Risks and Opportunities in Human-Machine Teaming in Operationalizing Machine Learning Target Variables

MENGTIAN GUO, University of North Carolina at Chapel Hill, USA
DAVID GOTZ, University of North Carolina at Chapel Hill, USA
YUE WANG, University of North Carolina at Chapel Hill, USA

Predictive modeling has the potential to enhance human decision-making. However, many predictive models fail in practice due to problematic problem formulation in cases where the prediction target is an abstract concept or construct, and practitioners need to define an appropriate target variable as a proxy to operationalize the construct of interest. Selecting an appropriate proxy target variable is a challenging process in practice, requiring both domain knowledge and iterative data modeling. While emerging prototyping tools promise to accelerate this process, it remains unclear how rapid iterations influence human judgment in problem formulation. In this work, we conducted a controlled user study ($N = 48$) to investigate the impact of human-machine teaming on proxy target selection. We instantiate a system offering three recommendation strategies: Relevance-First (prioritizing conceptual alignment), Performance-First (prioritizing model performance), and Pareto-Front (considering both). We find that while rapid iterations can significantly improve exploration efficiency, they also tend to amplify a "*performance bias*": the tendency to favor well-performing proxy targets even when they are not aligned with the modeling goal. However, systems that explicitly estimate and communicate the relevance of proxy targets can mitigate this bias. Our study highlights the risks and opportunities of human-machine teaming in operationalizing machine learning target variables, yielding insights for future research to explore the opportunities and mitigate the risks.

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

Additional Key Words and Phrases: Algorithmic Decision Support, Human-Machine Teaming, Performance Bias, Proxy Target Selection, Problem Formulation, Construct Validity

ACM Reference Format:
Mengtian Guo, David Gotz, and Yue Wang. 2026. Risks and Opportunities in Human-Machine Teaming in Operationalizing Machine Learning Target Variables. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3805689.3806512

1 Introduction

Predictive models aim to forecast individuals' future outcomes from historical data. When properly designed, they can assist human decision-making by considering more factors than a single decision-maker typically can and help reduce subjectivity and increase consistency. However, many predictive machine learning (ML) applications fail to support decision-makers in real settings, and some have had negative social impacts. While multiple factors contribute to these failures, prior work pointed out that problematic problem formulation and model specification are recurrent reasons causing failed predictive ML applications [20, 38, 48, 52].

Authors' Contact Information: Mengtian Guo, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, mtguo@unc.edu; David Gotz, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, gotz@unc.edu; Yue Wang, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, wangyue@unc.edu.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3806512

Problem formulation determines *what* is predicted and *how*. It's a critical first step in the predictive model development process that shapes how the model will later be interpreted and used. In fields such as medical treatment, child welfare, and criminal justice, problematic formulations have been documented, leading to disuse, misinterpretation, and systematic decision errors [5, 15, 26, 37]. Case studies, contextual inquiries, and human-AI interaction literature have surfaced several recurring challenges in problem formulation.

Large space of candidate problem formulations and long turnaround between iterations: Real-world datasets are often high-dimensional, and many variables (and composites) can plausibly approximate the modeling goal. Practitioners need to consider multiple potential target variables and framings—a need discussed in various ML applications [5, 40, 50]. However, navigating this space is constrained by the long turnaround time of model development. The implications of target variable choices made at the problem formulation stage may surface only after the time-consuming model training and evaluation, when the impact of hidden issues (e.g., label noise, missing data, and selection bias) becomes visible [37, 50]. These slow iteration cycles force practitioners to rely on intuition rather than systematic search, limiting exploration to a handful of options that may not be optimal.

Multiple objectives when defining problem formulations: Even when candidates are identified, practitioners face the challenge of optimizing multiple conflicting objectives: quantitative metrics (e.g., accuracy, efficiency) and qualitative values (e.g., construct relevance, fairness). At the problem formulation stage, the definition of the prediction target can influence which objectives can be achieved [8, 13, 36]. For instance, in sepsis prediction, different temporal framing structures (e.g., earliness) can shift class imbalance ratio and missing value rate, yielding widely different AUPRC scores [28]. Compared to quantitative metrics, qualitative objectives are hard to optimize during problem formulation. Using target variables with low relevance to the modeling goal can lead to predictions that misalign with intended use and even lead to ethical concerns [5, 26].

Human-machine teaming is a promising approach to solving the challenges in machine learning problem formulation, yet it presents a fundamental sociotechnical risk. In many domains, defining an appropriate target variable is a "wicked problem" [32], where the construct of interest is unobservable, and practitioners must select a proxy target based on contextual knowledge. These choices cannot be fully automated since human judgments remain indispensable in determining which target best reflects the construct of interest, foreseeing ethical consequences in deployment, and interpreting tradeoffs between competing values. However, the emergence of advanced systems—including AutoML techniques and LLM-based coding tools—introduces new capabilities that could reshape this dynamic [2, 3, 21, 54]. By lowering the technical barriers to prototyping, these tools offer the promise of broadening exploration: practitioners can potentially test and evaluate a wider range of problem definitions in the time it previously took to implement just one. This efficiency could lead to better-informed formulations.

However, this increased efficiency brings a critical open question on the quality of the resulting formulations. While research on interactive formulation tools has demonstrated their ability to accelerate iteration, we lack a quantitative understanding of how this rapid feedback loop influences human judgment [10, 46]. Specifically, as these tools make quantitative metrics immediately visible and easy to optimize, they create a potential tension with the qualitative goal, such as construct validity. Does the ability to quickly "solve" a prediction problem help users identify the most useful formulation, or do the high performance scores drive users toward targets that are easier to predict but less relevant to the real-world goal? Furthermore, how do specific system conditions, such as recommendations prioritizing performance versus those surfacing relevance or trade-offs, influence the user's exploration behavior and final decision-making? As automation lowers the cost of trial and error, it is crucial to know how these design choices shape a practitioner's path through the problem space.

To address this gap, we conducted a controlled quantitative study ($N = 48$) investigating how human-machine teaming influences problem formulation quality. We designed experimental scenarios that simulated the formulation process as a multi-objective decision problem, instantiating human-machine teaming through three recommendation strategies: relevance-based, performance-based, and Pareto-based (considering both). We found

that while automation indeed improved efficiency, it introduced a salient "*performance bias*", a phenomenon where human decisions drifted toward high-performing formulations even when they were misaligned with the modeling goal. Our results demonstrate that without explicit guardrails, the efficiency of modern tools may come at the cost of the resulting model's validity.

The contributions of this study can be summarized as follows.

1. We provided quantitative evidence of how human-machine teaming can influence the quality and efficiency of problem formulation. We demonstrate that while automation facilitates rapid prototyping and increases the volume of candidates explored, the design of the recommendation mechanism influences the quality of the outcome. We show that systems surfacing "Relevance" (construct validity) or Pareto Front (trade-offs) can guide users toward formulations that align with real-world goals, whereas performance-first and unguided exploration lead to misalignment.
2. We observed "performance bias", the behavior where practitioners prioritize quantitative model metrics over conceptual alignment. Our findings reveal that without explicit signals for relevance, users drift toward high-performing but invalid formulations, even when users freely explore without recommendations.
3. Our study generated various design insights for iterative formulation tools. We argue that iteration speed alone is insufficient. Systems should explicitly support the evaluation of user objectives and facilitate multi-criteria decision-making. Systems should provide recommendations that jointly consider all objectives to reduce the efforts required to examine different candidates.

2 Related Works

2.1 Problem Formulation for ML Applications

Predictive modeling has been leveraged in a wide range of applications to support human decision-making, including education [29], medical treatment [28], hiring [51], child welfare [50], and criminal justice [5]. However, issues of problem formulation, also called model specification, have been a major obstacle to the adoption of predictive models. For example, Lauritsen et al. [28] showed that alternative formulations for Sepsis prediction meaningfully change clinical usefulness. In education, experts critique that problem formulations often rely on narrow quantitative metrics misaligned with multi-faceted educational goals [29]. In child welfare, researchers have pointed out that there is often a misalignment between what the models were trained to predict and the social workers' priorities, resulting in low model utility [26, 45].

The issue of problem formulation has also been discussed in the AI fairness literature. For instance, researchers discovered that using re-arrest as a prediction target can result in racial bias in the recidivism prediction algorithm, as the selected outcome variable can reflect systemic biases in policing, arrests, and sentencing [5]. Obermeyer et al. [37] demonstrated the bias in an algorithm used to predict complex health needs due to problematic problem formulations. Future cost was used as the prediction target, ignoring its relation to income, which incorporates disparities in employment and salary. Tal [48] conceptualizes all predictive targets as inherently imperfect approximations, shaped by the complex negotiations and compromises that occur between data scientists and domain experts [38].

2.2 Challenges of Problem Formulation

The challenges of the problem formulation stage have been illustrated by case studies and contextual inquiries [35, 55]. Below, we review and summarize different aspects of this challenge.

Understanding ML and data constraints (performance): Understanding data availability and limitations are essential for developing solutions that fit the application context [25, 34]. Several factors can render a problem formulation unsolvable with an ML model, including limited data size [38], class imbalance [28], missing values [49], and label noise [36]. Data availability and machine learning capability limitations are

frequently identified during initial data analysis or model prototyping [38, 50]. Practitioners typically perform lightweight data analysis and experiments using existing data and simple models to assess the feasibility of the problem formulation [23, 56]. When a problem formulation is found to be unsolvable, practitioners commonly respond by collecting more data or exploring alternative problem formulations that may have high-quality data available [28, 38, 50].

Asking the right question (relevance): Problem formulations that are misaligned with the actual goal of the application can cause unexpected outcomes and biases [5, 11, 20, 31, 37, 38]. Recent work proposed normative perspectives on ML problem formulation [12, 41, 52] and applied modern validity frameworks to make the misalignment more detectable and actionable [12, 20, 43]. For instance, Guerdan et al. [20] proposed a framework that summarizes different sources of problem formulation validity issues, e.g., measurement error, intervention effects, and selection bias. Informed by statistics and quantitative social sciences methods, the authors suggest that problem validity can be evaluated through construct reliability, construct validity, etc. Despite involving quantitative analysis, the methods for validity evaluation all require domain knowledge and subjective human judgments.

