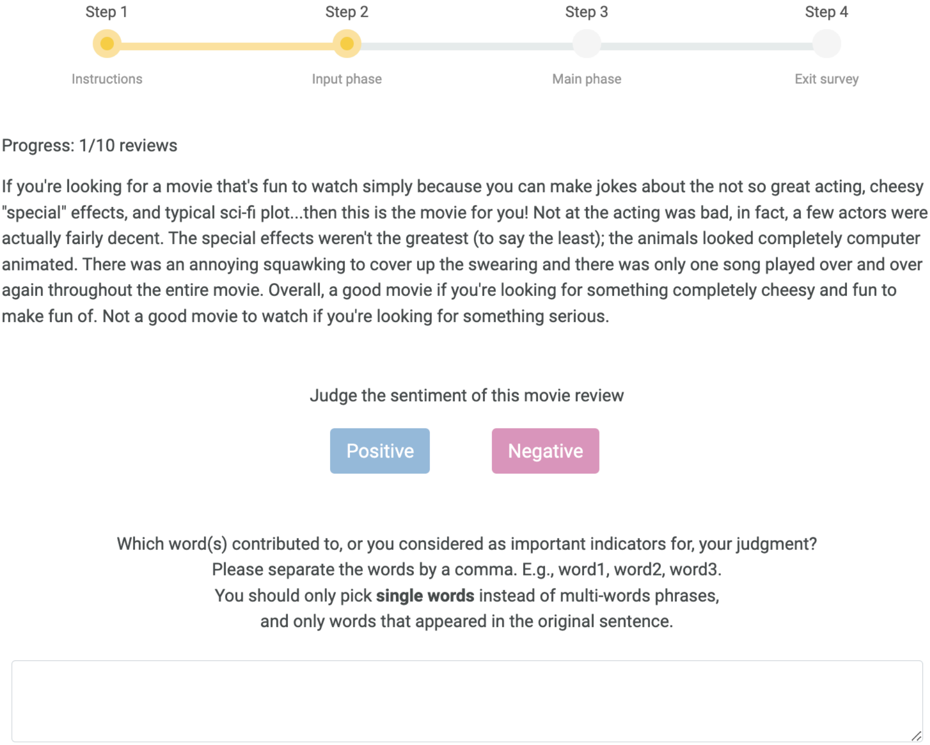
What is the purpose of the user study?

○ To classify deceptive reviews for a movie.
○ To classify whether movie reviews are positive or negative.
○ To write good and bad movie reviews.
○ To write a fake job description.

In the input phase, my task is to give input on whether the model's explanation makes sense—whether each keyword the model picks up should be considered an important indicator for predicting a sentiment.

○ False
○ True

Fig. 12. The participant is briefed on what is expected of the task. This figure shows the instruction of a participant who is tasked to provide input on their beliefs in the explanations and perform the prediction task.



