Detecting when Users Disagree with Generated Captions

Omair Shahzad Bhatti
Interactive Machine Learning, German Research Center for Artificial Intelligence (DFKI)
Germany
omair_shahzad.bhatti@dfki.de

Harshinee Sriram
University of British Columbia
Canada
hsriram@cs.ubc.ca

Abdulrahman Mohamed Selim
Interactive Machine Learning, German Research Center for Artificial Intelligence (DFKI)
Germany
abdulrahman.mohamed@dfki.de

Cristina Conati
University of British Columbia
Canada
conati@cs.ubc.ca

Michael Barz
Interactive Machine Learning, German Research Center for Artificial Intelligence (DFKI)
Germany
Applied Artificial Intelligence, University of Oldenburg
Germany
michael.barz@dfki.de

Daniel Sonntag
Interactive Machine Learning, German Research Center for Artificial Intelligence (DFKI)
Germany
Applied Artificial Intelligence, University of Oldenburg
Germany
daniel.sonntag@dfki.de

Abstract

The pervasive integration of artificial intelligence (AI) into daily life has led to a growing interest in AI agents that can learn continuously. Interactive Machine Learning (IML) has emerged as a promising approach to meet this need, essentially involving human experts in the model training process, often through iterative user feedback. However, repeated feedback requests can lead to frustration and reduced trust in the system. Hence, there is increasing interest in refining how these systems interact with users to ensure efficiency without compromising user experience. Our research investigates the potential of eye tracking data as an implicit feedback mechanism to detect user disagreement with AI-generated captions in image captioning systems. We conducted a study with 30 participants using a simulated captioning interface and gathered their eye movement data as they assessed caption accuracy. The goal of the study was to determine whether eye tracking data can predict user agreement or disagreement effectively, thereby strengthening IML frameworks. Our findings reveal that, while eye tracking shows promise as a valuable feedback source, ensuring consistent and reliable model performance across diverse users remains a challenge.

CCS Concepts

• **Human-centered computing** → **User studies**; **User models**; • **Computing methodologies** → *Supervised learning by classification.*

Keywords

disagreement detection, interactive machine learning, eye tracking, gaze, emotion detection, user disagreement

ACM Reference Format:
Omair Shahzad Bhatti, Harshinee Sriram, Abdulrahman Mohamed Selim, Cristina Conati, Michael Barz, and Daniel Sonntag. 2024. Detecting when Users Disagree with Generated Captions. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI Companion '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3686215.3688382

ICMI Companion '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0463-5/24/11
https://doi.org/10.1145/3686215.3688382

1 Introduction

As the use of Artificial Intelligence (AI) increases in various aspects of daily life, there is also a growing demand for AI-based systems to operate autonomously while considering user preferences and feedback. A promising concept to meet this demand is interactive Machine Learning (IML). This approach allows users, including non-experts, to dynamically and incrementally steer and train models by integrating human feedback into machine learning processes [1]. Much like IML, many AI systems across different fields rely on explicit feedback to learn user preferences and adapt system behavior. However, frequent requests for explicit feedback can quickly become a source of frustration for users [7], reducing their trust in the AI system and negatively impacting their perception of its accuracy [13].

Existing literature has proposed strategies to improve interaction within IML systems [9, 12, 32]. Specifically, works such as Dudley and Kristensson [9] argue for minimizing the frequency of user-AI interactions, suggesting that feedback should only be requested when it is critical to the model's learning process. Motivated by this perspective, we explore using implicit signals, such as eye tracking, to capture user feedback implicitly. Eye tracking is an unobtrusive method of capturing a user's gaze and can provide insight into the user's cognitive processing and focus of attention. These implicit signals could help identify instances of perceived agreement or disagreement with the output of an AI system, providing valuable feedback for system improvement and potentially helping decide

when to trigger explicit feedback requests. This approach could reduce user effort and make interactions with AI systems more user-friendly.

Our work focuses on an image captioning scenario, where AI systems generate textual descriptions (captions) for images. Despite recent advancements, these models can generate incorrect captions, leading to disagreement with the AI system's output. We aim to predict these occurrences using implicit signals. Therefore, we conducted a user study with thirty participants who interacted with a simulated image captioning system while we captured their eye movements and video-recorded their faces. Participants were instructed to evaluate captions (half of which were intentionally flawed) using images and captions sourced from the FOIL-COCO dataset [28]. By observing participants' reactions to correct and intentionally incorrect captions, we aimed to identify markers of disagreement. Specifically, the contributions of this work are as follows:

(1) A dataset with 30 participants' interactions with image-caption pairs, recording eye tracking and pupil dilation data, along with their binary ratings indicating agreement or disagreement with the generated captions.
(2) A cross-user experiment implementing a Leave-One-User-Out 10-fold cross-validation approach to examine the potential for generalizability in disagreement detection models across different users.
(3) A within-user experiment aimed at investigating the effectiveness of personalized model adaptations, assessing whether customized models can enhance the performance of disagreement detection in a user-specific context.

2 Background & Related Work

Central to our research is the notion of user disagreement, which we define as instances where the system's output during a specific task does not align with a user's expectation. These instances can serve as potential feedback signals or triggers to request further feedback in machine learning systems. To enhance our understanding of disagreement, we draw on relevant research from affective computing and try to find a connection to affective states.