Iterative prototyping and testing: Problem formulation is widely recognized as an iterative process, requiring repeated refinement based on data exploration and evaluation in ML workflows such as CRISP-DM, KDD, and TDSP [1, 17, 22, 27, 30, 33, 47]. Practitioners are often faced with multiple problem formulation options [15, 28, 50, 53]. Prototyping and testing help people realize the limitations of the current problem formulation, leading to reformulations of the problem. However, due to the complexity of data and uncertainty of ML outcomes, ML prototyping is time-consuming, often involving a labor-intensive inner-loop of ML development (i.e., data preprocessing, model training, hyperparameter optimization, and model selection). The prototyping process slows down the feedback loop, which in turn slows down problem (re)formulation. Consequently, practitioners typically have time to explore only one or a few problem formulations, limiting their ability to evaluate and compare all potential options [37].

2.3 Tools Designed to Support Problem Formulation

While there is a plethora of tools and systems that support developing and refining ML models, most of them support model development rather than problem formulation itself. There exist human-guided machine learning systems that support rapid construction of machine learning pipelines through an interactive interface without writing code [9, 14, 24, 39, 44]. However, these systems focus on the entire ML pipeline instead of the problem formulation stage. AutoML assumes that the ML problem has already been well-defined and aims to automate the inner loop of ML development (i.e., data preprocessing, model training, hyperparameter optimization, and model selection) [7, 16, 18]. Although AutoML itself does not produce problem formulations, it standardizes the ML inner loop and enables rapid prototyping, and thus can be a useful component in accelerating the problem formulation process.

Two systems were explicitly designed to support problem formulation. Tempo [46] supports iterative problem formulation by enabling quick model prototyping and subgroup-based model evaluation. Cashman et al. [10] proposed the Exploratory Model Analysis (EMA) system, which facilitates the user to discover meaningful problems to solve on a given dataset. EMA automatically experiments on all potential proxy targets in a dataset and presents them to the user to support proxy selection. This system implicitly influences users' proxy selection through predictive performance. Both works conduct qualitative evaluation through case studies; in contrast, our work conduct quantitative evaluation on how rapid prototyping affects practitioners' proxy target selection.

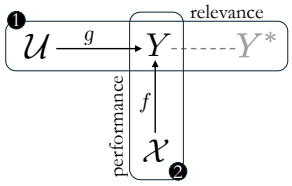


Fig. 1. Concept relationships in the proxy target selection problem. The function g uses observed outcomes $\mathcal{U}$ to construct the proxy target Y, which is a surrogate of the unobserved target outcome Y^* (Box 1). The machine learning model f uses predictors $\mathcal{X}$ to predict the proxy target Y (Box 2). The problem is to construct Y that is both relevant to Y^* and can be accurately predicted using $\mathcal{X}$.

3 Proxy Target Selection in Problem Formulation

3.1 Concept Definition

We study the core activity in ML problem formulation: selecting a proxy target variable to approximate a modeling goal. We define the *Target Outcome* (Y^*) as the theoretical construct of interest (e.g., "having long COVID"), which is conceived by stakeholders but unobserved in the task-specific dataset. In contrast, the *Observed Outcomes* ($\mathcal{U} = \{U_1, ..., U_m\}$) are the measurable variables pre-identified in the dataset (e.g., ICD codes or specific survey responses). A *Proxy Target* (Y) is a function of observed outcomes, $Y := g(\mathcal{U})$, constructed to serve as a quantifiable surrogate for Y^* based on domain knowledge. Finally, *Predictors* ($\mathcal{X}$) are the observed features used by a machine learning model f to predict this proxy: $Y \leftarrow f(\mathcal{X})$. Therefore, the *Proxy Target Selection* problem is to construct a proxy Y that is both semantically relevant to the unobserved Y^* and reasonably predictable by $\mathcal{X}$.

3.2 Problem Formulation as a Multi-Objective Decision-making Task

The relationships between concepts in the problem are illustrated in Figure 1. The function g in Box 1 (the horizontal rectangle) plays a central role as it produces the proxy target Y. g cannot be trained or evaluated using data because Y^* is unobserved. This is in contrast with the function f in Box 2 (the vertical rectangle), which can be trained and evaluated using data because Y is observed in data. We considered two aspects when evaluating g (and therefore the proxy target Y): *relevance* and *performance*. The relevance aspect (whether the proxy target faithfully represents the target outcome) has to be judged based on domain knowledge in the application context, which is better suited as a human task. The performance aspect (whether the proxy target can be accurately predicted using a set of predictors) can be evaluated using standard supervised machine learning training and evaluation procedures, which is better suited as a machine task (e.g., through AutoML).

In many real-world tasks, the relevance and performance objectives can conflict, and the most "relevant" formulation from a conceptual standpoint may not be the one that yields the most "useful" model. For instance, in early sepsis prediction [28], clinicians ideally want the earliest prediction — predicting sepsis 12 hours before onset is more relevant than predicting sepsis 6 or 3 hours before onset. However, early prediction may result in low model performance due to sparse signals. By adjusting the temporal framing to predict sepsis 6 hours or 3 hours before onset, we reduce conceptual relevance, but may be able to train a model that produces reliable predictions, which ends up having overall greater clinical utility. Thus, proxy target selection is inherently **multi-objective**: the relevance and performance objectives need to be jointly considered.

In practice, practitioners often approach the problem formulation task in a relevance-first strategy: they would first test problem formulations that are relevant to the application goal. However, a significant challenge in ML deployment is that the most "desirable" target variable may fail to yield a "feasible" model due to data sparsity or low signal-to-noise ratio [28, 38]. This creates the fundamental motivation for practitioners to adjust to another

relevant problem formulation. Since the performance information is not known ahead of time, it may take some trial and error to identify a problem formulation that is both relevant and yields well-performing [15, 50, 53]. Without automated support, practitioners may settle for suboptimal formulations due to the slow model iteration cycle.

Machines can help improve the efficiency of identifying problem formulations that satisfy both objectives by automatically evaluating the model's performance resulting from a problem formulation. When a practitioner constructs a problem formulation, machines can quickly evaluate the resulting model's performance to help the practitioner decide whether this is a good candidate (Baseline). In addition, machines can also improve efficiency by recommending potentially good candidates to practitioners, which reduces trial and error. In this study, we considered three recommendation criteria, each represents a different subset of objectives considered.

- **Relevance First**: In this condition, the system recommends proxies that are most likely to be relevant, i.e., semantically most relevant to the construct of interest.
- **Performance First**: In this condition, the system recommends proxies that lead to the highest model performance.
- **Pareto Front**: The system generates recommendations by prioritizing proxies on the Pareto front of relevance and performance. These recommended candidates dominate other proxies, ensuring both objectives are considered.

4 Experimental Evaluation

We designed our study to answer the following research questions:

RQ1 (Quality): How do different recommendation criteria influence the performance and relevance of the selected proxy target variable due to the change in people's decision-making?

RQ2 (Efficiency): How do different recommendation criteria influence people's efficiency in proxy selection?

RQ3 (User Perception): How do different recommendation criteria influence the perceived usefulness of the recommendation and users' satisfaction, confidence, and the perceived easiness of the proxy selection task?

4.1 Tasks

We aimed to design problem formulation tasks for realistic application scenarios that can be completed in a controlled study environment by a large pool of participants who have sufficient domain expertise. Since it is challenging to recruit a large number of experts from highly specialized domains, we designed application scenarios that student participants could meaningfully engage in. We used a survey dataset that examined the impact of COVID-19 on college students [4]. This dataset naturally contains features and a diverse set of outcome variables (see details in Appendix A).

Under the first application scenario, participants were tasked to develop a binary classification model to understand how many students' academic performance were negatively impacted by COVID-19. The target outcome is thus a binary variable Y_1^* = "whether a student's academic performance is negatively impacted by COVID-19." For the second application scenario, the target outcome was a binary variable Y_2^* = "whether a student's mental health is negatively impacted by COVID-19." This domain is directly relevant to our participants and provides a high-stakes decision context that allows them to understand both the model's purpose and the implications of choosing different formulations. While our participants were not professionally trained in this domain (e.g., school administrators), they possessed experience regarding the application scenarios, and thus could assess the semantic relevance of a proxy variable. This helps mitigate the gap between student participants and professional decision-makers.

Under each application scenario, participants were provided with a simplified version of the survey dataset with a fixed set of features ($\mathcal{X}$) and a set of 10 binary outcome variables ($\mathcal{U}$) with various relevance to the application

scenario and resulting model performance. A proxy variable Y involves a subset of the outcome variables $\mathcal{V} \subseteq \mathcal{U}$, $\mathcal{V} = \{V_1, \cdots V_k\}$. The function g combines the subset outcome variables using logic operators (e.g., OR, AND, NOT), resulting in binary proxy variables. To further simplify the task, we restricted participants to include at most two outcome variables and only combine them by OR. This results in $\binom{10}{2} = 55$ candidate proxy targets. We curated the two sets of outcome variables such that there is a tradeoff between relevance to the application scenario and the resulting model's performance.

Participants were asked to experiment with different proxy targets and select one that they think is the most applicable in practice based on both relevance and model sensitivity. Participants were explained with examples that an effective proxy target should both align with the application's objective and produce a model sensitive enough to identify the target group. We didn't provide an explicit optimization goal for participants since there is no standard answer as to which objective is more important. We expected a natural difference in interpretations of the relative importance between objectives among participants. We aimed to capture the impact of different conditions on participants' decision-making with their own interpretation rather than defining the criteria for optimization. The original description of the tasks is provided in Appendix A.

4.2 System Design

We designed a user interface to surface problem formulation recommendations and facilitate the problem formulation task. The interface contains three major components. The left panel of the interface shows the recommended proxies in a ranked list (see Figure 2 (a)). The ranking order varies between conditions. Under the Baseline condition, the left panel is empty, so no recommendations were provided. The right panel of the interface provides the basic components that allow users to construct proxies (see Figure 2 (b)). It displays all the available outcome variables and provides the function for users to manually construct proxies and train models. The proxy detail panel (see Figure 2 (c)) provides details of the selected proxy target, either through the proxy recommendation view or manually constructed by the user. The specific design choices are made based on the following principles: 1) include visual encoding and features that are standard for multi-attribute decision tasks, and 2) keep the design as comparable as possible through a consistent presentation of proxy target variables and objectives, a consistent visual design, and a consistent interaction design across conditions.

4.2.1 Objective Quantification. **Performance.** Given the labels Y derived from a proxy target variable, the system trains and evaluates the prediction model f with train-test split. We selected sensitivity (recall) as the primary metric because it aligns with the application goal of identifying as many target cases (e.g., at-risk students) as possible, while remaining intuitively interpretable for participants. We trained and evaluated a logistic regression model because of its efficiency and decent predictive performance. However, the model training process can be extended to other ML models and AutoML techniques.

