Do You (Dis)agree With Me? Modelling Implicit User Disagreement in Human–AI Interaction Using Gaze Data

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

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

Amr Gomaa
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
Saarland Informatics Campus
Saarbrücken, Germany
amr.gomaa@dfki.de

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

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

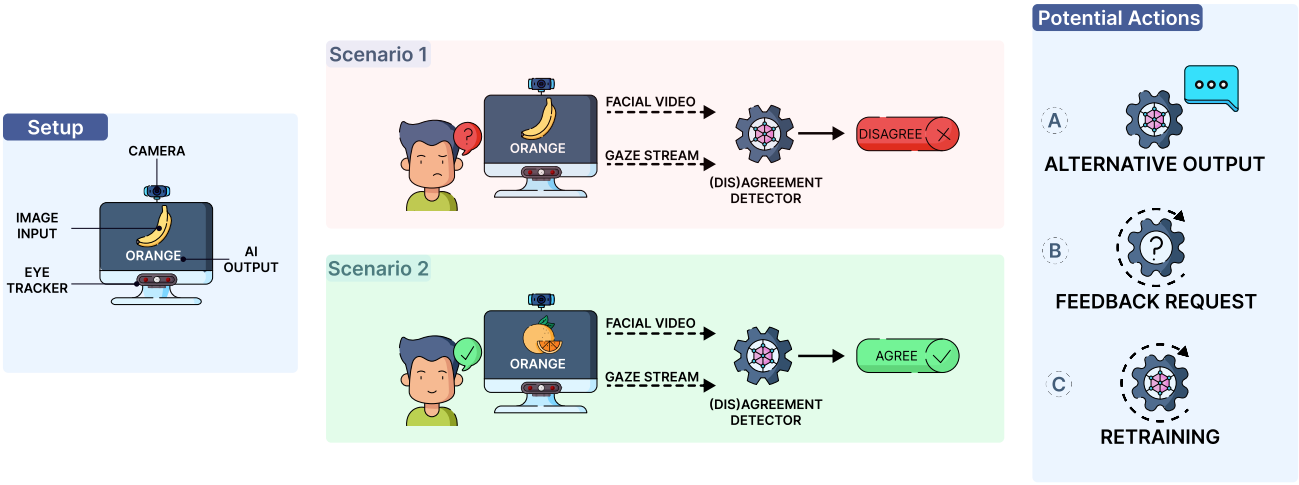


Figure 1: An overview and future outlook for implicit disagreement detection in human–AI interaction. The system records facial data using a webcam and gaze data using an eye tracker while a user views an AI prediction. The (dis)agreement detector fuses the two streams to infer whether the user agrees with the output: in Scenario 1 (banana captioned "ORANGE"), it predicts "DISAGREE"; in Scenario 2 (orange captioned "ORANGE"), it predicts "AGREE". When DISAGREE is detected, the system can trigger interventions by (A) suggesting alternative outputs, (B) requesting lightweight feedback, and (C) queuing the example for retraining.

*These authors contributed equally to this work.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790594

Abstract

The widespread use of generative AI has led to increased focus on human–AI interaction. However, AI systems can generate unexpected outputs, leading to disagreement or human–AI conflict. This paper focuses on modelling user disagreement using machine learning (ML) by observing users' implicit viewing behaviour. We conducted a controlled study with 30 participants evaluating captions from a simulated ML image-captioning system. Participants indicated agreement or disagreement with each caption while we recorded their gaze and facial-expression data, which we used to predict (dis)agreement. We show that unimodal gaze-based personalised modelling (0.684 average balanced accuracy) outperforms generalised modelling (0.570), whereas multimodal approaches did not improve performance. Our exploratory post hoc gaze-based analysis highlights the importance of feature selection and temporal dynamics, which help guide system design and future work. We release the dataset to support reproducibility and further work. Due to the nature of this research, we also discuss the potential ethical and privacy implications of continuous passive gaze and facial monitoring.

CCS Concepts

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

Keywords

Dataset, Disagreement Detection, Eye Tracking, Human–AI Conflict, Interactive Machine Learning, User Modelling

ACM Reference Format:

Abdulrahman Mohamed Selim, Omair Shahzad Bhatti, Amr Gomaa, Michael Barz, and Daniel Sonntag. 2026. Do You (Dis)agree With Me? Modelling Implicit User Disagreement in Human–AI Interaction Using Gaze Data. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3772318.3790594

1 Introduction

The rapid adoption of Artificial Intelligence (AI) in everyday life has increased attention to Human–AI interaction, encompassing end-user collaboration, as well as research approaches such as human-in-the-loop methods and Interactive Machine Learning (IML) [64]. Human–AI collaboration represents a joint activity in which humans offer intuition, domain expertise and creative direction, and AI provides computational support and data-driven suggestions [18]; this partnership allows users to steer the process and achieve outcomes that exceed what either partner could accomplish alone [65, 78, 96]. IML refers to methods that integrate user feedback into the Machine Learning (ML) training loop to adapt the system, enabling non-technical users to guide model behaviour through direct labelling or corrections and supporting rapid, incremental model updates [4, 35, 97, 104].