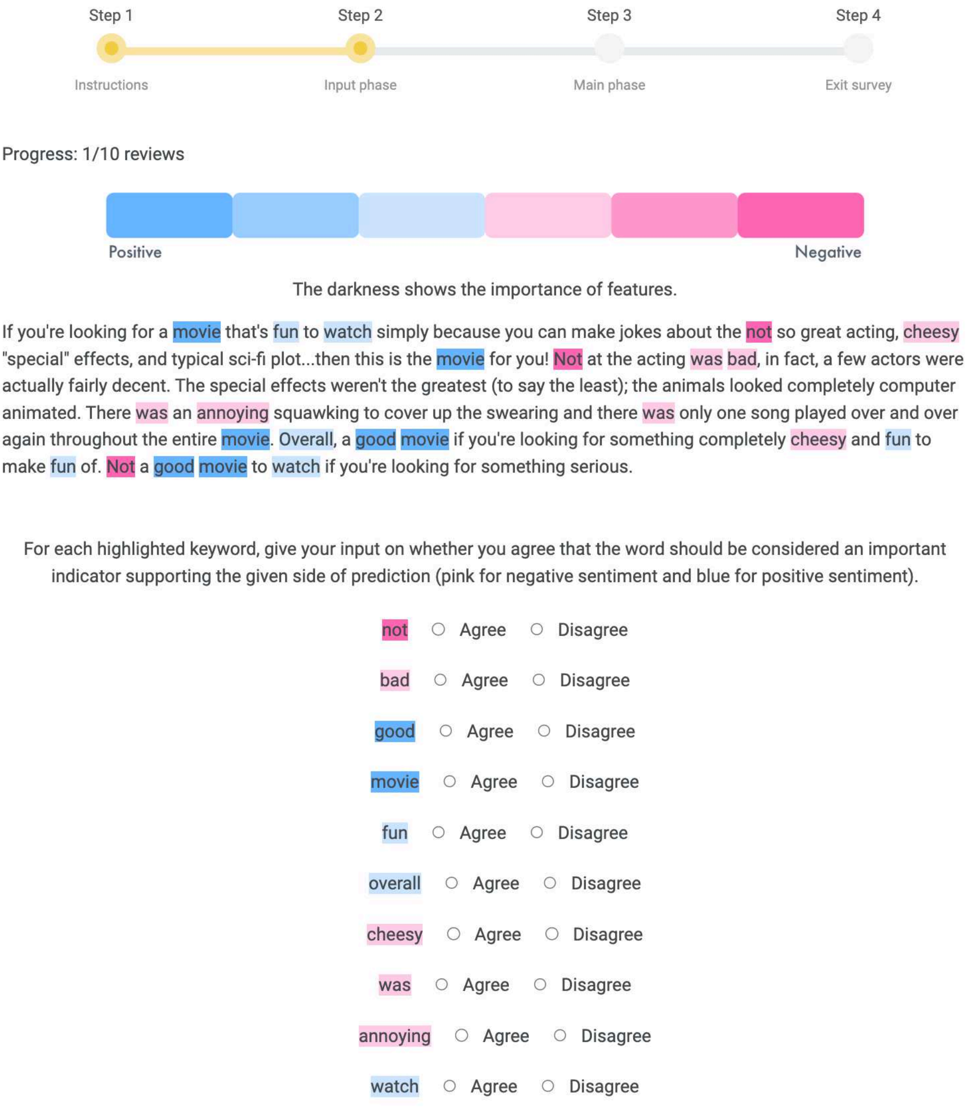
Progress: 1/10 reviews

If you're looking for a movie that's fun to watch simply because you can make jokes about the not so great acting, cheesy "special" effects, and typical sci-fi plot...then this is the movie for you! Not at the acting was bad, in fact, a few actors were actually fairly decent. The special effects weren't the greatest (to say the least); the animals looked completely computer animated. There was an annoying squawking to cover up the swearing and there was only one song played over and over again throughout the entire movie. Overall, a good movie if you're looking for something completely cheesy and fun to make fun of. Not a good movie to watch if you're looking for something serious.

Judge the sentiment of this movie review

Which word(s) contributed to, or you considered as important indicators for, your judgment?
Please separate the words by a comma. E.g., word1, word2, word3.
You should only pick **single words** instead of multi-words phrases,
and only words that appeared in the original sentence.

Fig. 13. There are two ways of providing input. This figure shows the interface of selective explanations with open-ended input (Open-ended).



Progress: 1/10 reviews

The darkness shows the importance of features.

If you're looking for a movie that's fun to watch simply because you can make jokes about the not so great acting, cheesy "special" effects, and typical sci-fi plot...then this is the movie for you! Not at the acting was bad, in fact, a few actors were actually fairly decent. The special effects weren't the greatest (to say the least); the animals looked completely computer animated. There was an annoying squawking to cover up the swearing and there was only one song played over and over again throughout the entire movie. Overall, a good movie if you're looking for something completely cheesy and fun to make fun of. Not a good movie to watch if you're looking for something serious.