Relevance. In order to surface potentially relevant candidates, we used a language model to calculate the semantic similarity between the proxy target variable and the target outcome to estimate their relevance. The description d of a proxy target is generated by combining the included variables' names using the selected logic operators (e.g. "feel worried about mental health OR feel angry while attending classes"). Given the description d of a proxy and a target outcome description c (e.g. "Student experience mental health issues"), the system embeds d and c into vectors (v_d, v_c) using a Sentence Transformer Model (SBERT) [42]. Specifically, we used a pre-trained model.[1] The relevance score of the proxy to the target outcome is calculated as the cosine similarity of their embeddings. We emphasize that the calculated relevance score acts as a quantitative approximation to assist humans in navigation and to provide guardrails for judgment instead of replacing human judgment. During the study, we clearly stated to the participants that they should rely on their own judgment of proxy relevance.

[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

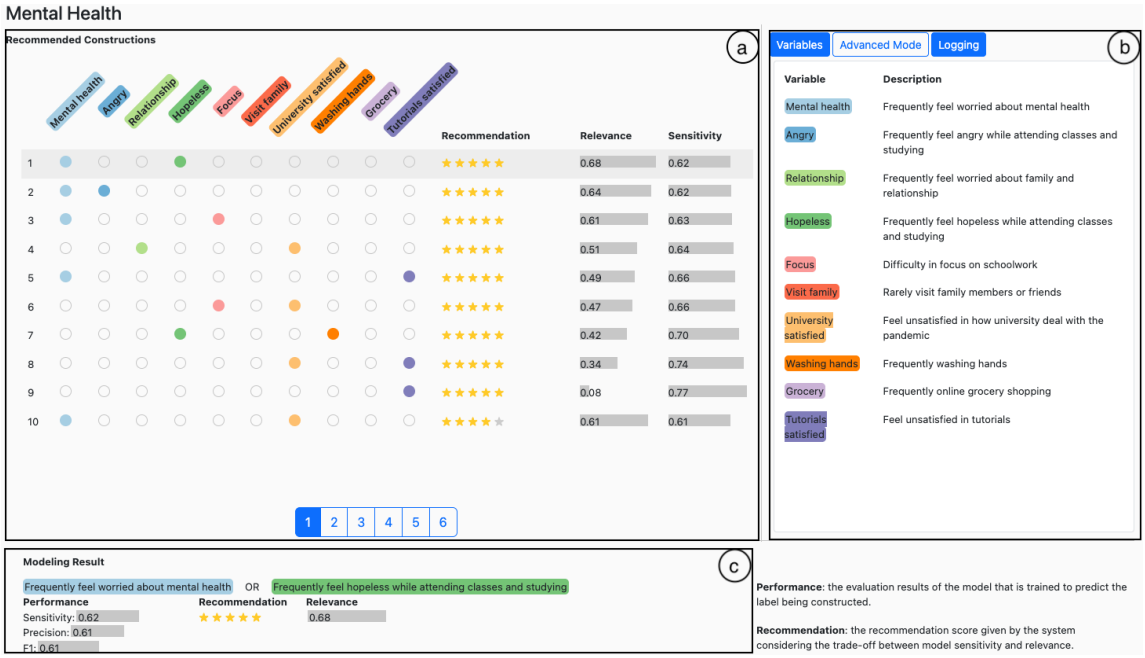


Fig. 2. Interface for the Pareto Front condition.

To validate this automated approach, we first conducted formal alignment study. We recruited four student participants with a mix of undergraduate and graduate levels to ensure a breadth of perspectives on academic and mental health contexts. They were briefed on the study's goals but remained blind to the model's scores. They then performed pairwise relevance assessments on 30 randomly selected proxy target pairs per scenario, evaluating which candidate in a pair was more conceptually relevant to the outcome. We compared these human preferences against the model's relevance scores. For the topic of "Mental Health", the average Cohen's Kappa between humans and the embedding model was 0.442, closely matching the average inter-human agreement of 0.454. For the topic of "Academic Performance", the average Cohen's Kappa between humans and the embedding model reaches 0.643, which is higher than the inter-human agreement of 0.491.

In addition, during the study, we collected relevance judgments from each participant on 30 randomly sampled pairs of proxies and compared them with the relevance scores. The average Cohen's Kappa between participants and the relevance scores is 0.425 on the topic of "Mental Health" and 0.528 on the topic of "Academic Performance". These results suggest that the embedding model aligns with human judgments to a similar degree as humans align with each other.

4.2.2 Proxy Recommendation View. Given a list of outcome variables that can be used to construct proxies, the system automatically enumerates all possible proxies, considering all combinations of outcome variables. Users can select a candidate from the list, and the details of the proxy target variable are shown in the proxy detail panel. The set of candidates is presented in a matrix view to facilitate the understanding and comparison of different proxies. Each variable is associated with a unique color to help users quickly interpret a proxy target.

The relevance and performance of each candidate are encoded by the length of the bars, a common feature in tabular visualizations for decision support [19]. In the Relevance First condition, only the relevance level is

Sub-session 1	Sub-session 2	Num.	Sub-session 1	Sub-session 2	Num.
Baseline	Relevance First	4	Relevance First	Baseline	4
Performance First	Pareto Front	4	Pareto Front	Performance First	4
Baseline	Performance First	4	Performance First	Baseline	4
Relevance First	Pareto Front	4	Pareto Front	Relevance First	4
Baseline	Pareto Front	4	Pareto Front	Baseline	4
Relevance First	Performance First	4	Performance First	Relevance First	4

Table 1. Experimental design with counterbalancing. Participants were split into 12 groups, corresponding to 12 permutations of two out of four interface conditions. Among the four participants assigned to each group, two worked on the mental health task in sub-session 1, and the other two worked on the academic performance task in sub-session 1.

displayed to explain the order of the ranking. Similarly, in the Performance First condition, only the performance information is shown. In the Pareto Front condition, both relevance and performance are presented. However, as a single objective is insufficient to explain the ranking or distinguish between non-domination fronts, the Pareto-optimality is encoded using stars. The candidates in the first non-domination front receive five stars, and the number of stars decreases progressively until the fourth front.

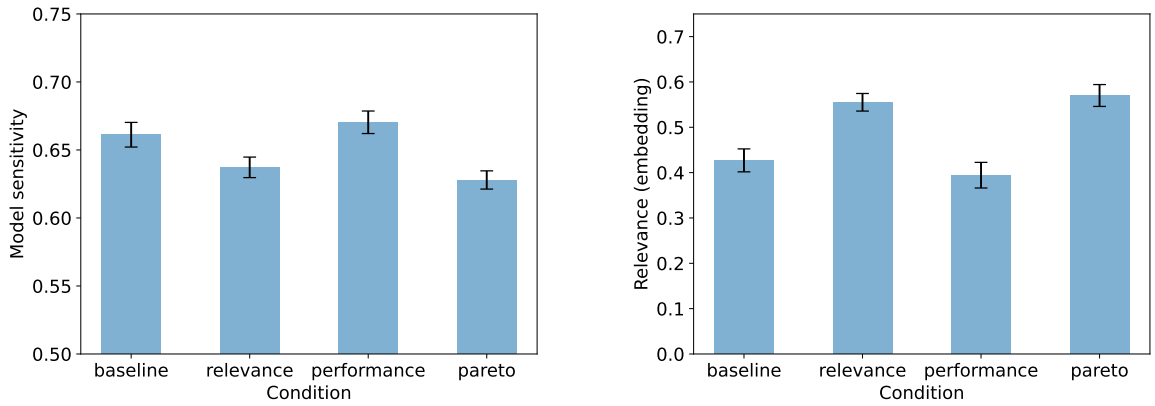
Under Relevance First, candidates are ranked by their relevance to the target outcome, where highly relevant proxies are shown at the top. Under Performance First, proxies are ranked by their performance, and well-performing proxies are ranked at the top. Under Pareto Front, candidates are not ranked based on a single objective. Instead, the system first organizes all candidates into multiple non-domination fronts, and the first front (containing the Pareto-optimal candidates) is ranked at the top. To maintain consistency between different conditions, the non-domination fronts are presented in a list format, where candidates in the same non-domination front are ranked by relevance. We presented all candidates based on non-domination fronts since this does not carry any assumption regarding the relative importance of the two objectives.

4.2.3 Proxy Detail View. In Proxy Detail View, we provided detailed performance metrics to facilitate the user in judging the quality of a proxy target. In the Baseline and Performance First conditions, only the model performance metrics are displayed. In the Relevance First condition, model performance metrics and the relevance score are displayed. In the Pareto Front condition, the recommendation level is displayed in addition to model performance and relevance score.

4.3 Study Overview

There are in total four interface conditions. We adopted a study design where each participant works on two tasks, each with a different condition. We adopted this design since the within-subjects components allow participants to serve as their own baseline, effectively controlling for individual differences. At the same time, it reduces cognitive fatigue since each task requires significant mental effort. $N = 48$ participants were randomly assigned to the 12 condition permutations as explained in Table 1. As a result, each of the four interface conditions was used by 24 participants. We strictly counter-balanced the interface conditions and tasks (application scenarios) to mitigate learning effects and topic bias. We used the Kruskal-Wallis test and Dunn's post-hoc tests for a high-level cross-condition comparison. To address the dependencies inherent in this mixed design, we utilized linear mixed-effects models (LMEMs) for our analysis [6]. By treating Participant ID as a random effect, we effectively controlled for individual-level variance, ensuring that our findings regarding system influence are statistically robust and account for the nested nature of the data.

Each study session lasted roughly one hour. After providing informed consent, the study coordinator provided detailed information about the application context, dataset, and proxy syntax. A study session contained two



(a) Resulting models' sensitivity.

(b) Resulting models' relevance.

Fig. 3. Average model sensitivity and relevance to the application scenario of participants' final proxy selections for each condition. Significant differences exist between conditions on model sensitivity and relevance (more details in Section 5.1).

sub-sessions. Using an example scenario, each sub-session began with a hands-on tutorial. Participants were then informed of the application scenario to work on. Participants would first finish the relevance judgment task. Then, participants would complete the proxy selection task. To encourage the participants to take the task seriously, the study coordinator would ask them to verbalize their reasoning and justify their final choices after each sub-session. Participants would also finish a questionnaire after each sub-session to report their experience. Participants could spend at most 15 minutes on each task. Finally, after completing the two sub-sessions, participants provided additional feedback about their experience using the tools through an exit interview.

The study sessions were conducted either in-person or online through Zoom with a study coordinator. All participants followed the same scripted tutorial, utilized the same web interface, and completed identical questionnaires. Furthermore, the study coordinator monitored all sessions in real-time to ensure consistent engagement.