Existing literature has shown that human gaze and facial expressions can be used for affect recognition [16, 33] and serve as sources for implicit user feedback [2, 3, 27]. For instance, Lallé et al. [15] utilize gaze data to predict states of *confusion*. According to D'Mello and Graesser [10], confusion "is hypothesized to occur when there is a mismatch between incoming information and prior knowledge [...], thereby initiating cognitive disequilibrium" (p. 292). Thus, we hypothesize that signs of confusion could act as indicators of user disagreement with a model's output. This state of confusion intersects with our understanding of user disagreement—essentially, a mismatch between what users expect and what the system delivers. Nonetheless, while confusion might signal disagreement, not every case of disagreement is necessarily tied to confusion. Further, Pollak et al. [22] investigated the use of facial emotion recognition technologies to distinguish between user satisfaction and dissatisfaction, thereby establishing a clear relationship between emotional responses and user feedback. Inspired by their findings, our study is specifically designed to elicit user disagreement and use eye trackers and cameras to record user behavior and reactions.

Early research on confusion detection originates from the field of educational computing [6, 8], where detection techniques often involve analyzing facial expressions of students, posture or interface interaction, and their studying behavior. Pachman et al. [21] propose using gaze data for predicting confusion in digital learning environments, by tracking the progression of the user's puzzle-solving tasks. Their goal was to detect the buildup of confusion during the problem-solving process. Our focus shifts from these studies by concentrating on the immediate affective state of confusion that results from the user processing the information of the model's output. Detecting this type of *immediate* confusion is especially relevant in Human-Computer Interaction (HCI) as it impacts user experience and satisfaction [19]. Salminen et al. [23] develop a confusion predictor partly derived from gaze data within their persona information visualization tool, using metrics such as the number of fixations, transitions between Areas-of-Interest (AOIs), and users' demographic information to predict confusion with 80% accuracy. They later enhance this predictor by solely using gaze data, achieving a 70% accuracy rate in identifying confusion, which boosts to 99% when demographic details are integrated. This indicates a strong correlation between demographic factors and confusion instances. However, while demographic features can help model the frequency of confusion, they may not be effective for real-time monitoring. Notably, they highlight that confusion predominantly affects inexperienced, older male users in contrast to younger participants — a finding that hints at a possible correlation between confusion and demographic traits such as age and gender. However, while demographic features can help model the frequency of confusion across different user groups, they may not be as effective for real-time confusion monitoring.

Lallé et al. [15] created a predictor of user confusion during interaction with their interactive data visualization tool ValueChart, an interactive data visualization tool designed to aid users in making well-informed decisions (such as finding rental property) aligned with their preference In their study, with 136 participants, gaze and mouse movement were collected as users performed tasks with the tool. Users could indicate confusion by clicking a dedicated button in the tool's interface in the top-right corner. Using a Random Forest Classifier, the authors' model achieved a 61% accuracy in predicting confusion. A more recent contribution from the same group [29] uses deep learning based on raw eye movements to predict confusion on the same dataset as [15]. They shifted from pre-processed features to raw sequential gaze data, fed into a Recurrent Neural Network (RNN). According to the author, this method enabled the RNN to uncover subtle patterns indicative of confusion, outperforming the previous model with an accuracy of 82% — a noteworthy improvement over the initial 61%. The success of this approach supports the potential of combining deep learning with unprocessed sequential gaze data for more accurate affect recognition. However, the dataset is highly imbalanced, with instances of no confusion overwhelmingly outnumbering confusion cases (99% vs 1%). This skew could potentially bias the model's ability to identify confusion accurately. Additionally, the interface's confusion self-report button might affect users' gaze behavior, introducing further data collection complexities. In response to these issues, we propose utilizing a handheld trigger to capture user feedback to minimize disruption to their gaze [5]. To further enhance the reliability of our study, we aim to balance cases of agreement and disagreement,

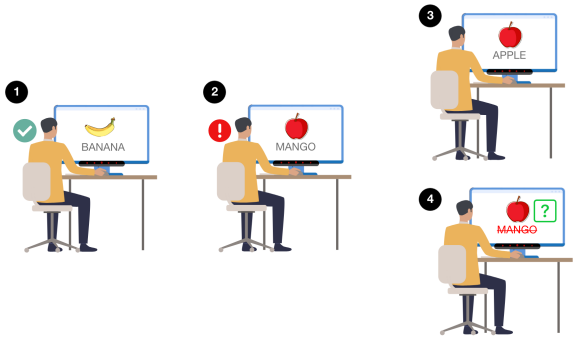

Figure 1: (1) User interacts with an IML system; (2) a predictor picks up that the user disagrees with the output; (3) the IML system reacts by returning an alternative solution or (4) triggers a feedback request. Steps (3) and (4) illustrate possible future integration in an IML system

thus addressing the disproportion found in the earlier dataset in the data collection study.

2.1 Application in Machine Learning

The concept of implicit feedback for artificial agents is a recent idea, with limited literature on the topic. In an explorative study, Pollak et al. [22] investigated the potential of user emotional feedback serving as a reward signal for a reinforcement learning agent. This feedback, determined by facial emotion recognition, was designed to reflect the user's *satisfaction* level, categorizing emotions into negative (such as 'angry', 'disgust', 'fear', 'sad'), positive ('happy'), or neutral ('neutral', 'surprise') categories according to [11]. They enabled a user to control a virtual drone, which, informed by the user's emotional reactions, adapted its movements to align with correct actions. Their preliminary results indicate that emotional feedback could indeed be integrated as a functional part of a reinforcement learning agent's reward system. However, they also observed significant variances in the intensity of emotional feedback from participant to participant, which presented challenges in distinguishing between positive and negative reactions accurately.

Krause and Vossen [14] suggest using signs of user confusion or uncertainty as cues to provide explanations in interactions between humans and AI agents. They argue that explanations should not only be provided when the user explicitly asks for it but also when the system identifies signs of the user's uncertainty or confusion. Further, they identify additional triggers like, belief conflicts, or misunderstandings of the agent's output, which are in line with the indicators of user disagreement that our research aims to explore. While Krause and Vossen [14]'s approach focuses on delivering explanations, we propose integrating these indicators into the interactive machine learning cycle. This integration might involve providing alternative model outputs and, when necessary, soliciting additional user feedback to facilitate continuous learning (see Figure 1).