It is important to keep in mind that AI-powered systems can behave unpredictably and may produce outputs that are disruptive, confusing, or offensive [5]. Such failures can lead to user disagreement with system outputs, i.e., human–AI conflict, which is a state of mismatch between user expectations and system behaviour [36]. To mitigate this, systems should enable mechanisms for corrective feedback and behaviour redirection [5]. At the same time, frequent explicit requests for feedback can frustrate users and reduce their trust in the system [15, 39]; these interruptions can also disrupt the natural flow of interaction and induce increased cognitive load [69, 102]. This raises an important question: **How can systems sense user disagreement without requiring explicit input?** A potential solution is to use passive, i.e., implicit, feedback channels that sense user disagreement without requiring conscious user action. Passive interaction is a system-initiated monitoring of a user's state that does not require explicit input from the user and can capture subtle cues of their reaction [63]. Nonetheless, detecting disagreement from passive cues is challenging because affective and cognitive responses do not always manifest as visible signals, and expressions of agreement or disagreement are typically multimodal, requiring the combination of multiple signal sources [13, 85, 86].

In this study, we investigate how to implicitly detect user disagreement with AI-generated outputs. We conducted a controlled user study in which 30 participants evaluated the perceived correctness of captions within a simulated ML image-captioning task. The captions contained deliberate errors sourced from the FOIL-COCO dataset [88]. As passive input modalities, we collected: (1) facial video recordings as facial expressions can provide cues for predicting disagreement [85, 86]; and (2) eye-tracking and pupil data, which have been shown to reflect underlying cognitive processes [62, 77, 80], such as confusion [90]. We designed our initial set of experiments to address the following research questions:

- **RQ1:** Can perceived disagreement with image–caption outputs be reliably detected from passive gaze and facial signals collected during a simulated human–AI interaction task?
- **RQ2:** How do personalised and generalised disagreement-detection models compare in terms of accuracy, reliability, and robustness to inter-participant variability?

Our initial results showed that multimodal approaches incorporating facial data did not outperform gaze-only models. To better understand the predictive signals within the gaze data, we conducted more exploratory post-hoc analyses using two approaches: *feature selection* and *time-window selection*. **Feature selection**, a critical step in machine learning experiments [100], enabled us to identify the gaze behaviours most predictive of disagreement. **Time-window selection** was motivated by the temporal nature of cognitive-affective processes. When inspecting a stimulus, gaze behaviour has been shown to unfold in stages; for example, in visual search tasks, gaze progresses from an initial orientation phase to deeper analysis [98]. More broadly, gaze patterns evolve as users process information and anticipate task-relevant details [109], highlighting the dynamic role of gaze in decision-making [75]. This temporal progression mirrors neurobiological models of decision-making, such as preference assessment and action selection [26], each with its own characteristic temporal signature [50]. Accordingly, we hypothesised that the predictive power of gaze patterns would vary across the viewing window. Therefore, we investigated which temporal periods most strongly predict users' final judgments of caption correctness. We formulated the following additional research questions to guide our post-hoc analysis:

- **RQ3-A:** Which gaze-based features are most predictive of perceived disagreement, and how consistent are these predictive features across participants?
- **RQ3-B:** Which time window prior to a participant's decision best discriminates perceived disagreement using passive gaze signals, and how does the duration of this window impact classification performance?

In this paper, we present our methods and results, followed by a discussion of limitations and implications for designing adaptive human–AI systems. This work takes an initial step towards modelling implicit disagreement, with the aim to enhance the responsiveness and trustworthiness of interactive AI systems. To this end, our primary contributions are: (1) a publicly available dataset of processed eye-tracking and facial-tracking data from 30 participants with binary annotations of agreement or disagreement; (2) exploratory findings into how gaze-based feature selection and time-window selection influence disagreement detection; and (3) an evaluation of personalised and generalised machine learning models for disagreement detection. It is important to note that our findings are exploratory. Although personalised gaze-only analyses improved performance, overall accuracy remained modest, reflecting the difficulty of detecting subtle (dis)agreement signals from passive cues. Therefore, we do not present an interactive system; instead, we provide our dataset and modelling approach as a baseline, together with insights and lessons learned to inform future research and system design. The full dataset is publicly available on GitHub to support reproducibility and further work.

2 Related Work

As AI systems become more widely adopted, situations in which their outputs conflict with human expectations or preferences have become increasingly apparent, a phenomenon termed human–AI conflict. Flemisch et al. [36] defined human–AI conflict as a state of misalignment, opposition, or incompatibility between humans and AI systems. Jiang et al. [41] further explains that such conflicts may arise in both collaborative and competitive contexts; this leads to two main forms of human–AI conflict: task conflict, which relates to concrete issues such as different goals or decision-making strategies, and relationship conflict, which occurs due to factors such as differences in values or interaction styles.

To model these conflicts effectively in human–AI interaction, it is important to understand disagreement from human behaviour. Our assumption is that behavioural patterns observed in human–human disagreement can inform the design of systems for detecting similar signals in human–AI scenarios; this view is based on research in human–robot and virtual-agent interaction showing that principles of human–human interaction can be transferred to agent design [2, 31]. This section begins by examining disagreement from psychological and user perspectives, then briefly shows how gaze has been used as a passive interaction modality through various applications, and finally provides examples of how implicit feedback signals have been used in ML applications.

2.1 Understanding Disagreement

Disagreement is a complex phenomenon to model, partly because we could not find a universally accepted definition. For computational purposes, disagreement is often defined as the belief that one holds an opinion contrary to that of an interlocutor [13, 72]. Fundamentally, disagreement represents an oppositional stance to a preceding action or proposition, involving differing opinions [7, 16, 70]. Within our study, we can view disagreement as reflecting a perceived semantic mismatch between visual information and its accompanying textual description.