For each highlighted keyword, give your input on whether you agree that the word should be considered an important indicator supporting the given side of prediction (pink for negative sentiment and blue for positive sentiment).

not ○ Agree ○ Disagree

bad ○ Agree ○ Disagree

good ○ Agree ○ Disagree

movie ○ Agree ○ Disagree

fun ○ Agree ○ Disagree

overall ○ Agree ○ Disagree

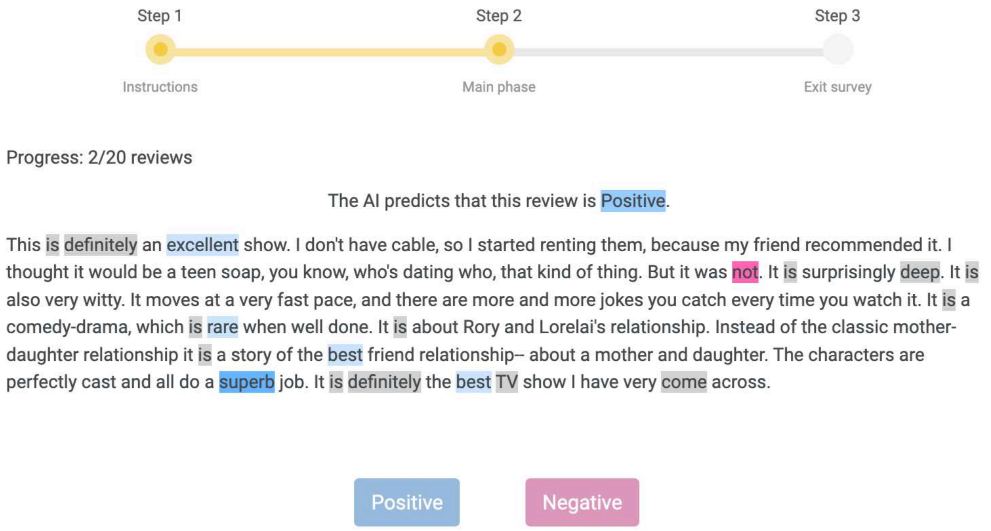
cheesy ○ Agree ○ Disagree

was ○ Agree ○ Disagree

annoying ○ Agree ○ Disagree

watch ○ Agree ○ Disagree

Fig. 14. There are two ways of providing input. This figure shows the interface of selective explanations with model explanation critiques (Critique-based).

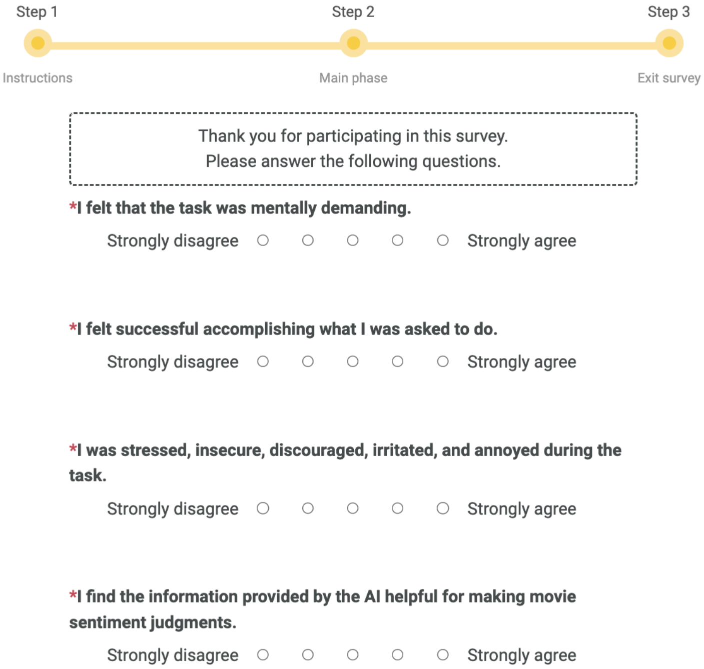


Progress: 2/20 reviews

The AI predicts that this review is Positive.

This is definitely an excellent show. I don't have cable, so I started renting them, because my friend recommended it. I thought it would be a teen soap, you know, who's dating who, that kind of thing. But it was not. It is surprisingly deep. It is also very witty. It moves at a very fast pace, and there are more and more jokes you catch every time you watch it. It is a comedy-drama, which is rare when well done. It is about Rory and Lorelai's relationship. Instead of the classic mother-daughter relationship it is a story of the best friend relationship-- about a mother and daughter. The characters are perfectly cast and all do a superb job. It is definitely the best TV show I have very come across.

Fig. 15. This figure shows the interface of the prediction task. Selective explanations are grayed out and are predicted to be misaligned with what the user would consider as relevant for judging review sentiment.