3 Data Collection

In this section, we detail our data collection study. This study was designed to explore the potential of eye movement as an implicit signal for detecting user disagreement in machine learning interaction. We are particularly interested in the context of image captioning tasks, as this setting allows for the occurrence of disagreements due to model-generated errors or unsuitable captions. Hence, in the study, we presented participants with a series of images and their associated captions from the FOIL COCO dataset [28], with half of the captions intentionally containing errors to elicit disagreement. Meanwhile, the participants were recorded using an eye tracker and a camera. Next, we provide information about the participants, the specifics of the task, the stimuli involved, the apparatus setup, and the procedure followed for data collection. Finally, an overview of the collected dataset is presented.

3.1 Participants

We designed and conducted this study following guidelines provided by our Ethics Committee. The experiment was reviewed and approved by the Committee before any recruitment or data collection. We recruited 31 potential participants via email and university postings. However, due to complications with the eye tracker, one participant's data could not be included, resulting in a final count of 30 participants (21 males, 9 females, avg. age 26.4). Nineteen participants had used an eye tracker before. Each participant was fluent in English and had normal or corrected-to-normal vision. For their contributions, participants were compensated at a rate of 15 Euros per hour. The study took around 60 minutes.

3.2 Task

Participants in our study were primarily asked to rate the accuracy of the images and captions paired. They were given a series of images, each linked with a corresponding caption derived from the FOIL-COCO dataset. Half of these captions contained deliberate errors to mitigate the issue of data imbalance prevalent in related research. As participants viewed each image-caption pair, they were instructed to provide a binary rating of 'agree' or 'disagree' based on their perception of the caption's correctness. Once a rating was provided, they could proceed to the next pair in the series.

3.3 Stimuli

We selected a total of 154 images, 134 from the FOIL-COCO training set and 20 from the FOIL-COCO validation set. The FOIL-COCO dataset builds upon the standard COCO dataset by providing 'foil' captions — captions that are identical to the original captions but with one intentional error. To ensure a diverse range of categories, we included two images from each of the dataset's supercategories. We primarily selected captions around ten words in length and standardized the image resolution across all stimuli. Participants were split into two groups — Group A and Group B. Both groups were presented with the same images to maintain uniform visual stimuli. The key distinction between the two groups' experiences was the presentation of 'foil' captions: if Group A saw the correct caption for a given image, Group B would see the foil caption for that same image, and vice versa. To reduce order effects, we randomized the image sequence for each participant.

Figure 2: Screenshot of the study interface showcasing an image-caption pair. Here the induced error in the caption is the word 'keyboard' instead of 'phone'.

3.4 Apparatus

The setup included a Tobii Pro Fusion eye tracker [1], operating at 250Hz, mounted on a 27-inch monitor. Directly below the eye tracker, a Luxonis OAK-D[17] camera was positioned to record the participant. The interface interaction was facilitated using a Logitech Presenter, selected for its intuitive design and the ability to be used without diverting gaze from the screen, minimizing influence on gaze behavior. Consistent lighting was maintained across sessions to reduce the impact on pupil dilation. A height-adjustable table was used to optimize eye tracker accuracy by accommodating varying participant heights. Moreover, the participant-to-screen distance was controlled at 60cm.

3.5 Procedure

As participants arrived, they were first presented with a consent form, which they signed to acknowledge their voluntary participation and understanding of the study's nature. They were also asked to fill out a demographic questionnaire to collect relevant background information. Following the acquisition of consent, participants received a thorough briefing about the task at hand. This briefing included the key detail that the captions they were to evaluate had been generated by an Image Captioning Model. After making sure that participants fully grasped the task and its objectives, we introduced them to the study system. We then proceeded with the calibration of the eye tracking device, utilizing the Tobii Pro Eye Tracker Manager for a precise 9-point calibration process. To confirm the accuracy of the eye tracker, we manually checked the data using the provided gaze visualization tool, performing recalibrations when necessary.

Before beginning the main task, a training phase was conducted to familiarize participants with the study environment and procedure. A countdown on the interface was shown at the center of the screen. This ensured that all participants began their task with their gaze focused on the center. Upon completion of the countdown, an image-caption pair was presented. Participants were instructed to determine the correctness of the caption and then proceed to the next screen to enter their decision. Participants advanced to the main task phase once they confirmed their understanding of these instructions. Upon its completion, a debriefing session was conducted, and participants were compensated for their contribution to the study.

[1]https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-fusion, [Accessed 16-08-2024].

3.6 Dataset Overview

Our dataset consists of 4,620 samples, collected from 30 participants. In line with the study's design, the dataset was structured to achieve an even distribution of perceived correctness, targeting a 50/50 split between agreement and disagreement with the presented image caption pairs. This objective is reflected in the dataset, with 50.5% of the samples rated as incorrect (disagree) and 49.5% as correct (agree). The average accuracy compared to ground truth from the FOIL COCO dataset across all participants was 90.24% with a variance of 8.8%. Additionally, the response times across trials indicated an average duration of 5.16 seconds per trial, a median duration of 4.45 seconds, and a standard deviation of 2.92 seconds. Lastly, the robustness of the eye tracking data was confirmed, with a minimum recognized gaze signal rate of 91% and an average rate of 98.5%. The dataset is available at https://github.com/DFKI-Interactive-Machine-Learning/disagreement-detection-dataset.

4 Method

In the following section, we describe the preprocessing of the eye tracking data and the feature extraction method applied. The preprocessing was aimed at preparing the data for the feature extraction process, while the latter focused on identifying the attributes from the eye tracking data that could be used for classification. Additionally, the classification methods proposed for disagreement detection are outlined. This includes traditional machine learning algorithms as well as a deep learning method based on VTNet [29], designed specifically to process raw gaze data.