2.1.1 Disagreement as a Cognitive Appraisal. Another approach to defining disagreement is to understand how disagreement manifests. We draw on Scherer's appraisal theory [85, 86], which characterises disagreement as a cognitive process that can trigger affective responses under certain conditions. In our context, we can understand that disagreement emerges when participants assess the alignment between an image and its caption, forming a judgement of agreement or disagreement. These assessments or appraisals can generate actions that show up as behavioural expressions. While disagreement typically involves low-intensity appraisals rather than full-blown emotions, a noticeable mismatch may trigger subtle affective states such as confusion or frustration, potentially leading to micro-expressions, e.g., furrowed brows [86]. However, such facial signals vary considerably across individuals and cultures, making detection based solely on facial expressions a challenging task.

2.1.2 Disagreement as a Communicative Act. Although appraisal theory clarifies the internal evaluation process, it does not fully explain how disagreement is expressed and communicated. To address this gap, we need to consider disagreement as a communicative act as well, which signals a misalignment in understanding. Based on Clark's concept of common ground [25], we recognise that shared understanding between conversational partners is built through interaction. In our human–AI interaction, we assume that common ground is established as the user interprets both the visual stimulus and the AI-generated text. Disagreement, therefore, signals a breakdown in this shared understanding, i.e., a perceived gap between the user's mental model and the system's output [24, 25]. When explicit feedback channels are absent, users may communicate this misalignment implicitly through their behaviour.

This perspective of disagreement as a communicative function is further supported by research in conversation analysis, which identifies disagreement as a *dispreferred response* [46, 94], i.e., an action that deviates from the expected flow of an interaction. Dispreferred responses are characterised by distinctive patterns in timing and multimodal behaviour, such as delays or gaze aversion [71]. These patterns serve a practical purpose by implicitly communicating resistance to the established common ground. Building on findings from human–human interaction [93], we expect that when users encounter a mismatch with AI output, their implicit behaviours will signal their assessment, and resist the system's interpretation.

2.1.3 Multimodal Nature of Disagreement. Given that disagreement is both an internal evaluation and a communicative act, it is inherently a multimodal phenomenon [13]. Ekman's framework on basic emotions [32] emphasises that while disagreement does not qualify as a basic emotion, it may co-occur with emotions such

as anger or disgust when the perceived mismatch is personally relevant. These emotions have distinct physiological and expressive signatures, often decoded through the activation of facial Action Units (AU), which are the basic components of facial expressions and provide a reliable framework for interpreting emotional states [23]. Bousmalis et al. [13] concluded in their survey that no single cue, such as a head nod or shake [44], can be reliably matched to disagreement. Instead, disagreement emerges from the interplay and combination of multiple signals over time, including vocal cues, gestures, and facial expressions [13, 46].

To detect such multimodal disagreement signals, we need to consider how human–human social interaction relies on multiple channels of communication, such as speech and body language; participants combine multiple cues to coordinate actions and establish alignment, which is a fundamental requirement for successful communication [46, 71, 92]. Interactional misalignment occurs when an action is inappropriate for the current context, which could disrupt the interaction's progress [52, 71]. Within this multimodal system, gaze is often regarded as an important signal for conveying disengagement and other complex social attitudes [46, 79, 94]. People monitor the gaze of their partners to infer their state of engagement, and recent work has consistently shown that disagreeing responses correlate strongly with gaze aversion [45, 52, 71]. Building on this evidence that misalignment is systematically expressed through gaze patterns, we hypothesise that gaze-based markers will characterise disagreement in human–AI interaction [52].

2.1.4 Our Understanding of Disagreement. Building on all these perspectives, we conceptualise disagreement as an internal cognitive appraisal of a semantic mismatch when viewing image–caption pairs, which manifests externally as a multimodal communicative act. Therefore, our study employs both gaze and facial analysis to detect these implicit signals, building on the established roles of gaze in signalling communicative misalignment and facial expressions in conveying cognitive-affective states.

2.2 Gaze-based Implicit User Feedback

Building on its established role in social communication [45, 52, 71], this section reviews the literature on using eye-tracking technology to infer user states. Eye tracking is a technology that records eye movements and gaze locations over time [17]. It has long been used as a passive input modality to infer user states across various use cases [67]. Gaze indicates where attention is directed and how information is processed. Researchers have applied gaze to predict perceived relevance estimation [8, 10, 11, 68], predict confusion [54, 82, 83, 90], and model signals and states such as measuring confidence [91], stress [48], cognitive workload [53], and need for assistance [3]. These studies show that gaze can serve as a reliable passive signal for monitoring cognitive and behavioural states.

Gaze is also used in affective computing, both in unimodal and multimodal setups. Several studies report that gaze patterns relate to emotions such as anger or happiness [19, 43, 55, 89]. Combining gaze with physiological and behavioural measures, for example, electroencephalography (EEG), electrocardiography (ECG), electrodermal activity, and facial AUs, proved successful for tasks such as emotion recognition [1, 22, 42, 106], mind-wandering detection [14], and user modelling [34]. These findings further show that complex cognitive-affective states are best captured through multimodal interaction, and gaze can be successfully utilised for such applications. As described in Section 2.1, disagreement often co-occurs with states such as uncertainty and confusion, and it can elicit subtle affective responses; therefore, gaze-based features for monitoring different cognitive and emotional states can be reasonable proxies for detecting a perceived disagreement, which is the focus of this work.

2.3 Implicit User Feedback in Machine Learning

Using human feedback to improve ML systems is an established paradigm, especially in reinforcement learning and human–robot interaction [21, 49, 60, 101]. However, these approaches traditionally require explicit user input to guide a model's behaviour. As previously noted, frequent requests for such explicit feedback can frustrate users, disrupt the natural flow of interaction, and increase cognitive load [15, 39, 69, 102]. To mitigate these issues, a growing body of work has focused on leveraging implicit user feedback as a more natural signal for creating adaptive and interactive ML systems.