This study was exempt from review by our university's Institutional Review Board (IRB) according to the category: Survey, interview, public observation; Benign Behavioral intervention, under 45 CFR 46.104. All participants provided informed consent and were compensated with a $20 Amazon gift card for their time.

4.4 Participants

We recruited $N = 48$ participants from STEM majors via campus-wide mailing lists. To participate, individuals were required to have completed at least a machine learning course or tutorial, as prior ML knowledge was necessary for the task. The sample included 28 males, 18 females, and 2 participants who chose not to disclose their gender. Among them, 33 had prior experience applying ML to real-world problems. Participants came from diverse educational backgrounds, with the most common fields of study being Information Science (12), Computer Science (7), Biostatistics (4), and Health Informatics (4).

5 Results

5.1 RQ1: Quality

We analyzed how different system conditions influence the quality of the proxy target selected by participants. The results are shown in Figure 3. Kruskal-Wallis test shows a significant difference between conditions in proxies' relevance ($p < 0.001$) and resulting model's sensitivity ($p < 0.01$). Dunn's post-hoc test showed that participants using Performance First or Baseline selected proxies with significantly higher sensitivity ($p < 0.03$) but lower relevance compared to those in the Relevance First and Pareto Front conditions ($p < 0.01$). For model sensitivity, Dunn's post-hoc test shows a significant difference between the Baseline and the Pareto Front condition ($p = 0.015$, Hedges' $g = 0.838$), Relevance First and Performance First ($p = 0.021$, Hedges' $g = -0.84$), Performance First and Pareto Front ($p = 0.001$, Hedges' $g = 1.13$). For proxies' relevance, Dunn's post-hoc test shows a significant difference between the Baseline and Relevance First ($p = 0.007$, Hedges' $g = -1.14$), Baseline and Pareto Front ($p = 0.0004$, Hedges' $g = -1.16$), Relevance First and Performance First ($p = 0.001$, Hedges' $g = 1.33$), Performance First and Pareto Front ($p = 0.00004$, Hedges' $g = -1.34$). Note that the relevance score is calculated using the same text embedding model as in the system recommendation.

As participants may disagree with the embedding model, we fitted a Bradley-Terry model to generate a relevance score for each proxy target given participants' relevance judgments. We observed the same pattern as using the relevance score given by the embedding model. In summary, when participants did free trial and error or were provided with recommendations based on model performance, they were more likely to select proxy targets that had higher model performance but were less relevant to the modeling goal. More detailed analysis can be found in Appendix B.

We applied a mixed linear-effects model to account for the dependencies inherent in the mixed study design. The fixed effects include sub-session number (i.e., first or second sub-session), interface condition, and the application scenario or topic (i.e., mental health or academic performance). The random effect is the participant ID, which accounts for the fact that each individual will perform similarly in the two sessions (i.e., use a similar amount of time and put similar weights on relevance and performance). We fitted four mixed linear models, one for the time spent on the task, one for the resulting model's sensitivity, one for the resulting model's relevance based on the embedding model, and one for the resulting model's relevance based on the relevance judgments. We applied a log transformation to the time spent on the task to meet the assumptions of the model.

For model sensitivity, there is a significant difference between the Baseline and the Pareto Front ($p = 0.001$ based on Wald test), between the Baseline and the Relevance First ($p = 0.007$), between the Relevance First and the Performance First ($p < 0.001$), and between the Performance First and the Pareto Front ($p < 0.001$). For proxies' relevance, there is a significant difference between the Baseline and the Pareto Front ($p < 0.001$), between the Baseline and the Relevance First ($p < 0.001$), between the Relevance First and the Performance First ($p < 0.001$), and between the Performance First and the Pareto Front ($p < 0.001$). We found that there was a significant effect of session ID on the time spent on the task. Participants spent less time in the second session, which implies a learning effect ($p = 0.002$). However, the session ID does not have any effect on the quality of the resulting selections. No topic effects were detected. Though there was a learning effect, the analysis results remain valid since we counter-balanced the order of interface conditions and tasks.

To better understand how participants made their selections, we analyzed their exploration journeys. For each participant, we collected all the proxies they examined during the session (via manual proxy construction, clicking, or hovering), each associated with a relevance score and model sensitivity. To compute an average exploration pattern across participants, we first normalized each participant's timeline to a common scale. We then defined a shared time grid of 100 points and applied linear interpolation to estimate relevance scores and model sensitivities at each time point. The average trajectories of all participants are shown in Figure 4. The color

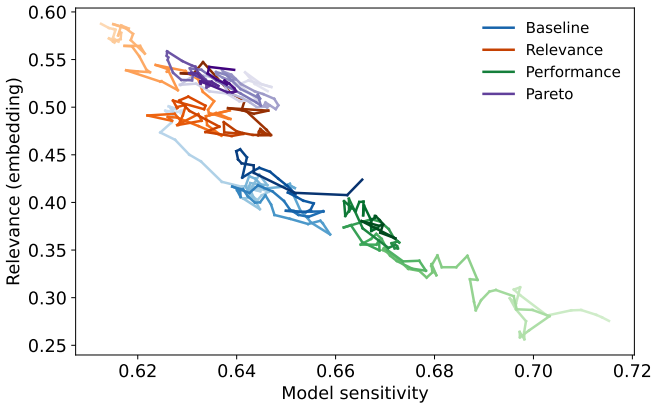


Fig. 4. Participants' selection path in a sensitivity vs. relevance plot. Light color indicates the beginning of the session, and dark color indicates the end of the session. We collected all the proxies participants examined during the study session and applied interpolation to obtain an average trajectory for each condition.

darkness indicates the time point in a session, where light colors indicate early stages and dark colors indicate late stages.

We observed that different human-machine teaming strategies significantly influenced the candidates examined by participants. Participants' exploration behaviors show that they aimed to achieve both high model performance and proxy relevance. Under Baseline, participants started with highly relevant proxies, but gradually working towards those with high model sensitivity. Participants might be motivated to improve model performance and try less relevant candidates. This optimization behavior is enabled by the fast feedback loop allowed by fast and automatic model training and evaluation. In practice, when testing a formulation takes longer, the exploration might be restricted to those highly relevant candidates. The same performance optimization behavior is also observed under Relevance First. However, participants had a higher threshold for proxy relevance under the Relevance First condition compared to Baseline. This is likely due to the presence of the estimated relevance score, which provides a red flag for their exploration. Under Performance First, participants explored increasingly relevant candidates but were anchored in the high performance region. Under Pareto Front, participants' exploration is limited to a relatively small spectrum of relevance and performance. This reflected that Pareto Front provided an overview of the problem space, reducing the need for exploration.

We can also observe from Figure 4 that different human-machine teaming strategies influenced participants' criteria when selecting the proxy target. Under Relevance First and Pareto Front, participants were restricted to candidates with high relevance (relevance score > 0.45). Their final selections are also more relevant. Under Performance First and Baseline, participants were more open to proxies that provide high model performance. One potential explanation is that under those two conditions, the quantitative performance objective was easier to track and optimize than the unquantified, subjective relevance objective. This made participants more willing to select well-performing proxies. Under Relevance First and Pareto Front, the system provided a quantitative estimate of relevance, making it easier to track and optimize. Therefore, participants appeared to be more strict about the relevance objective.

5.2 RQ2: Efficiency

We measured the time participants spent on the task when using different conditions. Kruskal-Wallis Test shows a weakly significant difference between conditions ($p = 0.0708$). Participants spent the longest time under the Baseline condition and the shortest time under the Pareto Front condition ($t_{baseline} = 402s$, $t_{pareto} = 286s$, $p = 0.098$). This reflects that when recommending candidates based on both relevance and performance information, it takes less time for people to identify a satisfying proxy.

We also examined the total number of candidates examined by participants to understand the effort required to identify a satisfactory proxy. We counted proxies manually specified, clicked, and moused over for more than 2 seconds as examined.

Participants manually specified significantly more proxies under the Baseline condition ($Avg = 23$) compared to the other conditions ($p < 0.05$). Participants examined significantly more proxies under the Relevance First condition than the Pareto Front conditions ($Avg_{relevance} = 47$, $Avg_{pareto} = 16$, $p = 0.003$). There is no significant difference between the other conditions. Participants, on average, checked three pages of recommendations under all conditions. Participants checked significantly more unique proxies under Baseline and Relevance First than Pareto Front ($Avg_{baseline} = 14.8$, $Avg_{relevance} = 20$, $Avg_{pareto} = 8.5$, $p \leq 0.05$).

In summary, participants mainly examined candidates in the ranked list when recommendations were available. There are some evidence showing that providing recommendations considering both relevance and performance (Pareto Front) helps improve the efficiency of proxy selection, reducing the time spent on the task and the required effort. Providing recommendations based on relevance (Relevance First) encourages exploration — participants examined significantly more candidates. One potential explanation is that the presence of the relevance score reduces the effort on relevance judgment, as mentioned by multiple participants.

5.3 RQ3: User Perception

There was no significant difference between conditions in terms of the questionnaire results. Despite the difference in the selected proxies' relevance score, participants consistently indicated that their final selections are "moderately relevant" to the modeling goal. System conditions did not show a significant impact on users' perceived satisfaction, confidence, easiness of the task, and the helpfulness of the system. In the questionnaire, participants indicated the relative importance of model performance and relevance when they made their decisions. On average, participants gave similar importance to the two factors. Participants gave slightly lower importance to model performance under the Pareto Front condition. Detailed analysis of user perception and qualitative feedback can be found in Appendix D.

6 Discussion

In this work, we investigated applying human-machine teaming to facilitate machine learning problem formulation. Our findings suggest that while automation significantly accelerates the iteration of proxy target selection, it introduces specific risks regarding how users prioritize quantitative performance over semantic relevance. Below, we discuss the implications of problem formulation quality and efficiency, the phenomenon of "performance bias", and design insights for future tools.

6.1 Effect of Human-Machine Teaming on the Quality and Efficiency of Problem Formulation

Our results demonstrate a tension between the efficiency of the formulation process and the quality of the resulting problem definition. Consistent with prior work [46], we found that tools designed for fast prototyping enables rapid exploration of the modeling option space. Participants were able to examine more candidates and explore divergent ideas that might not have been obvious initially. With automatic model training and evaluation, participants can examine more than 10 or even 20 different proxy candidates during a short study session.