4.1 Data Preprocessing

During preprocessing, we addressed inconsistencies in the dataset's timestamps. Initially, these timestamps varied slightly, with differences ranging from 3.99 to 4.01 milliseconds. These were adjusted to a fixed interval of exactly 4 milliseconds to establish temporal consistency across the dataset. Gaze points were then computed as the average of the gaze coordinates obtained from each eye. In circumstances where data from one eye was missing, the gaze point was inferred from the coordinates of the other eye, ensuring continuous data representation.

For the categorization of gaze events into fixations and saccades, our implementation closely followed the methodology outlined by Tobii [20]. We classified fixations using the Identification by Velocity Threshold (I-VT)[25] fixation detection algorithm. This method classifies fixations as sequences of the raw gaze signal, where gaze velocity stays below a predefined threshold, indicating a relatively stable gaze. We chose the Savitsky-Golay[26] filter for calculating velocities, with an order of 2 and a span of 40 ms, following recommendations from literature [31]. A default velocity threshold of 20 degrees/s was applied. However, after manually inspecting the event detection results, we adjusted the threshold to 30 degrees/s

Table 1: Detailed Gaze, Pupil, and Transition Features. Features marked with an asterisk (*) are calculated per Area of Interest (AOI), including the whole visit, visits on the Image, and visits on the Caption.

Category	Description
*Fixation-based**	Total number of fixations Number of fixations per unit time Average duration of fixations Standard deviation of fixation durations Total duration of all fixations
*Saccade-based**	Average length of saccades Standard deviation of saccade lengths Average of relative angles of saccades Standard deviation of relative saccade angles Average of absolute saccade angles Standard deviation of absolute saccade angles
*Pupil**	Average width of the left pupil Standard deviation of left pupil width Maximum width of the left pupil Minimum width of the left pupil Average width of the right pupil Standard deviation of right pupil width Maximum width of the right pupil Minimum width of the right pupil Pupil width of the left eye at the first fixation Pupil width of the left eye at the last fixation Pupil width of the right eye at the first fixation Pupil width of the right eye at the last fixation
Transition	Number of transitions between AOIs

for three participants where data exhibited higher noise levels. In addition, for each fixation calculated, we determined the location of its occurrence in relation to the predefined AOIs—specifically, the image, the caption, or the background.

4.2 Features

Our features are retrieved from Barz et al. [2] and Lallé et al. [15]. These features are calculated across various segments of the user interface for the entire duration of each task, focusing on areas where users' attention is most indicative of their decision-making process. We concentrate on three AOIs: the whole screen, the image area, and the caption area, as shown in Figure 2.

The selection of these AOIs is strategic, as they represent key elements where user interaction is most telling of their agreement or disagreement. The whole screen provides a general overview of user engagement, the image area relates to visual content processing, and the caption area pertains to textual content processing.

The features are categorized into three groups: Fixation-based, Saccade-based, and Pupil-based, each providing a different perspective on the user's gaze behavior. Fixation-based features, for example, might indicate the points of interest or confusion, while saccade-based features could reflect the user's search patterns or hesitations. Pupil-based features offer an additional layer, potentially correlating with cognitive load or emotional response. Additionally, Transition features are included to capture the dynamic aspect of user interaction, tracking how users move between different AOIs. These movements can be revealing of how users process and evaluate the image-caption pairs.

Table 1 presents a detailed breakdown of these features, including parameters such as fixation count, fixation rate, mean and standard deviation of fixation durations, saccade lengths, angles, and various pupil dimensions. Each feature marked with an asterisk (*) indicates calculation on a per-AOI basis.

4.3 Feature-based predictor

The algorithms Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Logistic Regression (LR) were utilized to develop models for predicting user disagreement from eye-tracking features. These methods are widely recognized for their robust performance in handling eye-tracking data [2, 4, 15, 24]. For each algorithm, hyperparameter tuning was conducted to optimize the models.

For XGBoost, the hyperparameters that were considered included the number of estimators (100, 200, 300), learning rate (0.01, 0.1), maximum depth (3, 6), and minimum child weight (1, 2). RandomForest models were tuned using the number of estimators (100, 200, 300), maximum depth (None, 10, 20, 30), and the number of features considered at each split (None, 'sqrt'). For Logistic Regression, the hyperparameters included the regularization strength (C) with values 0.1, 1, 10, 100, and the type of penalty ('l1', 'l2').

4.4 The VTNet and VTNet_att models

In parallel with traditional machine learning algorithm models, we also employed a deep learning method that uses the raw gaze data. The *VTNet* model, initially presented in [29], was developed to detect user confusion by learning from raw Eye Tracking (ET) data. The model integrates a single-layer Gated Recurrent Unit (GRU) sub-model with a two-layer Convolutional Neural Network (CNN) sub-model, each operating independently. The GRU sub-model is responsible for processing the ET data sequentially, whereas the CNN sub-model processes its corresponding spatial representation by learning from scan path images, which is a representation of the ET samples' X and Y gaze coordinates and the transitions between them. In [30], a self-attention layer was added before the GRU to allow it to focus on the more important segments within the ET sequences, thereby enhancing the model's ability to discern long-term dependencies. This model, now termed *VTNet_att* to indicate the addition of attention, incorporates a self-attention layer with a dimensionality of 6 to match the dimensionality of the input sequential ET data and a single attention head to preserve model simplicity and computational efficiency, which is important to prevent overfitting on small datasets (a typical characteristic of ET datasets). The output from the GRU, characterized by a 256-unit hidden state, concatenates with a 50-element vector from the CNN to form a combined vector of 306 elements. This vector serves as the input for a simple neural network comprising a hidden layer and a SoftMax output layer, which classifies the input into two categories: Disagreement or Agreement. The hyperparameters of the *VTNet_att* model remain consistent with those specified in

[29] and [30], and the model undergoes end-to-end training as a cohesive unit.