Pollak et al. [73] investigated emotional feedback as a reinforcement signal for training an agent, using facial emotion recognition to infer user satisfaction and adapt a virtual drone's behaviour; although their findings suggest that emotional cues can be integrated into reinforcement learning, they also highlight challenges related to variability in emotional intensity across individuals. Similarly, Krause and Vossen [51] proposed leveraging signs of user confusion or uncertainty as triggers for providing explanations in human–AI interaction, arguing that systems should proactively respond to implicit indicators rather than waiting for explicit queries. Additionally, Xu et al. [107] and Kim et al. [47] used EEG-detected error potentials (i.e., brief EEG signals that appear when the brain notices a mistake or error events) as a direct feedback source to accelerate reinforcement learning tasks. Beyond these examples, other studies have examined how naturally occurring reactions, including gestures and facial expressions, can serve as rich feedback channels for improving agent performance without imposing additional cognitive load on users [27].

Building on the concept of implicit feedback from a different perspective, Sumers et al. [95] treated natural language instructions as a source for implicit feedback, using sentiment analysis to understand instructions and infer the underlying reward function. Broadening the definition of feedback further, Dennler et al. [29] demonstrated that monitoring users' exploration of a system can serve as a source of feedback; they employed contrastive learning to infer user interests from which robot behaviours users chose to investigate, thereby creating a model of preferences without requiring any direct input.

Although this review is not exhaustive, it highlights a clear trend toward improving human–AI interaction by leveraging implicit signals, such as gaze and facial cues, to interpret affective states. This presents a promising direction for building adaptive systems capable of responding to disagreements in real-time. The main challenge of our task lies in the complex nature of disagreement, as established in our theoretical grounding in Section 2.1.

3 Data Collection Study

We conducted a within-subjects user study to investigate how to implicitly detect user disagreement with AI-generated outputs using gaze and facial data. To elicit disagreement, we used an image captioning task in which participants were told that the captions were generated using an ML model. This design prompted disagreement when captions contained errors or were otherwise inappropriate.

We obtained ethics approval from the ethical review board of the Faculty of Mathematics and Computer Science at Saarland University prior to data collection. Our dataset contains data from 30 participants (21 male, 9 female; mean age 26.4 years). Participants were recruited via email and university campus postings, and the resulting skewed gender distribution reflects the composition of the respondent pool from this convenience sample. All participants were fluent in English, had normal or corrected-to-normal vision, and 19 participants reported prior eye-tracking experience. The study lasted around 60 minutes, and the participants were compensated at a rate of 15 Euros per hour. As previously mentioned, the processed dataset is publicly available on GitHub to support reproducibility and further work.

3.1 Apparatus

We recorded the eye-tracking data using a Tobii Pro Fusion[1] operating at 250 Hz, which follows prior recommendations for reliable saccade detection ($\geq$ 120 Hz) [56] and for computing saccade-based features ($\geq$ 200 Hz) [6]. Additionally, we captured the facial video recordings using a Luxonis OAK-D camera[2] operating at 30 Hz. Both devices were connected to the same laptop, and their timestamps were synchronised to the laptop's clock.

We maintained consistent lighting throughout the sessions to reduce effects on pupil size. We used a height-adjustable table to optimise eye-tracker accuracy for different participant heights. The eye tracker and camera were mounted on the screen to record each participant, and the participant-screen distance was fixed at 60 cm.

3.2 Stimuli

We used 154 image–caption pairs from the FOIL-COCO dataset [88]; it pairs each MS-COCO dataset image [58] with a correct caption and a foil caption, i.e., a caption that has exactly one incorrect (foil) word. We assigned participants to two groups, Group A and Group B; both groups viewed the same images, but caption assignment was counterbalanced so that when Group A saw the correct caption for an image, Group B saw the corresponding foil, and vice versa. Figure 2 shows an example of the stimuli we used, where if Figure 2a was used for Group A, then Figure 2b was used for Group B. We included deliberate single-word errors in half of the captions (i.e., 77 out of 154) to reduce class imbalance. We randomised the image sequence for each participant to reduce order effects. We included at least two images from each of the dataset's 73 categories, to ensure coverage across different types of images and foil words. In addition, we selected captions of approximately ten words and standardised image resolution across stimuli.

[1]https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-fusion (Accessed August 18, 2025)
[2]https://shop.luxonis.com/products/oak-d (Accessed August 18, 2025)

We decided to use FOIL-COCO because it contains real photographs, and its single-word foils had been validated by human annotators. Therefore, it allowed us to measure how small changes to a single word can make participants disagree with a caption. FOIL-COCO, therefore, provides naturalistic, tightly controlled stimuli that isolate semantic mismatch while preserving grammatical form and sentence context. Our counterbalancing preserves identical visual input across conditions, so differences in behavioural responses can be attributed to caption content rather than image properties.

3.3 Procedure & Task

Upon arrival, we explained the study to the participants and then presented them with a consent form, which they signed to confirm voluntary participation and understanding of the study. They then completed a demographic questionnaire to provide relevant background information. Next, participants were introduced to the study system, and we calibrated the eye tracker using the Tobii Pro Eye Tracker Manager with a 9-point calibration procedure. Calibration accuracy was verified manually using the gaze visualisation tool, and recalibration was performed when necessary.