However, higher efficiency does not necessarily translate into higher relevance of selected proxy targets. Our analysis reveals that accelerated iteration cycles can expand the space of explored candidates towards areas with degraded relevance. Participants tended to select proxies with higher performance under the Baseline condition, where the exploration process was driven by participants rather than guided by system recommendations. Participants demonstrated the behavior of chasing after performance gains, which could undermine the proxy's relevance. One potential explanation is that when users are exposed to a large volume of candidates via fast feedback loops, they would drift toward proxies that optimize quantitative metrics (e.g., sensitivity) at the expense of qualitative objectives (e.g., relevance).

6.2 Performance Bias in Problem Formulation

We observed "performance bias" in some conditions from our participants: the tendency for them to select proxies that yield high quantitative metrics, even when those formulations are less aligned with the underlying modeling goal. This might suggest a nuanced form of *automation bias* where users trust the perceived "objectivity" of machine-generated metrics over their own subjective domain judgment. We posit that performance bias arises from the cognitive disparity between evaluating quantitative metrics and assessing conceptual relevance. Quantitative metrics are numerical and therefore cognitively easy to compare. In contrast, conceptual relevance are nuanced, multi-faceted, involving domain expertise and not readily quantifiable. When a system provides rapid performance feedback without an equivalent signal for relevance, users might engage in a form of *streetlight effect*, optimizing where the light is brightest (the quantifiable objective is the highest). While our study demonstrates this bias within a proxy selection task, the degree to which this generalizes to expert practitioners—who may have stronger professional norms or institutional guardrails—remains an open question for future field studies.

In our study, conditions showing relevance scores helped mitigate performance bias. Using semantic relevance to approximate proxies' relevance to the modeling goal, though not perfect, can put resistance in the performance optimization loop, serving as a soft threshold that signaled when a user was sacrificing too much relevance for a gain in model performance. This indicates that performance bias might be mitigated when qualitative values are made computationally visible.

6.3 Design Insights to Support Problem Formulation

To support efficient and valid machine learning problem formulation, systems should move beyond simply enabling fast iterations. We propose three key design implications:

1. Support Multi-Objective Decision Making: Problem formulation is inherently a multi-objective optimization task involving trade-offs between performance, conceptual relevance, fairness, and many other objectives. Our findings show that ranking candidates solely by performance or providing only performance feedback can encourage bias. We also demonstrated that supporting users to consider multiple objectives improved the efficiency of their exploration. Future systems should leverage Multi-Criteria Decision Making (MCDM) frameworks to explicitly present the trade-offs between conflicting objectives. By visualizing the "cost" of performance in terms of relevance, systems can help users make more informed, holistic decisions.

2. Assist, Do Not Replace, Subjective Judgment: While relevance is a subjective construct best judged by humans, our study shows that users benefit from automated assistance. The Relevance First and Pareto Front condition might suggest that providing a quantitative signal for relevance helps users calibrate their internal mental models. Future tools could use large language models (LLMs) and domain-specific knowledge representation (e.g., semantic embeddings, knowledge graphs) to assist relevance judgment and potentially allow a human-in-the-loop validation process where the machine suggests, and the human verifies.

3. Provide Guardrails for Exploration: Fast iteration requires guardrails to prevent users from optimizing along the wrong targets. Systems should implement mechanisms that flag or filter candidates that violate basic

validity constraints such as relevance and fairness, regardless of their predictive performance. Furthermore, providing an retrospective overview of the user's exploration trajectory can help them reflect on the big picture of the problem space and their journey so far.

6.4 Limitations

A primary limitation of this study is the simplification of the problem formulation task and process. This was necessary to reduce technical complexity and support rapid prototyping in a short study session, allowing us to focus on the core decision-making process during proxy target selection. Despite these simplifications, our study still captures two major complexities of the problem formulation process: the large space of candidates and the multi-objective nature of deciding the best problem formulation. The behavioral patterns we observed, such as users' tendency to prioritize quantifiable objectives and the influence of machine recommendations on decision-making, are deeply rooted in human decision-making processes, not specific to the problem formulation process studied here. Therefore, we argue that the core behavioral patterns identified in this study are likely to generalize to more complex, real-world systems.

Another limitation of the study is that we used proxy participants (i.e., students) instead of domain experts. Our observations during the study show that participants were mentally involved in the task since they demonstrated independent and critical thinking when the system's recommendations conflicted with their own understanding. However, it is important to note that the generalizability of our findings to high-stakes professional contexts may be limited. Domain experts bring deeper domain and contextual knowledge of the task and the data. Therefore, they may feel that the system-generated relevance scores are not informative or not aligned with their own understanding. Similarly, domain experts may be less influenced by model performance metrics, as they have a clearer sense of what constitutes a useful model. This stronger control of the conceptual alignment with the real-world goal could lead to problem formulations with a higher practical utility than what we observed under the Baseline and Performance First conditions. Future research should explore how such systems influence the domain expert's efficiency, problem formulation quality, and experience.

The third limitation of the study is the use of semantic relevance as a proxy for target relevance. While our validation study demonstrated that SBERT embeddings align with student judgments at a level comparable to inter-human agreement, semantic similarity models primarily identify linguistic and taxonomic overlaps. They may fail to capture nuanced causal relationships or the deep contextual relevance that a domain expert might prioritize. Future work should explore tools from measurement and modern validity theory, such as convergent, discriminant, and predictive validity, to supplement purely semantic relevance [12, 20].

7 Conclusion

In this work, we explored the risks and opportunities of human-machine teaming in machine learning problem formulation. Our controlled study reveals a tension between efficiency and validity: while automation accelerates the exploration of proxy targets, it introduces a "performance bias," driving practitioners to prioritize the performance of a proxy target over its relevance to the modeling goal. This tendency to optimize for measurable objectives persists even without explicit recommendations. However, we demonstrate that interface designs that quantify relevance act as an effective guardrail, mitigating this bias while maintaining exploration efficiency. We conclude that future AI prototyping tools should go beyond speed to explicitly support multi-objective decision-making, ensuring that models are not just accurate but align with their real-world goals.

Generative AI Usage Statement

Gemini 3 Pro was used for grammar editing. The content (text and image) of the paper was not generated using generative AI tools.

Acknowledgments

This material is based upon work supported in whole or in part with funding from the Department of Defense (DoD). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DoD and/or any agency or entity of the United States Government.

References

[1] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software Engineering for Machine Learning: A Case Study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. 291–300. doi:10.1109/ICSE-SEIP.2019.00042

[2] Anthropic. 2025. Claude Code. https://claude.ai AI coding assistant, Version 2.0.

[3] Anysphere. 2025. Cursor. https://cursor.sh AI-powered code editor.

[4] Aleksander Aristovnik, Damijana Keržič, Dejan Ravšelj, Nina Tomaževič, Lan Umek, Toyin Cotties Adetiba, Adetutu Deborah Aina, Oluwatoyin Ayodele Ajani, Bibi Alajmi, Sultan Ghaleb Aldaihani, Magdalena Waleska Aldana-Segura, Said Aldhafri, Jogymol Alex, Fahad Ahmed Al-Harbi, Yusuf Alpayadin, Parag Amin, George Kofi Amoako, Octavian Andronic, Sorin Gabriel Anton, Arheiam Arheiam, Alex Riolexus Ario, Maja Arslanagić-Kalajdžić, Sofia Asonitou, Roxana Pamela Balbontín Alvarado, Martin Mabunda Baluku, Mohammad Bashaar, Joy Benatov, Naima Benkari, Syed Ahmad Helmi Bin Syed Hassan, Isaac Mensah Boafo, Roberto Burro, Michael P. Cameron, Silvia Cantele, Maria Cheraghi, Yi-Lin Chiang, Andy Choi Yeung, Simeon-Pierre Choukem, Özkan Çikrikci, Michaela Cortini, Baye Dagnew, Denilson da Silva Bezerra, Vera Dimitrievska, Beata Dobrowolska, Jadranka Đurović Todorović, Diena Dwidienawati, Falk Ebinger, Arri Eisen, Maha El Tantawi, Mahmoud M. Emam, Ibeawuchi K. Enwereuzor, Adeniyi Francis Fagbamigbe, Stefania Fantinelli, MoezAlIslam E. Faris, Ali Farooq, Maria Fedorova, Paulo Ferrinho, Barbara Fogarty-Perry, Morenike Oluwatoyin Folayan, Thais França, Bongani Thulani Gamede, Yongtao Gan, Manuel Gericota, Belinka González-Fernández, Luz María González-Robledo, Paul Gorczynski, Muji Gunarto, Adam Gyedu, Soumeyya Halayem, Sarah J. Halvorson, Nazir S. Hawi, Shiva Heidari, Azita Hekmatdoost, Meeri Hellsten, Meirav Hen, Evelyne Hübscher, Fany Inasius, Takashi Inoguchi, Yariv Itzkovich, Ervin Iusein, Telesphore Kabera, Sedighe Sadat Hashemi Kamangar, Sujita Kumar Kar, Konstantinos Karampelas, Elham Kateeb, Amrita Kaur, Kerefu Lawrence Joseph, Aleksandar Kešeljević, Pavol Kráľ, Hiroko Kudo, P. A. P. Samantha Kumara, Murodbek Laldjebaev, Kornélia Lazányi, Florin Lazăr, Paul H. Lee, Poliana Mihaela Leru, Aurora Lopez-Fogues, Rataya Luechapudiporn, Philippe N. Lukanu, Prosper Lutala, Juan D. Machin-Mastromatteo, Marwa Madi, Piotr Major, Maria Malliarou, Niko Männikkö, João P. Maroco, Bertil P. Marques, João Matias, Oliva Mejía-Rodríguez, Jana Meloska Petrova, Silvia Mariela Méndez Prado, Milena Milićević, Marek Milosz, José Joaquín Mira, Marta Miret, Alpana Mishra, Masoud Mohammadnezhad, Cristina Mollica, Immanuel Azaad Moonesar, Nicolas J. Mouawad, Elfi Mu'awanah, Dilbar Mukhamedova, Lillias Hamufari Natsai Mutambara, Joseph Muthiani Malechwanzi, Silvana G. Navarro, David Musyimi Ndetei, Nga Nguyen, Singhanat Nomnian, Alka Obadić, Ryan Michael Oducado, Olawale Festus Olaniyan, Izabela Ostoj, Efstathia Papageorgiou, Nino Paresashvili, Shirona Patel, Susan Kane Patton, Lidia Perenc, Virtudes Pérez-Jover, Harm Peters, Justyna Podgórska-Bednarz, Eka Sunarwidhi Prasedya, Bo Pu, Sumayyah Qudah, Daniela Raccanello, Agustine Ramie, Luis Armando Ramos Palacios, Mamun Ur Rashid, Vijayalakshmi Reddy, Iveta Reinholde, Maya Roche, Ana Sofia Rodrigues, Danilo V. Rogayan, Piotr Rzymski, Fahad Saleem, Roberta Sammut, Grover Sandeep, Oana Săndulescu, Rinku Sanjeev, Muhammad Saqib, Pavlos Sarafis, Muthupandian Saravanan, Mariano Schlez, Abdul-Aziz Seidu, Akkaya Senkrua, Abdel-Aziz Sharabati, Bidhan Shrestha, Aggrey Siya, Ricarda Steinmayr, Eveline Surbakti, Rajanikanta Swain, Vanphanom Sychareun, Snežana Šćepanović, David Špaček, Ivana Tadić, Kathy W. Tannous, Sanja Tatalović Vorkapić, Harold Jan Terano, Mehmet S. Tosun, Chinaza Uleanya, Olga Ushakova, Thomas Varghese, Daina Vasilevska, Tengiz Verulava, Giada Vicentini, Sornkanok Vimolmangkang, Jeffrey Dawala Wilang, Angelique Wildschut, Nikolay N. Yagodka, Guo-liang Yang, Chunlin Yao, Norhafezah Yusof, Ana-Maria Zamfir, Shehla A. Yasin, Adrian P. Ybañez, Özlem Yorulmaz, Yunquan Zhang, Oksana Zhirosh, and Al Et. 2021. Impacts of the Covid-19 Pandemic on Life of Higher Education Students: Global Survey Dataset from the First Wave. 5 (Dec. 2021). doi:10.17632/88y3nffs82.5 Publisher: Mendeley Data.