As a data augmentation method, we cyclically split the eye tracking sequences, following the approach used by Sims and Conati [29] and Sriram et al. [30]. In these works, the cyclical splitting process produced four new data points from each original one by grouping samples collected at 120Hz that were four steps apart into the same new data point. This method preserved the temporal structure because contiguous samples showed minimal variation due to the high sampling rate, and it expanded the dataset fourfold. However, since our eye tracker had a higher sampling rate of 250Hz, we cyclically split each data point into eight separate ones, thereby increasing the dataset eightfold.

5 Evaluation

In this section, we present our experiment setup. We aim to explore the possibilities of predicting user disagreement through eye tracking data using traditional machine learning algorithms and a deep learning method. Our exploration is split into two key experiments: the *Cross-user* experiment and the *Within-user* experiment.

The *Cross-user* experiment attempts to determine the generalizability of the predictive model—can it effectively use eye tracking data from a pool of users to predict disagreement for any given user? This experiment will reveal the model's capability to apply learned patterns of disagreement from the collective data to unseen individuals.

Whereas the *Within-user* experiment focuses on the model's capacity to predict disagreement when trained and tested on data from the same user. Here, the aim is to understand how well a model can learn individual-specific patterns of eye movements and whether these personalized models lead to improved performance over generalized models.

The fundamental questions guiding our experiments are:

(1) Is eye tracking data a viable source of implicit feedback for predicting user disagreement with an AI-generated caption, and can such a prediction model generalize across different users?
(2) How effective are personalized models, tailored to individual users, when using eye tracking data to predict disagreement?

In both cross-user and within-user experiments, we statistically compared the results using a one-way MANOVA test, where the model type served as the independent variable, while the performance metrics acted as the dependent variables. For post-hoc pairwise comparisons, we used the Tukey's HSD test. We report statistical significance when the p-value is less than 0.05.

5.1 Cross-User evaluation

The Cross-User evaluation aims to assess the generalizability of our models in predicting user disagreement. To achieve this, we implemented a 10-fold Leave-Groups-Out cross-validation (CV) strategy. Under this validation method, the dataset was divided into 10 exclusive groups, with each group acting as a hold-out test set at different iterations. Every iteration ensured that a particular user's data was included in the test set just once, thus guaranteeing that the training set did not contain any data from the user being tested. This separation is vital for ensuring that our evaluation of the model's generalizability is not compromised by information leakage. In the context of feature-based predictors, such as Random Forest, Extreme Gradient Boosting, and Logistic Regression, we incorporated recursive feature elimination with cross-validation (RFECV) within the CV framework. This method is intended for feature selection and is executed in tandem with hyperparameter tuning on the training set.

For the VTNet model, which accepts raw gaze data without prior feature selection, no feature selection step was used within its validation loop. The model was evaluated based on its ability to learn from raw data as provided.

Table 2: Average Performance Metrics with Standard Deviation for Each Model

Model	F1 Score	Accuracy	Precision	Recall
LR	0.59 ± 0.04	0.53 ± 0.04	0.53 ± 0.05	0.70 ± 0.17
RF	0.59 ± 0.05	0.54 ± 0.03	0.53 ± 0.03	0.66 ± 0.12
XGBoost	**0.63** ± 0.04	**0.55** ± 0.01	0.54 ± 0.02	**0.76** ± 0.13
VTNet	0.53 ± 0.04	0.54 ± 0.03	0.53 ± 0.06	0.55 ± 0.10
VTNet_att	0.50 ± 0.07	**0.55** ± 0.04	**0.56** ± 0.07	0.47 ± 0.13

Algorithm 1 Cross-User evaluation

$X, Y, G \leftarrow$ Dataset, Targets, Groups
$M \leftarrow$ Set of Models
$H \leftarrow$ Hyperparameters for Models in M
$F \leftarrow$ Set of Features
for $(train, test) \in$ LeaveGroupsOut(G) **do**
  $X_{train}, Y_{train} \leftarrow X[train], Y[train]$
  $X_{test}, Y_{test} \leftarrow X[test], Y[test]$
  $F_{selected} \leftarrow$ FeatureSelection$(X_{train}, Y_{train}, F)$
  $X^{fs}_{train} \leftarrow X_{train}[F_{selected}]$
  $X^{fs}_{test} \leftarrow X_{test}[F_{selected}]$
  for $model \in M$ **do**
    $best\_model \leftarrow$ ParamTuning$(X^{fs}_{train}, Y_{train}, model, H)$
    Scores$_{model} \leftarrow$ Evaluate$(best_model, X^{fs}_{test}, Y_{test})$
  end for
end for
return Scores$_{model}$

5.1.1 Results. This section reports on the findings from our experiment designed to test the generalizability of various models in predicting user disagreement using eye tracking data. Table 2 presents the average performance metrics for each model across the 10-fold Leave-Groups-Out cross-validation setup, providing detailed insight into their precision, accuracy, F1 scores, and recall rates, along with the variability of these metrics.

The data in Table 2 highlights the average performance achieved by each model. The Logistic Regression model recorded an average F1 Score of 0.59 with a standard deviation of 0.04 and featured an average Recall of 0.70. Similarly, the Random Forest model demonstrated an average F1 Score of 0.59 and a Recall of 0.66. Among traditional machine learning models, XGBoost performed the best

with the highest average F1 Score of 0.63 and the highest average Recall of 0.76.