Participants were informed that **"The captions were generated using an ML model"** and for each image–caption pair, they had to provide a binary judgement of **agree** or **disagree** based on their perception of the caption's correctness before proceeding to the next trial. The task was self-paced to avoid inducing stress from time constraints. Participants could proceed to the next image–caption pair immediately after providing their rating, spending on average 4.75 ± 2.18 seconds (mean $\pm$ SD) per trial. Before starting the main task, participants completed a short training phase to familiarise themselves with the interface and procedure, ensuring they fully understood the task and its objectives before proceeding. Once participants confirmed their understanding of these instructions, they advanced to the main task. Each trial began with a countdown displayed at the centre of the screen to ensure their initial gaze fixation was on the centre. Following this, an image–caption pair appeared, and participants made a binary judgement of **agree** or **disagree** regarding the perceived caption's correctness. After completing the main task, participants took part in a debriefing session and were compensated for their time.

4 Methods

In this study, we modelled participants' perceived disagreement with the output of an ML model within an image captioning task, as described in Section 3.3. As established in Section 2.1, disagreement is commonly treated as a multimodal phenomenon, which motivated us to collect our dataset in a multimodal fashion. However, the collected facial data did not lead to any meaningful improvements; instead, it resulted in a drop in performance when combined with the gaze data. To keep the manuscript focused on our research questions, this section concentrates on the gaze modality. However, for completeness and to support reuse of the dataset, we provide additional details in the appendix for the facial data processing in Appendix B and for our initial multimodal setup in Appendix C. Below, we describe the data pre-processing, feature extraction, and the ML methods used for analysis.

(a) Stimulus with correct caption **(b) Stimulus with foil caption**

Figure 2: An example from the FOIL-COCO dataset [88], showing the correct and foiled captions for the same stimulus.

4.1 Gaze Data Pre-Processing & Event Detection

For the collected eye-tracking data, we resampled timestamps to a fixed 4 ms interval to ensure temporal consistency (original range: 3.99-4.01 ms). Gaze coordinates were computed as the mean of both eyes' gaze data streams; when one eye's data stream was missing, we used the other eye's stream alone to preserve continuity.

Gaze points are usually categorised into different event types. **Fixations** describe the state when the eyes remain relatively still for a time period lasting somewhere between a few tens of milliseconds up to a few seconds, while **saccades** are the rapid eye movements from one fixation to another [38]. The combination of alternating fixation and saccadic events produces a **scanpath** [12], which refers to the trace of a person's eye movements across space over time [38].

We identified fixation and saccadic events using the Dispersion-Threshold Identification (I-DT) algorithm [84]. In this method, gaze points are grouped into a fixation if their spatial dispersion (i.e., the maximum spread between points within a time window) remains below a predefined threshold for at least a minimum duration; otherwise, the points are classified as part of a saccade. We set the minimum fixation duration to 50 ms. To determine the dispersion threshold for each participant, we implemented an adaptive procedure based on the Median Absolute Deviation (MAD) of gaze velocities. In short, we computed the MAD of each participant's velocity distribution to quantify noise: a larger MAD indicated greater velocity variability, leading to a higher dispersion threshold, whereas a smaller MAD indicated more stable velocities, allowing a lower threshold. We manually selected user-specific thresholds by maximising the proportion of samples labelled as fixations while balancing over- and under-segmentation. The full adaptation procedure is described in Algorithm 1. Additionally, we merged consecutive fixations separated by fewer than 20 pixels to reduce fragmentation. As a verification step, we cross-validated event labels against Area of Interest (AOI) mappings, for example, to distinguish text from image fixations.

The final gaze output consists of event timelines with labelled fixations and saccades, event durations, and AOI associations. We validated these outputs using visual overlays on the stimuli and quantitative metrics such as cluster scores. Additionally, two eye-tracking experts performed manual checks on a number of trials to ensure quality.

4.1.1 Dataset Summary & User-Generated Ground Truth. After pre-processing, we obtained 4,092 valid samples from 30 participants across both modalities. Although the stimuli were evenly split between correct and foil captions, participants did not achieve perfect identification. On average, they correctly distinguished between correct and foil captions with an accuracy of 0.886 ± 0.092, ranging from a maximum of 0.974 to a minimum of 0.455 when comparing their agree/disagree labels with the ground truth. Since our focus is on participants' perceived agreement or disagreement, we used their self-generated labels as user-generated ground truth data, regardless of correctness, as these reflect their actual judgments.

4.2 Gaze-based Feature Extraction

We extracted gaze-based features informed by prior research on cognitive state monitoring. Following a review publication [67], we included features (n=17) that appeared in two or more relevant studies, as these represent widely adopted indicators of visual attention and cognitive processing. These features capture fixation behaviour, saccadic dynamics, and pupil responses, which together provide a comprehensive view of how users allocate attention and process visual information. Fixation-based features, such as fixation count and duration, reflect the depth of processing and attentional focus; saccade-based features, including amplitude and velocity, indicate search strategies and information scanning; and pupil-based features are often linked to cognitive load and emotional arousal. To complement these, we incorporated additional features (n=22) from related work [8, 54, 76]. These include transition metrics between

Algorithm 1: Adaptive dispersion threshold optimisation using MAD

Input: Gaze samples $\{(x_t, y_t, \tau_t)\}_{t=1}^{T}$; base threshold D_{base}; tuning parameter α (default 0.1); small constant ε