Thank you for participating in this survey.
Please answer the following questions.

***I felt that the task was mentally demanding.**

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

***I felt successful accomplishing what I was asked to do.**

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

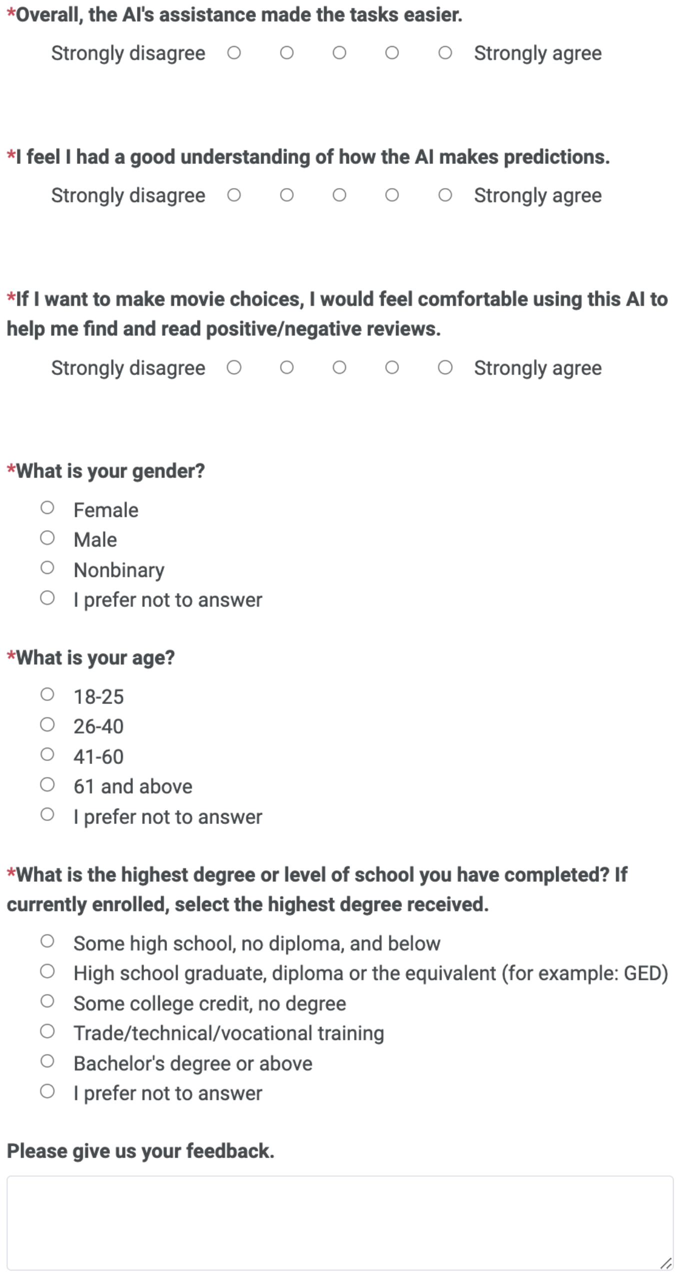
***I was stressed, insecure, discouraged, irritated, and annoyed during the task.**

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

***I find the information provided by the AI helpful for making movie sentiment judgments.**

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

Fig. 16. Due to length constraints, this figure shows the first part of the survey which features some of the subjective questions.



*Overall, the AI's assistance made the tasks easier.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

*I feel I had a good understanding of how the AI makes predictions.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

*If I want to make movie choices, I would feel comfortable using this AI to help me find and read positive/negative reviews.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

*What is your gender?

○ Female
○ Male
○ Nonbinary
○ I prefer not to answer

*What is your age?

○ 18-25
○ 26-40
○ 41-60
○ 61 and above
○ I prefer not to answer

*What is the highest degree or level of school you have completed? If currently enrolled, select the highest degree received.

○ Some high school, no diploma, and below
○ High school graduate, diploma or the equivalent (for example: GED)
○ Some college credit, no degree
○ Trade/technical/vocational training
○ Bachelor's degree or above
○ I prefer not to answer

Please give us your feedback.

Fig. 17. This figure shows the second and remaining part of the survey which features more subjective questions and demographic questions.