In contrast, the deep learning approaches, represented by VTNet and VTNet_att, exhibited worse performance metrics compared to the more traditional machine learning models. VTNet achieved an average F1 Score of 0.53 and a Recall of 0.55, while VTNet_att displayed slightly lower scores with an average F1 Score of 0.50 and a Recall of 0.47. Furthermore, the standard deviations for these models indicate a higher variability in performance, particularly for VTNet_att.

The MANOVA found significant effects of the type of model on the F1 Score($F_{4,45} = 8.620$, partial $\eta^2 = 0.434$) and Recall ($F_{4,45} = 6.254$, partial $\eta^2 = 0.357$). Pairwise comparisons showed that, based on the F1 scores, the XGBoost and Random Forest models were equivalent in performance and they both outperformed the other three models (Logistic Regression, VTNet, and VTNet_att). These three models were found to be equivalent to one another. In terms of Recall, the XGBoost, Logistic Regression, and Random Forest models were found to be equivalent to one another and they all outperformed the VTNet and VTNet_att models which, in turn, were equivalent in performance.

Algorithm 2 Within-User Model Training and Evaluation

```
P ← Set of Participants
D ← Dataset
M ← Models
for p ∈ P do
    d_p ← getDataOfParticipant(D, p)
    X_train, X_test, Y_train, Y_test ← Split(d_p)
    for model ∈ M do
        Train(model, X_train, Y_train)
        Evaluate(model, X_test,test )
    end for
end for
return Evaluation scores for each model
```

5.2 Within-User evaluation

To assess the effectiveness of personalized models for each individual user, we conducted an experiment where a distinct model was trained for each participant using the full array of methods previously introduced, including both feature-based and VTNet algorithms.

For this purpose, we organized the dataset comprising of 154 samples from each user, splitting them into training and testing subsets. The division designated 134 samples for training purposes, while 20 samples were set aside for evaluation, adhering to a pre-established split based on the image sets. Subsequently, a unique model employing both, the feature-based predictor as well as VTNet, was trained for each participant. This approach allows us to explore and compare the performance of models personalized to individual users. The specific steps for the training and evaluation process for each user are detailed in Algorithm 2.

5.2.1 Results. The summarized results of the within-user evaluations are presented in Table 3, which illustrates the average performance of all personalized models trained on data for a single user.

Table 3: Average Performance Metrics with Standard Deviation for Each Model

Model	F1 Score	Accuracy	Precision	Recall
LR	0.51 ± 0.19	0.54 ± 0.19	0.52 ± 0.30	0.38 ± 0.18
RF	0.50 ± 0.33	0.55 ± 0.24	0.52 ± 0.20	0.42 ± 0.20
XGBoost	0.57 ± 0.16	**0.59** ± 0.12	**0.57** ± 0.20	**0.62** ± 0.18
VTNet	0.55 ± 0.18	0.57 ± 0.13	0.55 ± 0.19	0.57 ± 0.21
VTNet_att	**0.58** ± 0.14	0.58 ± 0.12	**0.59 ± 0.15**	0.59 ± 0.17

For the Logistic Regression model, an average F1 Score of 0.51 ± 0.19 was obtained. The Random Forest model had an average F1 Score of 0.50 ± 0.33. XGBoost had an average F1 Score of 0.57 ± 0.16, with the highest average Accuracy of 0.59 ± 0.12 and the highest average Precision of 0.57 ± 0.20 among the feature-based models.

In the category of deep learning approaches, the VTNet model achieved an average F1 Score of 0.55 ± 0.18, while the VTNet_att model had an average F1 Score of 0.58 ± 0.14. The VTNet_att model also achieved the highest average Precision of 0.59 ± 0.15 when compared with other models. The standard deviation values indicate variability in model performance across different user data. However, the MANOVA test revealed that there was no significant effect of the type of model on any of the performance metrics. Hence, they were all statistically equivalent to one another.

6 Discussion

The results of the cross-user experiment reveal that among the models evaluated, XGBoost achieved the best performance. However, the overall average accuracy score of 0.55 is quite low. The high recall rate (0.76), suggests that the model is good at identifying most instances where the user disagrees with the caption. This means that most user disagreement instances are captured and can be used as feedback to the IML system. However, the low precision also implies that the model may produce false positives, i.e. predicting disagreement where there may be none. This can be problematic if the model's predictions are used directly to trigger requests for user feedback, potentially leading to an excessive number of interruptions. Such interruptions can decrease user experience by prompting users to provide feedback too frequently, particularly in cases where they might perceive the AI system's output as satisfactory.

Therefore, it might be beneficial to explore a different approach that incorporates additional implicit sources of information. By integrating data such as facial expressions, the model could gain insight into a wider range of implicit user feedback signals, possibly allowing for a more accurate determination of when to request explicit user input. With these enhancements, it would be possible to maintain the benefits of detecting disagreement for IML while reducing the risk of unnecessary feedback prompts.

In the within-user experiment, both the XGBoost and the VTNet models achieve the best average accuracies of 0.59 and 0.58, respectively. However, these results were accompanied by high variances in performance between users. Such variances are further emphasized by the considerable standard deviations between precision and all other metrics, as shown in Table 3.

Table 4: Comparative Performance Metrics for Top and Bottom 5 Users Based on Balanced Accuracy of XGBoost and Gaze Robustness Score

User ID	Balanced Accuracy	Gaze Robustness
B13	0.3000	**0.97230**
A06	0.4000	**0.94518**
B07	0.4141	**0.91343**
B11	0.4394	0.99513
B03	0.4596	**0.95736**
B05	0.7083	0.99408
A04	0.7143	0.97802
B14	0.7473	0.99496
B09	0.7500	0.99438
B01	0.8333	0.99935

XGBoost worked well for certain users and poorly for others, indicating the presence of individual differences in disagreement behavior (see Table 2). For 7 users, the models provided balanced accuracies exceeding 0.70, and more than half of the users had accuracies greater than 0.60. However, for other users we observed performance with balanced accuracies falling below 0.50, highlighting a potential disparity in how effectively models can capture indicative behavior.