Function *ComputeVelocity*($\{(x_t, y_t, \tau_t)\}_{t=1}^{T}, \varepsilon$)
- Remove samples with NaN in x, y, or τ; // Clean input data
- $\Delta x \leftarrow \text{diff}(x)$; // $\Delta x_t = x_t - x_{t-1}$
- $\Delta y \leftarrow \text{diff}(y)$; // $\Delta y_t = y_t - y_{t-1}$
- $\Delta \tau \leftarrow \text{diff}(\tau)$; // $\Delta \tau_t = \tau_t - \tau_{t-1}$
- Set $\Delta \tau_i = \varepsilon$ where $\Delta \tau_i = 0$; // Avoid division by zero
- $V \leftarrow \left[\sqrt{\Delta x_i^2 + \Delta y_i^2} / \Delta \tau_i \right]_{i=1}^{T-1}$; // Instantaneous speed
- **return** V;

Function *MAD(V)*
- Remove samples with NaN in V; // Clean input data
- **if** *V is empty* **then**
  - **return** 0; // No valid velocity intervals
- $m \leftarrow \text{median}(V)$; // Robust central tendency
- $D \leftarrow [\ |v - m| \ \forall v \in V\ ]$; // Absolute deviations
- **return** median(D); // Median absolute deviation (MAD)

Function *Main()*
- **Input:** observed gaze sequence $\mathcal{S} = \{(x_t, y_t, \tau_t)\}_{t=1}^{T}$, base threshold D_{base}, tuning α, constant ε;
- **Goal**: compute an adapted dispersion threshold D_{adapt} that scales with observed velocity dispersion;
- **Step 1 (velocity computation)**: $V \leftarrow \text{ComputeVelocity}(\mathcal{S}, \varepsilon)$;
  // Produces a vector of speeds between consecutive valid samples
- **Step 2 (dispersion estimation)**: $mad \leftarrow \text{MAD}(V)$;
  // Computes the median absolute deviation of V
- **Step 3 (threshold adaptation inline)**: $D_{\text{adapt}} \leftarrow D_{\text{base}}$; // Start from base threshold
- **if** $mad > 0$ **then**
  - $D_{\text{adapt}} \leftarrow D_{\text{base}} \times (1 + \alpha \times mad)$;
  - // Increase proportionally to velocity dispersion
- // If $mad = 0$, leave D_{adapt} at D_{base} (no adaptation)
- **return** D_{adapt}; // For downstream fixation detection

AOIs to capture how users navigate between the image and caption, and detailed pupil statistics that may reveal subtle variations in engagement. This combination ensures both coverage of established indicators and inclusion of measures relevant to our experimental setting. Table 1 summarises the full feature set.

4.3 Machine Learning Processing

Our initial investigation explored the use of multimodal gaze and facial data across different ML algorithms (see Appendix C). We found that several deep and transfer-learning architectures failed to generalise reliably, a common challenge with small datasets such as ours. Therefore, we focused on using statistical features from gaze (see Section 4.2) and facial (see Appendix B) data with classical ML algorithms (e.g., Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Linear Discriminant Analysis (LDA), and Random Forest (RF)). However, this multimodal approach revealed that adding facial features offered no performance gain over gaze-only baselines and introduced instability across fusion strategies. Based on these findings, we refined our methodology to focus on classical supervised learning algorithms using only gaze-based features. This focused approach provides a more thorough investigation of the feature and time-window effects central to this work.

To systematically evaluate a wide range of ML algorithms and accelerate our workflow, we used an Automated Machine Learning library, called PyCaret[3]. It is an open-source ML library that automates model training and evaluation. It is built on established frameworks, such as scikit-learn[4], and supports model selection, hyperparameter tuning, and performance comparison. We used PyCaret to evaluate 18 different classification algorithms under identical conditions and to rank them based on performance. We

[3]https://pycaret.gitbook.io/docs (Accessed August 18, 2025)
[4]https://scikit-learn.org/stable/ (Accessed 18 Feb 2025)

Table 1: Gaze-based features extracted per trial. The Source column cites the literature where each feature was reported. The [+ per AOI] indicates that the features were computed both globally across the entire stimulus and on a per AOI basis. For features with statistics in parentheses (e.g., "Mean, Std, Total"), all listed summary statistics are computed for that feature.

Feature Type	Feature	Units	Source
Fixation-based	Fixation Count (Total) [+ per AOI]	Count (unitless)	[8, 54, 67]
	Fixation Duration (Mean, Std, Total) [+ per AOI]	Time (ms)	[8, 54, 67]
	Fixation Rate per Second	Count/s	[54]
Saccade-based	Absolute Saccade Angles (Mean, Std)	Visual angle (deg)	[54]
	Relative Saccade Angles (Mean, Std)	Visual angle (deg)	[54]
	Saccade Count (Total)	Count (unitless)	[67]
	Saccade Duration (Mean, Std, Total)	Time (ms)	[67]
	Saccade Length (Mean, Std)	Visual angle (deg)	[8, 54, 67]
	Saccade Velocity (Max)	Deg/s	[67]
Scanpath-based	Scanpath Length (Total)	Visual angle (deg)	[67]
	Scanpath Duration (Total)	Time (ms)	[67]
AOI Transition-based	Dwell Time Before First Transition	Time (ms)	[76]
	Transitions Count Between AOIs	Count (unitless)	[54]
	Transitions Ratio Between AOIs	Ratio (unitless)	[54]
Pupil-based	Pupil Diameter (Mean, Std)	mm	[67]
	Right and Left Pupil Diameter (Min, Max, Mean, Std)	mm	[54]
	Right and Left Pupil Diameter During First and Last Fixation	mm	[54]

applied a Robust Scaler[5] to rescale the features. The scaler subtracts the median and scales the data according to the Interquartile Range (IQR), which is the range between the 1st quartile (25th percentile) and the 3rd quartile (75th percentile).