[5] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2022. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. http://arxiv.org/abs/2106.05498 arXiv:2106.05498 [cs].

[6] Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68, 3 (2013), 255–278. doi:10.1016/j.jml.2012.11.001

[7] James Bergstra, Daniel Yamins, and David Cox. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on Machine Learning*. PMLR, 115–123. https://proceedings.mlr.press/v28/bergstra13.html ISSN: 1938-7228.

[8] Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. 2022. The Effects of Data Quality on Machine Learning Performance. *arXiv preprint arXiv:2207.14529* (2022). https://arxiv.org/abs/2207.14529

[9] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable Machine: Approachable Web-Based Tool for Exploring Machine Learning Classification. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–8. doi:10.1145/3334480.3382839

[10] Dylan Cashman, Shah Rukh Humayoun, Florian Heimerl, Kendall Park, Subhajit Das, John Thompson, Bahador Saket, Abigail Mosca, John Stasko, Alex Endert, Michael Gleicher, and Remco Chang. 2019. A User-based Visual Analytics Workflow for Exploratory Model Analysis. *Computer Graphics Forum* 38, 3 (June 2019), 185–199. doi:10.1111/cgf.13681

[11] Hao-Fei Cheng, Logan Stapleton, Anna Kawakami, Venkatesh Sivaraman, Yanghuidi Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. How child welfare workers reduce racial disparities in algorithmic decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.

[12] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2023. A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. 690–704. doi:10.1109/SaTML54575.2023.00050

[13] Misha Denil and Thomas Trappenberg. 2010. Overlap versus Imbalance. In *Advances in Artificial Intelligence*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Atefeh Farzindar, and Vlado Kešelj (Eds.). Vol. 6085. Springer Berlin Heidelberg, Berlin, Heidelberg, 220–231. doi:10.1007/978-3-642-13059-5_22 Series Title: Lecture Notes in Computer Science.

[14] Ruofei Du, Na Li, Jing Jin, Michelle Carney, Scott Miles, Maria Kleiner, Xiuxiu Yuan, Yinda Zhang, Anuva Kulkarni, Xingyu Liu, Ahmed Sabie, Sergio Orts-Escolano, Abhishek Kar, Ping Yu, Ram Iyengar, Adarsh Kowdle, and Alex Olwal. 2023. Rapsai: Accelerating Machine Learning Prototyping of Multimedia Applications through Visual Programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–23. doi:10.1145/3544548.3581338

[15] Grant Duwe. 2012. Predicting first-time sexual offending among prisoners without a prior sex offense history: The Minnesota Sexual Criminal Offending Risk Estimate (MnSCORE). *Criminal Justice and Behavior* 39, 11 (2012), 1436–1456.

[16] Radwa Elshawi, Mohamed Maher, and Sherif Sakr. 2019. Automated Machine Learning: State-of-The-Art and Open Challenges. *arXiv:1906.02287 [cs, stat]* (June 2019). http://arxiv.org/abs/1906.02287 arXiv: 1906.02287.

[17] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM* 39, 11 (Nov. 1996), 27–34. doi:10.1145/240455.240464

[18] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and Robust Automated Machine Learning. In *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2015/hash/11d0e6287202fced83f79975ec59a3a6-Abstract.html

[19] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. 2013. Lineup: Visual analysis of multi-attribute rankings. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2277–2286.

[20] Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. 2023. Ground(less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 688–704. doi:10.1145/3593013.3594036

[21] Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. AutoML: A survey of the state-of-the-art. *Knowledge-based systems* 212 (2021), 106622.

[22] Charles Hill, Rachel Bellamy, Thomas Erickson, and Margaret Burnett. 2016. Trials and tribulations of developers of intelligent systems: A field study. In *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, Cambridge, 162–170. doi:10.1109/VLHCC.2016.7739680

[23] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and Visualizing Data Iteration in Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376177

[24] James Honaker and Vito D'Orazio. 2014. Statistical Modeling by Gesture: A graphical, Browser-based Statistical Interface for Data Repositories. (2014).

[25] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Enterprise Data Analysis and Visualization: An Interview Study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (Dec. 2012), 2917–2926. doi:10.1109/TVCG.2012.219

[26] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving human-AI partnerships in child welfare: understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[27] Sean Kross and Philip J. Guo. 2021. Orienting, Framing, Bridging, Magic, and Counseling: How Data Scientists Navigate the Outer Loop of Client Collaborations in Industry and Academia. http://arxiv.org/abs/2105.05849 arXiv:2105.05849 [cs].

[28] Simon Meyer Lauritsen, Bo Thiesson, Marianne Johansson Jørgensen, Anders Hammerich Riis, Ulrick Skipper Espelund, Jesper Bo Weile, and Jeppe Lange. 2021. The Framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. *NPJ digital medicine* 4, 1 (2021), 158.

[29] Lydia T. Liu, Serena Wang, Tolani Britton, and Rediet Abebe. 2023. Reimagining the machine learning life cycle to improve educational outcomes of students. *Proceedings of the National Academy of Sciences* 120, 9 (2023), e2204781120. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2204781120 doi:10.1073/pnas.2204781120

[30] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R. Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How Data ScientistsWork Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question? *Proc. ACM Hum.-Comput. Interact.* 3, GROUP, Article 237 (dec 2019), 23 pages. doi:10.1145/3361118

[31] Donald Martin Jr., Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. 2020. Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics. http://arxiv.org/abs/2005.07572 arXiv:2005.07572 [cs, stat].

[32] Raymond McCall and Janet Burge. 2016. Untangling Wicked Problems. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 30, 2 (2016), 200–210.

[33] Microsoft. [n. d.]. Team Data Science Process -TDSP. https://datascienceprocess.com/member-home-page/team-data-science-process-tdsp/

[34] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–15. doi:10.1145/3290605.3300356

[35] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kastner. 2022. Collaboration challenges in building ML-enabled systems: communication, documentation, engineering, and process. In *Proceedings of the 44th International Conference on Software Engineering*. ACM, Pittsburgh Pennsylvania, 413–425. doi:10.1145/3510003.3510209

[36] David F. Nettleton, Albert Orriols-Puig, and Albert Fornells. 2010. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review* 33, 4 (April 2010), 275–306. doi:10.1007/s10462-010-9156-z

[37] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (Oct. 2019), 447–453. doi:10.1126/science.aax2342 Publisher: American Association for the Advancement of Science.

[38] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 39–48. doi:10.1145/3287560.3287567 arXiv:1901.02547 [cs].

[39] Kayur Patel, Naomi Bancroft, Steven M. Drucker, James Fogarty, Amy J. Ko, and James Landay. 2010. Gestalt: integrated support for implementation and analysis in machine learning. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, New York New York USA, 37–46. doi:10.1145/1866029.1866038

[40] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 469–481. doi:10.1145/3351095.3372828

[41] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew D. Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972. doi:10.1145/3531146.3533158 arXiv:2206.09511 [cs].

[42] N Reimers. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084* (2019).

[43] Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. doi:10.48550/arXiv.2505.10573 arXiv:2505.10573 [cs].

[44] Aécio Santos, Sonia Castelo, Cristian Felix, Jorge Piazentin Ono, Bowen Yu, Sungsoo Ray Hong, Cláudio T. Silva, Enrico Bertini, and Juliana Freire. 2019. Visus: An Interactive System for Automatic Machine Learning Model Building and Curation. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics - HILDA'19*. ACM Press, Amsterdam, Netherlands, 1–7. doi:10.1145/3328519.3329134

[45] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2020. A Human-Centered Review of Algorithms Used within the U.S. Child Welfare System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376229

[46] Venkatesh Sivaraman, Anika Vaishampayan, Xiaotong Li, Brian R Buck, Ziyong Ma, Richard D Boyce, and Adam Perer. 2025. Tempo: Helping Data Scientists and Domain Experts Collaboratively Specify Predictive Modeling Tasks. *arXiv preprint arXiv:2502.10526* (2025).

[47] Stefan Studer, Thanh Binh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters, and Klaus-Robert Mueller. 2021. Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. http://arxiv.org/abs/2003.05155 arXiv:2003.05155 [cs, stat].

[48] Eran Tal. 2023. Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 312–321.

[49] Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. 2019. Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. *California Management Review* 61, 4 (Aug. 2019), 15–42. doi:10.1177/0008125619867910 Publisher: SAGE Publications Inc.

[50] Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. 2017. Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation. (2017). https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V1-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL.pdf Methodology V1 from 16-ACDHS-26 Predictive Risk Package.

[51] Elmira Van den Broek, Anastasia Sergeeva, and Marleen Huysman. 2021. When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. *MIS quarterly* 45, 3 (2021).

[52] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2024. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM Journal on Responsible Computing* 1, 1 (2024), 1–45.