To understand the variability in model performance, particularly for users where models underperformed (<.50), we examined metrics such as gaze robustness and noise level. We aimed to determine whether low performance could be associated with identifiable factors, such as a user's fixation on the background instead of the relevant areas of interest or inconsistency in gaze data. An analysis revealed that the accuracy of XGBoost has a moderate correlation (.43) with gaze robustness, a metric that assesses the quality and consistency of the gaze signal. Notably, 4 out of the 5 users with the worst model performance had gaze robustness scores that were lower than the group average (98.5). This finding suggests that gaze robustness may play a role in the effectiveness of the model in predicting user disagreement accurately.

In addition, we also investigated the influence of increasing training data using XGBoost, which has the best average performance. These results are graphically represented in Figure 3, where users' average accuracy is plotted against the increasing number of training samples provided to their personalized model and error bars showing the standard deviation. We found that, on average, there is a moderate increase in the XGBoost model's performance as the quantity of training data per user increases, indicating that additional training samples improve model accuracy. However, the high std suggests that there is inconsistent individual performance improvement.

6.1 Limitations

Although our study offers the first insight into the predictive capacity of gaze for user disagreement, it has its limitations. First, the data were collected in a controlled experiment setting which contrasts natural user interaction with AI, the latter of which can introduce a wider range of variability and complexity that is not fully replicated in a laboratory context. Second, our work does not explore alternative modeling techniques [18] or the integration of additional data sources (e.g., facial expressions), which could offer further dimensions to understanding user disagreement and improve model robustness. Lastly, the scope of the study was restricted to the context of image-caption pairs. The focus on image-caption pairs is context-specific and thus, future research should assess the transferability of these insights across different forms of Human-AI interaction to validate their broader applicability.

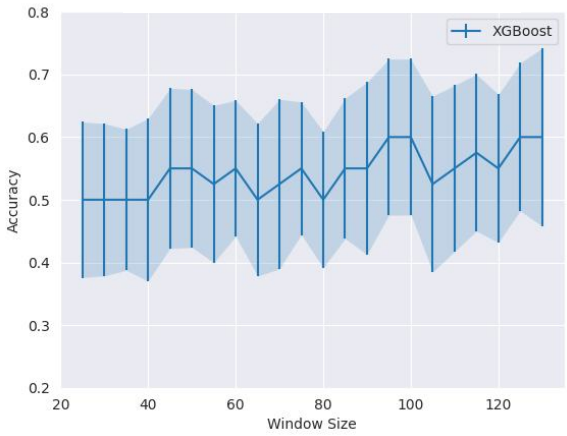

Figure 3: Average balanced accuracy performance of personalized models with increasing training samples

7 Conclusion & Future Work

This work investigated the potential of eye tracking signals as an implicit source of prediction of disagreement when interacting with an AI system. We collected a dataset with 30 participants in which they interacted with a simulated image captioning system, while their eye movements as they rated the AI-generated captions were recorded. We investigated the performance of machine learning models in both cross-user and within-user contexts. The findings indicate that while the best model (XGBoost) performs well in detecting instances of disagreement, its precision remains a challenge, underscoring the necessity for models that better capture individual user behaviors. Notably, XGBoost proved effective for some users while failing to capture the disagreement of others, demonstrating a disparity in model performance that should be further investigated. In conclusion, our research provides a first understanding of the relationship between eye movements and user disagreement in AI interactions. The integration of additional modalities and the application of advanced analytical techniques represent promising directions for future research.

Acknowledgments

This work was funded in part by the European Union under grant number 101093079 (MASTER), and the German Federal Ministry of Education and Research (BMBF) under grant number 01IW23002 (No-IDLE).

References

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (Dec. 2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

[2] Michael Barz, Omair Shahzad Bhatti, and Daniel Sonntag. 2021. Implicit Estimation of Paragraph Relevance From Eye Movements. *Frontiers Comput. Sci.* 3 (2021), 808507. https://doi.org/10.3389/fcomp.2021.808507

[3] Michael Barz, Sven Stauden, and Daniel Sonntag. 2020. Visual Search Target Inference in Natural Interaction Settings with Machine Learning. In *ETRA '20: 2020 Symposium on Eye Tracking Research and Applications, Stuttgart, Germany, June 2-5, 2020*, Andreas Bulling, Anke Huckauf, Eakta Jain, Ralph Radach, and Daniel Weiskopf (Eds.). ACM, 1:1–1:8. https://doi.org/10.1145/3379155.3391314

[4] Nilavra Bhattacharya, Somnath Rakshit, and Jacek Gwizdka. 2020. Towards Real-time Webpage Relevance Prediction UsingConvex Hull Based Eye-tracking Features. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 28, 10 pages. https://doi.org/10.1145/3379157.3391302

[5] Omair Bhatti, Michael Barz, and Daniel Sonntag. 2022. Leveraging Implicit Gaze-Based User Feedback for Interactive Machine Learning. In *KI 2022: Advances in Artificial Intelligence*, Ralph Bergmann, Lukas Malburg, Stephanie C. Rodermund, and Ingo J. Timm (Eds.). Springer International Publishing, Cham, 9–16.

[6] Nigel Bosch, Yuxuan Chen, and Sidney D'Mello. 2014. It's Written on Your Face: Detecting Affective States from Facial Expressions while Learning Computer Programming. In *Intelligent Tutoring Systems*, Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, and Kitty Panourgia (Eds.). Springer International Publishing, Cham, 39–44.