We used balanced accuracy as our primary evaluation metric because our aim was to detect perceived agreement and disagreement with a model's output. The F1-score is unsuitable in this context, as it does not account for true negatives and therefore assesses only the model's performance on the positive class. Although overall accuracy considers both classes, it can be misleading when class distributions are imbalanced; therefore, we used balanced accuracy (Equation 1), which assigns equal weight to the positive and negative classes. In addition to the balanced accuracy, we computed recall, precision, and the area under the ROC curve (AUC). For our use case, we assigned disagreement as the positive class (i.e., 1) and agreement as the negative class (i.e., 0).

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\underbrace{\frac{\text{TP}}{\text{TP}+\text{FN}}}_{\text{Sensitivity}} + \underbrace{\frac{\text{TN}}{\text{TN}+\text{FP}}}_{\text{Specificity}}\right) \tag{1}$$

5 Experimental Design & Evaluation

To investigate the detection of implicit user disagreement in our simulated human–AI interaction image-captioning task, we designed two sets of experiments. In our initial experiments, we explored the multimodal nature of disagreement by analysing both gaze and facial signals, as motivated by the literature (Section 2.1). Second, we compared the performance of personalised (within-participant) versus generalised (between-participant) models for disagreement detection. These experiments were designed to answer the following research questions:

- **RQ1:** Can perceived disagreement with image–caption outputs be reliably detected from passive gaze and facial signals collected during a simulated human–AI interaction task?
- **RQ2:** How do personalised and generalised disagreement-detection models compare in terms of accuracy, reliability, and robustness to inter-participant variability?

The findings from this initial phase motivated a deeper investigation into gaze-based modelling. Our second set of experiments focused solely on gaze data, examining the effects of feature and time-window selection on model performance. These additional experiments address the following research questions:

- **RQ3-A:** Which gaze-based features are most predictive of perceived disagreement, and how consistent are these predictive features across participants?
- **RQ3-B:** Which time window prior to a participant's decision best discriminates perceived disagreement using passive gaze signals, and how does the duration of this window impact classification performance?

5.1 Initial Experiments

We began with a leave-users-out 5-fold cross-validation, i.e., we trained and fine-tuned a model on data from a subset of participants and tested it on data from a separate set of participants. In this procedure, each participant's data were used exclusively for either training, validation, or testing. By using 5-fold cross-validation, we ensured that each participant's data appeared in the test set at least once. These experiments assess generalisability among participants by testing models on participants who were not present in the training subset. Afterwards, to investigate a personalised approach, we

[5] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html (Accessed August 18, 2025)

Table 2: The best generalised results from the initial gaze-based experiments using group-based 5-fold cross-validation.

BA (%)	F1 (%)	Recall (%)	Precision (%)	AUC (%)
54.42	59.77	62.92	56.91	56.08
53.75	52.83	61.09	46.53	54.57
53.16	58.38	61.87	55.25	53.81
59.52	69.16	90.52	55.95	62.64
54.56	60.78	63.27	58.49	57.02
55.08 ± 2.54	60.18 ± 5.88	67.93 ± 12.66	54.63 ± 4.69	56.82 ± 3.49

performed a user-based 5-fold cross-validation, i.e., we trained and tested models on each participant's data separately. This evaluates personalisation, i.e., whether we can model an individual and correctly classify their future, unseen samples. In all experiments, we used stratified folds to maintain balanced label distributions across training and test sets. We also applied a feature-selection strategy using both `SequentialFeatureSelector` and `SelectKBest`[6].

5.1.1 Results. Our initial experiments produced three sets of results: (1) multimodal results from models that combined statistical features from both gaze and facial data; (2) unimodal facial results from models using only facial features (see Appendix B); and (3) unimodal gaze results from models using only gaze-based features.

Multimodal Results. For the multimodal approach, we combined gaze and facial data using an early fusion strategy, i.e., concatenating the features from both modalities before passing them to the ML model (see Appendix C for more details). However, the generalised models remained close to chance level, with an average balanced accuracy of 51.75% ± 0.86% (chance ≈ 50.00%). The personalised models performed slightly better, reaching an average balanced accuracy of 54.00% ± 6.93%. However, only 7 out of 30 participants achieved a balanced accuracy above 60.00% across the evaluated ML algorithms. The detailed results are shown in Table D.1 and Table D.3 in Appendix D.

Facial Data Results. For the unimodal processing of the **facial data**, the generalised models produced chance-level performance, with an average balanced accuracy across the five folds of 49.94% ± 1.13%. The personalised models were only slightly better, with an average balanced accuracy of 51.76%±6.37%. Only three participants achieved balanced accuracies above 60.00%. The detailed results are shown in Table D.2 and Table D.4 in Appendix D.

Gaze Data Results. For the unimodal processing of the **gaze data**, the generalised models performed slightly above chance, with an average balanced accuracy of 55.08% ± 2.54%, as shown in Table 2. The personalised models showed a modest improvement, with an average balanced accuracy of 57.29% ± 6.67%. Across the different ML algorithms, ten participants achieved balanced accuracies above 60.00%.

From these results, we can see that personalisation led to consistent but limited performance gains. The unimodal gaze-based models achieved the highest balanced accuracies, although most remained below 60.00%. In contrast, facial data alone yielded chance-level performance, and when combined with gaze, it reduced performance compared to gaze-only models. Therefore, we conducted further experiments on personalised, gaze-based models to better understand why a bigger subset of participants achieved balanced accuracies above 60.00% and whether their performance can be modelled more effectively using gaze features.