[53] Samuel J. Weisenthal, Caroline Quill, Samir Farooq, Henry Kautz, and Martin S. Zand. 2018. Predicting acute kidney injury at hospital re-entry using high-dimensional electronic health record data. *PLOS ONE* 13, 11 (Nov. 2018), e0204920. doi:10.1371/journal.pone.0204920

[54] Liwenhan Xie, Chengbo Zheng, Haijun Xia, Huamin Qu, and Chen Zhu-Tian. 2024. WaitGPT: Monitoring and Steering Conversational LLM Agent in Data Analysis with On-the-Fly Code Visualization. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 119, 14 pages. doi:10.1145/3654777.3676374

[55] Qian Yang, Aaron Steinfeld, Carolyn Rose, 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*. ACM, Honolulu HI USA, 1–13. doi:10.1145/3313831.3376301

[56] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, Hong Kong China, 573–584. doi:10.1145/3196709.3196729

Appendix

A Application Scenarios and Variable Names

We chose the application scenarios that the participants need to build models to identify students who are negatively impacted by the pandemic in two life aspects: mental health and academic performance. However, to provide an overview of the entire problem space, we simplified the task by reducing the number of outcome variables and converting all the outcome variables into binary. A proxy variable Y involves a subset of the outcome variables $\mathcal{V} \subseteq \mathcal{U}, \mathcal{V} = \{V_1, \cdots V_k\}$. The function g combines the subset outcome variables using logic operators (e.g., OR, AND, NOT), resulting in binary proxy variables.

We focused on the challenging scenario where users must trade off between relevance and performance. To facilitate the comparison between conditions, we aim to identify two sets of outcome variables with similar data conditions. For the topic of mental health, there exists a tradeoff between model performance and proxy relevance in the original dataset, i.e., the relevant outcome variables result in relatively low model performance while the irrelevant outcome variables result in relatively high model performance. We selected 10 outcome variables with the selected performance metric uniformly distributed between 0.59 (min) to 0.77 (max).

As finding another set of outcome variables with similar characteristics is challenging, we created a synthetic dataset for the topic of academic performance. We retained the underlying data for the mental health scenario but altered the variable names. For an outcome variable in the mental health scenario with a specific relevance to mental health, we selected an original outcome variable name from the survey dataset with a similar level of relevance to academic performance based on the embedding model. This process resulted in 10 outcome variables with various levels of relevance to the academic performance scenario. To create a tradeoff between relevance and model performance, the variable names and outcome variables are matched with a reversed order of performance and relevance. The variable name with the highest relevance to the topic is assigned to the outcome variable that will result in the lowest model performance and vice versa.

Scenario 1: Mental Health. You are a data scientist in the student affairs office at UNC. During COVID-19, **you would like to identify students who are experiencing mental health issues and offer them therapy sessions**. List of outcome variables:

- Frequently feel worried about mental health
- Frequently feel angry while attending classes and studying
- Frequently feel worried about family and relationship
- Frequently feel hopeless while attending classes and studying
- Difficulty in focus on schoolwork
- Rarely visit family members or friends
- Feel unsatisfied in how university deal with the pandemic
- Frequently washing hands
- Frequently online grocery shopping
- Feel unsatisfied in tutorials

Scenario 2: Academic Performance. You are a data scientist in the student affairs office at UNC. During COVID-19, **you would like to identify students who are experiencing a drop in academic performance and offer them tutoring services**. List of outcome variables:

- Academic performance as a student worsen
- Cannot master the skills taught in class
- Frequently feel bored while attending classes and studying
- Frequently feel worried about studying issues
- Feel unsatisfied in lectures

- Rarely communicating with family and friends
- Feel not adapt well to the new teaching and learning experience
- Frequently avoiding public transport
- Frequently working from home
- Frequently feel worries about similar pandemic crisis

B Relevance Modeling

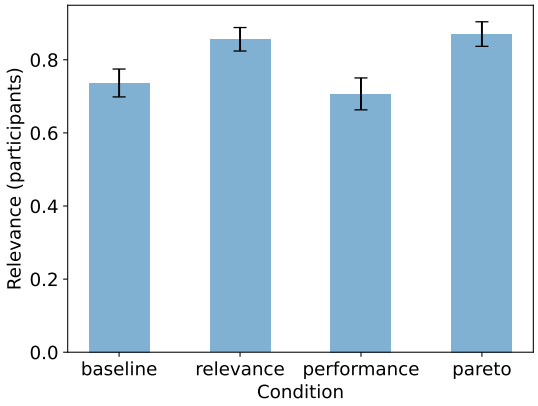


Fig. 5. Average relevance score of participants' final proxy selection, calculated using the Bradley-Terry model using participants' relevance judgment. The relative difference between conditions shows the same pattern as using the relevance scores based on the embedding model (Figure 3b)

As participants may disagree with the embedding model, we fitted a Bradley-Terry model to generate a relevance score for each proxy target given participants' relevance judgments. The Bradley–Terry model is a probability model for the outcome of pairwise comparisons between items. Given a pair of proxies i and j drawn from all potential proxies, it estimates the probability that the pairwise comparison $i > j$ (i is more relevant than j) turns out true, as

$$Pr(i > j) = \frac{p_i}{p_i + p_j} = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}} .$$

The fitted β_i and β_j can be used as the relevance score for proxy i and j. We fitted one Bradley–Terry model using the relevance judgments collected from all participants.

The final proxies' relevance using the derived relevance score for each condition is shown in Figure 5. There is a significant difference between conditions ($p = 0.003$). There is a significant difference between Relevance First and Performance First ($p = 0.036$, Hedges' $g = 0.783$), Performance First and Pareto Front ($p = 0.018$, Hedges' $g = -0.843$). The difference between the Baseline and Pareto Front is weakly significant ($p = 0.066$, Hedges' $g = -0.747$). Overall, the conclusion is consistent with the results calculated using the relevance score given by the embedding model.

C Study Details

C.1 Procedure

User study sessions last roughly one hour. After providing informed consent, the study coordinator provided detailed information about the application context, dataset, and proxy syntax. A study session contained two sub-sessions. Using an example scenario, each sub-session would begin with a hands-on tutorial. Participants will then be informed of the application scenario to work on. Participants would first finish the relevance judgment task. Then, participants would complete the proxy selection task. To encourage the participants to take the task seriously, the study coordinator would ask them to verbalize their reasoning and justify their final choices after each sub-session. Participants also finished a questionnaire after each sub-session to report their experience. Participants could spend at most 15 minutes on each task. Finally, after completing the two sub-sessions, participants provided additional feedback about their experience using the tools through an exit interview.

To help participants get mentally involved in the proxy target selection task and built a mental model of the system's recommendation mechanism, model performance, and relevance score, we provided a thorough tutorial at the beginning of the study session regarding the following aspects:

- **Machine learning model and evaluation:** what is a binary classification model, how to interpret different evaluation metrics, and how would model performance influence the usage of the model?
- **Proxy target selection task and objectives:** What is a proxy target, why do we need to define one, and why are both the relevance of the proxy and the resulting model's performance important? How could poor formulations lead to harmful or inefficient decisions?
- **Intended usage of the model in each application scenario:** What is the construct of interest, and how is the model going to be used?
- **System algorithms:** How does the system generate the relevance score and the model performance metric, how does the system generate the recommendation list in each condition, and how to interpret the scores and ranking given by the system?

Before each session, participants started with a hands-on tutorial where they use the system to do a proxy target selection task with an example scenario. Participants were encouraged to ask any clarification questions. This helped participants understand the system conditions and the task before the main task.

C.2 Evaluation Metrics

Different system conditions are evaluated using both subjective feedback from participants and objective evaluation of participants' performance. The evaluation metrics are organized based on the corresponding hypothesis.

RQ1: Quality.

Performance. The achieved model performance is measured using sensitivity (recall), which also serves as an objective for participants during the study. A single model performance metric is used as the study focuses on decision-making processes concerning two broad objectives: performance and relevance. Among the potential performance metrics (e.g., F1, accuracy), sensitivity was chosen for its simplicity and emphasis on coverage (identify all students who might need help). We evaluated the model sensitivity of the final proxy target selected and the saved proxies by each participant. In addition, to understand the intermediate decisions made by participants, we tracked the change in performance between successive selections.

Relevance. The relevance of the final proxy selection and the saved proxies were measured using the relevance score given by the language model (Section 4.2.1) and the relevance model derived from participants' relevance judgments (Appendix Section B). Similarly, we tracked the change in relevance between successive selections using the language model. Since participants might disagree with the relevance score, participants will be asked

to provide subjective evaluations of the final selection's relevance on a Likert scale from 1 "strongly irrelevant" to 7 "strongly relevant".

RQ2: Efficiency.

Time on task. This metric aims to understand the time participants spent on making a selection.

Number of constructed or inspected candidates. This metric is the number of proxies a participant needed to inspect to identify a satisfying one, serving as an indicator of efficiency.

Number of pages looked at. This metric serves as another indicator of efficiency, as going through more pages of recommendations implies that participants spent more effort in finding satisfying proxy targets.

Quality of the clicked candidates. This metric aims to evaluate whether a participant spent time in "high-value" explorations. Specifically, we evaluated whether the newly selected proxy target improved relevance, performance, or both compared to the previous choice. Additionally, we evaluated whether the new proxy explored new outcome variables or different areas in the objective space.

Time on different views. We measured the time duration of participants' mouse hovering over the proxy detail panel and the proxy recommendation view. Mouse activities can partially reflect the participant's focus, allowing us to understand whether a participant spent more time making decisions or interpreting individual candidates.

RQ3: User Perception. After completing each sub-session, participants would evaluate the choice they just made according to the following criteria:

Satisfaction. Participants would be asked to what extent they are satisfied with their choice, ranging from "not satisfied at all" to "very satisfied".

Confidence. They would be asked to what extent they are confident about their choice, ranging from "not confident at all" to "very confident".

Easiness. They would be asked to what extent they consider this choice easy to make, ranging from "very difficult" to "very easy".

Usefulness. They would be asked to what extent they consider the system recommendation to be useful, ranging from "very useless" to "very useful".

Preference. After the two study sessions, participants would be asked to choose which system they prefer.

Relative importance between performance and relevance. To understand how different recommendations influence people's decision-making, participants were asked to rate the relative importance they assigned to relevance and performance when making decisions, ranging from "relevance only" to "performance only".

Agreement with the embedding model. Participants were asked to give relevance judgments of proxy target pairs to calculate their agreement with the embedding model as described in Section 4.2.1. This measurement allows us to analyze how the difference in interpretation influenced users' performance and perception of the system.