[7] Maya Cakmak, Crystal Chao, and Andrea L. Thomaz. 2010. Designing Interactions for Robot Active Learners. *IEEE Transactions on Autonomous Mental Development* 2, 2 (2010), 108–118. https://doi.org/10.1109/TAMD.2010.2051030

[8] Sidney K. D'Mello, Scotty D. Craig, and Art C. Graesser. 2009. Multimethod Assessment of Affective Experience and Expression during Deep Learning. *Int. J. Learn. Technol.* 4, 3/4 (oct 2009), 165–187. https://doi.org/10.1504/IJLT.2009.028805

[9] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 8 (jun 2018), 37 pages. https://doi.org/10.1145/3185517

[10] SIDNEY K D'Mello and Arthur C Graesser. 2014. Confusion. In *International handbook of emotions in education*. Routledge, 299–320.

[11] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology* 53, 4 (1987), 712.

[12] Maliheh Ghajargar, Jan Persson, Jeffrey Bardzell, Lars Holmberg, and Agnes Tegen. 2020. *The UX of Interactive Machine Learning*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3419249.3421236

[13] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. 2020. Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (Oct. 2020), 63–72. https://ojs.aaai.org/index.php/HCOMP/article/view/7464

[14] Lea Krause and Piek Vossen. 2020. When to explain: Identifying explanation triggers in human-agent interaction. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*. 55–60.

[15] Sébastien Lallé, Cristina Conati, and Giuseppe Carenini. 2016. Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 2529–2535.

[16] Jia Zheng Lim, James Mountstephens, and Jason Teo. 2020. Emotion Recognition Using Eye-Tracking: Taxonomy, Review and Current Challenges. *Sensors* 20, 8 (2020). https://doi.org/10.3390/s20082384

[17] luxonis. 2020. OAK-D: Stereo camera with Edge AI. https://luxonis.com/ Stereo Camera with Edge AI capabilities from Luxonis and OpenCV.

[18] Abdulrahman Mohamed Selim, Michael Barz, Omair Shahzad Bhatti, Hasan Md Tusfiqur Alam, and Daniel Sonntag. 2024. A review of machine learning in scanpath analysis for passive gaze-based interaction. *Frontiers in Artificial Intelligence* 7 (2024). https://doi.org/10.3389/frai.2024.1391745

[19] Sucheta Nadkarni and Reetika Gupta. 2007. A Task-Based Model of Perceived Website Complexity. *MIS Quarterly* 31, 3 (2007), 501–524. http://www.jstor.org/stable/25148805

[20] Anneli Olsen. 2012. The Tobii IVT Fixation Filter Algorithm description. https://api.semanticscholar.org/CorpusID:52834703

[21] Mariya Pachman, Amaël Arguel, Lori Lockyer, Gregor Kennedy, and Jason Lodge. 2016. Eye tracking and early detection of confusion in digital learning environments: Proof of concept. *Australasian Journal of Educational Technology* 32, 6 (Dec. 2016). https://doi.org/10.14742/ajet.3060

[22] Manuela Pollak, Andrea Salfinger, and Karin Anna Hummel. 2022. Teaching Drones on the Fly: Can Emotional Feedback Serve as Learning Signal for Training Artificial Agents? *arXiv preprint arXiv:2202.09634* (2022).

[23] Joni Salminen, Bernard J. Jansen, Jisun An, Soon-Gyo Jung, Lene Nielsen, and Haewoon Kwak. 2018. Fixation and Confusion: Investigating Eye-Tracking Participants' Exposure to Information in Personas. In *Proceedings of the 2018 Conference on Human Information Interaction I&' Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 110–119. https://doi.org/10.1145/3176349.3176391

[24] Joni Salminen, Mridul Nagpal, Haewoon Kwak, Jisun An, Soon-gyo Jung, and Bernard J. Jansen. 2019. Confusion Prediction from Eye-Tracking Data: Experiments with Machine Learning. In *Proceedings of the 9th International Conference on Information Systems and Technologies* (Cairo, Egypt) *(icist 2019)*. Association for Computing Machinery, New York, NY, USA, Article 5, 9 pages. https://doi.org/10.1145/3361570.3361577

[25] Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. 71–78.

[26] Abraham. Savitzky and M. J. E. Golay. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36, 8 (1964), 1627–1639. https://doi.org/10.1021/ac60214a047

[27] Abdulrahman Mohamed Selim, Omair Shahzad Bhatti, Michael Barz, and Daniel Sonntag. 2024. Perceived Text Relevance Estimation Using Scanpaths and GNNs. In *Proceedings of the INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24)* (San Jose, Costa Rica) *(ICMI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3678957.3685736

[28] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurelie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. "FOIL it! Find One mismatch between Image and Language caption". In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*. 255–265.

[29] Shane D Sims and Cristina Conati. 2020. A neural architecture for detecting user confusion in eye-tracking data. In *Proceedings of the 2020 international conference on multimodal interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 15–23. https://doi.org/10.1145/3382507.3418828

[30] Harshinee Sriram, Cristina Conati, and Thalia Field. 2023. Classification of Alzheimer's Disease with Deep Learning on Eye-tracking Data. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 104–113.

[31] Benjamin Voloh, Marcus Watson, Seth Konig, and Thilo Womelsdorf. 2020. MAD saccade: statistically robust saccade threshold estimation via the median absolute deviation. *Journal of Eye Movement Research* 12 (05 2020). https://doi.org/10.16910/jemr.12.8.3

[32] Jan Zacharias, Michael Barz, and Daniel Sonntag. 2018. A Survey on Deep Learning Toolkits and Libraries for Intelligent User Interfaces. arXiv:1803.04818 [cs.HC]

[33] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 1 (2009), 39–58. https://doi.org/10.1109/TPAMI.2008.52