C.3 Data Analysis Methodology

Participants' questionnaire responses and the quality measurements for selected proxy targets were averaged within system conditions for comparison between conditions. Fisher's randomization test ($\alpha = 0.05$) was used to test for significant effects due to system condition differences. We also analyzed the effects of the application scenario and the order of the systems using a mixed ANOVA model.

D Additional Analysis

D.1 RQ3: User Perception

There was no significant difference between conditions in terms of the questionnaire results. Despite the difference in the selected proxies' relevance score, participants consistently indicated that their final selections are "moderately relevant" to the modeling goal. System conditions did not show a significant impact on users' perceived

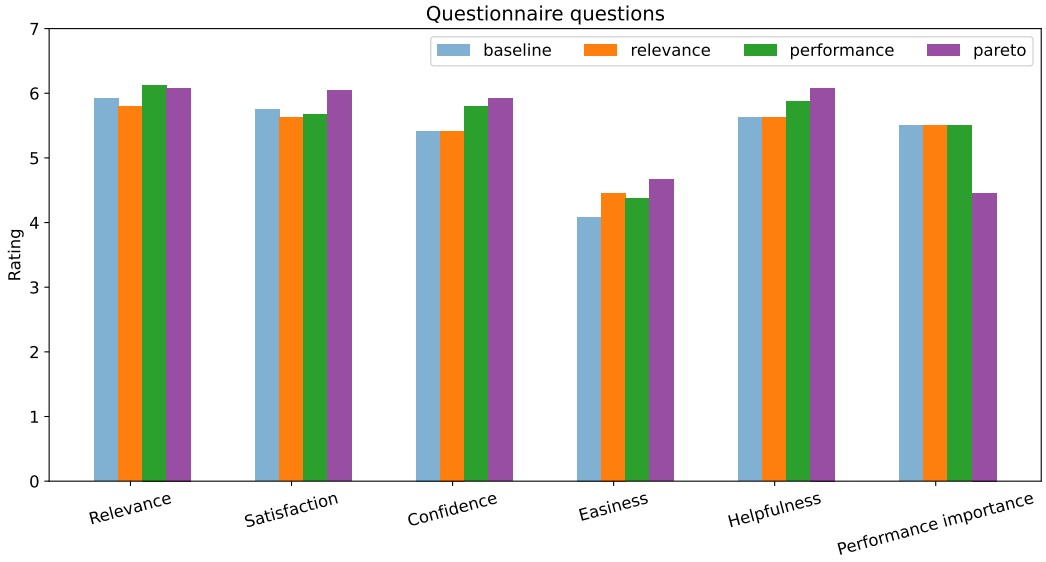


Fig. 6. Participants' responses in post-task questionnaire regarding their experience of using the system to accomplish the proxy selection task. There is no significant difference between conditions.

satisfaction, confidence, easiness of the task, and the helpfulness of the system. In the questionnaire, participants indicated the relative importance of model performance and relevance when they made their decisions. On average, participants gave similar importance to the two factors. Participants gave slightly lower importance to model performance under the Pareto Front condition.

D.2 Qualitative Analysis

At the end of the study section, participants described their strategy of approaching the task and explained what they liked or did not like about each system used. In this section, we focus on analyzing participants' answers on system preferences, as their answers reflect how they leveraged different information and how different conditions influenced their decision-making and experience.

D.2.1 Which system condition did the participants prefer to use? Participants indicated which of the two conditions they preferred. Note that each condition was used by exactly 24 participants. 18 of the 24 participants who used Pareto Front preferred it over the other. 13 of the 24 participants who used Relevance First preferred it over the other. Same for the Performance First condition. Only four of the 24 participants who used the Baseline condition preferred it over the other. These results show that participants preferred systems with recommendations. And among different recommendation conditions, Pareto Front was the most preferred.

D.2.2 What information was useful to users and how did they use it? During the interview, 18 participants mentioned that the machine-estimated relevance score was helpful. Although participants agreed with the relevance score to different degrees, they mentioned that the relevance score helped reduce the required efforts of relevance judgment and allowed them to narrow down their selection. Participants also preferred to have the resulting model's performance easily available (n = 13), which helped reduce the effort of going through different candidates and narrow down their selection. Many participants who have used the Pareto Front

condition appreciated seeing the performance and relevance information side by side ($n = 11$). This provides more information for their decision-making and helps some participants ($n = 3$) to see the tradeoff between model performance and relevance. However, participants showed divergent opinions on the recommendation level (star levels in Pareto Front). Six participants indicated that they liked the recommendation level, as it provided guidance for their selections and gave them confidence. Six participants felt the recommendation level was not helpful, as the levels were hard to understand or misaligned with their own preferences. Sometimes participants felt self-doubt if their selection was assigned a low recommendation level.

D.2.3 How did system recommendation or ranking influence participants' exploration and decision? **Relevance judgment**. Ten participants mentioned that the relevance score assisted them in judging which candidates are relevant to the modeling goal. Some participants thought relevance was "hard to quantify", therefore requiring deep thinking and some "guessing". But with the relevance score, the required efforts for relevance judgment are reduced, so they can focus on making their choices. This is especially true since the task requires participants to consider both problem relevance and the resulting model's performance. Having the relevance objective quantified and easily available allows fast decision-making. For instance,

P26: *I maybe prefer the one with the relevance, although I know it is a reference, it also helped me. It saved my time too. I don't need to do all the combination in my mind randomly. Instead, I can choose from the relevant rank, and then, according to rank, I can choose. Maybe the relevance score is not 100% correct. I can take it as a reference. So in this way I can only consider, like, the accuracy.*

The relevance score also served as another perspective, informing participants whether they are "in the right direction or not" and provided confidence in their selection. Participants were aware that these relevance scores cannot replace their own judgment. However, they appreciated having a reference.

P35: *You know, I have my own notion of what I think is relevant. But right, if you call that relatively objective, then it, you know, makes me sort of check my sort of gut instinct.*

P47: *Because the relevance score ranking help me to guide ... to make quick judgment about whether I'm in the right direction or not, so I can think more on the performance. If there is a ranking for the relevance, although it's generated by a large language model, it still can make me a bit easier to find the relevance.*

Filtering. Ten participants mentioned that ranking by the relevance score, model performance, and recommendation star levels helped narrow down their selections as this information allowed them to filter out "less ideal" options.

P12: *If I'm using the relevance ranking one I can just check, like the top, like maybe 5 high relevance ones and try to see which one has a really high sensitivity.*

P20: *I think on the 1st version I kind of like the star system because it says like these are it because it groups things that are like similar in like overall performance, so it tells me, like these are still all performing really well, and like these are performing less well, like, I probably wouldn't look at the 3 stars.*

P29: *It is helpful, because with a listing by sensitivities, I tried to avoid the ones with really low sensitivities, so like the second, or like the half with the low, the lower sensitivity. I try not to look at those.*

Some participants adopted the strategy to first filter candidates based on one objective, and then choose based on the second objective. This was thought to be an easier strategy than thinking about the tradeoff relationship between two objectives.

P28: *I only need to pay attention to that thing and then think other things, because kind of filter out like the ones I don't like. Yes, and then it's like, it's have less choices. I feel so it's easier. But the 1st one (the Pareto Front condition) is the one I need to trade off. Have trade off more. Yeah, that's a long time.*

Although the filtering approach reduced effort, it may also cause participants to miss their ideal options if those options were ranked lower. For instance, P11 was using the Performance First condition in one of the study sessions. His final choice was located on the fifth page of the ranked list. He commented:

P11: *So if I'm not patient enough, I would just go with someone in the 1st page, right? And then I feel like it might be a wrong choice. But then, after I scroll to like the very end, the 4th or 5th page, I just find the one that I want to choose. So yeah, I mean, if I'm not patient enough, I probably just gonna miss that one.*

Required exploration and efforts. In total, 17 participants mentioned that using certain system conditions required more effort. Ten participants mentioned that the Baseline condition required more efforts comparing to the other condition, as it required "testing out things one by one", "trials-and-errors", "going back and forth between selections", and remembering what was tried and what were the good options. Five participants mentioned that the Relevance First condition required more effort compared to the conditions with model performance information. It was because it required clicking on options to view model sensitivity scores.

P8: *It took me longer time to see the distribution of sensitivity scores.*

However, one participant mentioned that despite requiring more effort, the Relevance First condition encouraged them to explore more options.

P7: *I think the 1st version is better towards encouraging me to explore different options, because I have to actually click through more, and because it takes a lot of energy to click through 55 options like, I think, I'm more incentivized to mentally determine which features matter more.*

System bias, self-doubt, and agency. Eighteen participants mentioned that they felt the system ranking or the recommendation levels biased their selections. Participants felt that they were more inclined to select from the top of the ranked list or select the options with higher scores (recommendation levels or relevance scores). When their choices disagree with those of others, they feel self-doubt.

P3: *The recommendation was interesting, because I feel like it did sort of like make me question my choice, right, whether good or bad. So it either validated my choice like this one. I think it validated it because it was a 5 star. It was the 3rd choice, and I thought those 2 variables were very relevant. But the other one, what I thought I wanted was like a 4 star. So it made me kind of question that, had me go back in a little bit more. But I thought it was helpful to see like where the model thought it was right compared to what I thought.*

For some participants, this self-doubt or disagreement caused discomfort. For others, this changed their beliefs about what the good options were. For instance, P42 used the Performance First condition in one of the sessions. After noticing that some options had higher sensitivity, they gave extra thought on why the proxy target could be relevant.

P42: *I think I chose 2 different things. But then, my decision got changed a little bit because I realized these ones have higher sensitivity. And then when I give it more thought, I realized, okay, there's a reason that this could be relevant. And I kind of gave it more thought as to why that might be the case.*

Some participants noticed that when using different conditions, they weighted the two objectives differently.

P29: *I think the 1st one didn't give you a list of all the sensitivities. I think using that one I was more focused on like the relevance. And then the second one, you can see all the sensitivity, makes me more focused on the sensitivity and not as much relevance.*

Six participants mentioned that they felt they had more control when using the Baseline system versus other conditions. Participants reflected that the Baseline condition allowed them to apply their own strategy and made focused selections based on their preferences rather than going through a long list of options recommended by the system. Participants also mentioned that they were making "data-driven" decisions and were pushed to select the top-ranking candidates when there were recommendations. In comparison, when using the Baseline condition, the system cannot sway them.

P45: *I would say the second one (the Baseline condition) definitely got me into the like, the juggling part, the actual like changing thing part. I think that's very helpful, and it kind of feels like I am more actively choosing things versus the other one. I'm just like browsing through a list and picking one on balance.*

Received 13 January 2026; revised 22 April 2026; accepted 2 March 2026