[6]https://scikit-learn.org/stable/api/sklearn.feature_selection.html (Accessed August 18, 2025)

5.2 User Modelling Experiments & Post-hoc Analysis

Building on our initial results (Section 5.1), we explored, in these additional experiments, how feature and time-window selection affect user modelling to better understand the variability in gaze-based disagreement signals. To test whether these factors improved performance, we repeated the same setups: **generalised models** evaluated across participants on the whole dataset and separately for Group A and Group B, and **personalised models** trained and tested on each participant's data.

Table 3: Statistically significant features with a checkmark indicating membership in $\mathcal{F}_A$, $\mathcal{F}_B$, and $\mathcal{F}_{Pool}$. Note that the bilateral pupil features are listed once but counted as two separate features (Right, Left) in the subset sizes.

Feature	$\mathcal{F}_A$	$\mathcal{F}_B$	$\mathcal{F}_{Pool}$
Fixation Count	✓	✓	
Fixation Count per Image	✓	✓	
Fixation Count per Word	✓	✓	
Fixation Duration		✓	✓
Fixation Duration per Image	✓	✓	✓
Fixation Duration per Word		✓	
Relative Saccade Angles (Std)		✓	
Saccade Count		✓	
Saccade Duration		✓	
Saccade Length (Mean)		✓	
Scanpath Length	✓	✓	✓
Scanpath Duration	✓	✓	✓
Transitions Count Between AOIs	✓	✓	
Pupil Diameter (Std)	✓	✓	
Right and Left Pupil Diameter (Max)		✓	
Right and Left Pupil Diameter (Std)	✓	✓	
Right and Left Pupil Diameter During First Fixation		✓	

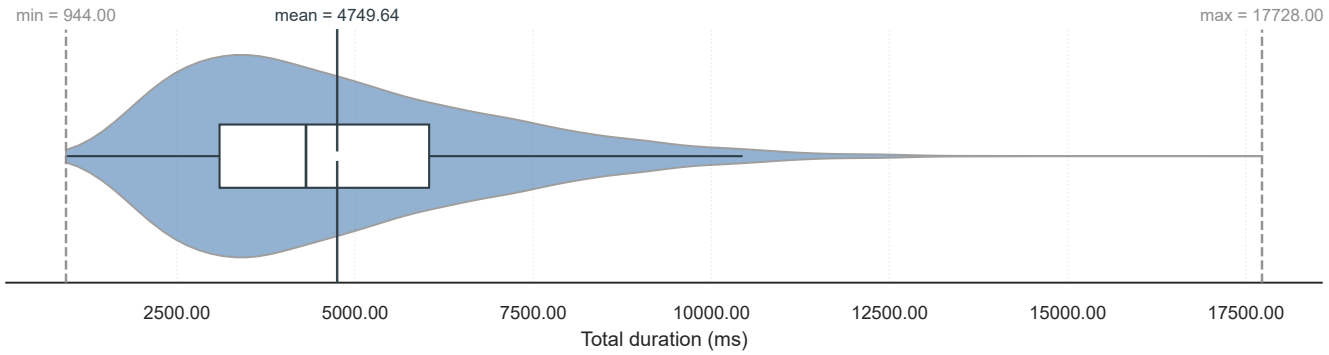


Figure 3: Violin plot for the total gaze duration of all samples across participants in milliseconds. Total gaze duration refers to the time spent on an image–caption pair from its onset until the participant's agree/disagree response.

5.2.1 Feature Selection. The automatic feature selection using `SequentialFeatureSelector` and `SelectKBest`, which we employed in the initial experiments (Section 5.1), did not lead to above-chance-level performance. Therefore, we decided to manually assess the features for statistical significance. We examined the distribution of each feature for the *agree* and *disagree* labels across the entire dataset and for each participant individually. To check for normality, we applied the Shapiro–Wilk test [87], and based on the outcome, we used either a paired t-test for normally distributed data or the Wilcoxon signed-rank test [103] for non-parametric cases (Algorithm A.1 in the Appendix outlines the logic behind this analysis.). Because we conducted multiple tests for each condition, we applied the Benjamini–Hochberg correction [9] to control the false discovery rate at $\alpha = 0.05$ and adjust the p-values.

When analysing participants individually, only seven participants had statistically significant features: three from Group A with 10 features in total, and four from Group B with 20 features in total. However, when analysing the dataset as a whole, we identified four statistically significant features. Based on these results, we defined three feature subsets: $\mathcal{F}_A$ contains features that were significant for at least one participant in Group A; $\mathcal{F}_B$ contains features that were significant for at least one participant in Group B; and $\mathcal{F}_{Pool}$ contains features that were significant across the dataset as a whole. Let F be the set of all the extracted gaze features, and let $\mathcal{P}_A$ and $\mathcal{P}_B$ be the participants within Group A and Group B, respectively. For each participant p, let $S_p \subseteq F$ be the set of features that are statistically significant for the agree–disagree contrast when testing p's data individually. Let $S_{\text{pool}} \subseteq F$ be the set of features that are statistically significant when the entire dataset is analysed jointly (all participants together). We define

$$\mathcal{F}_A = \bigcup_{p \in \mathcal{P}_A} S_p, \qquad \mathcal{F}_B = \bigcup_{p \in \mathcal{P}_B} S_p, \qquad \mathcal{F}_{Pool} = S_{\text{pool}}.$$

In our data, the subset sizes are $|\mathcal{F}_A| = 10$, $|\mathcal{F}_B| = 20$, and $|\mathcal{F}_{Pool}| = 4$ as shown in Table 3.

5.2.2 Time-window Selection. We began by examining gaze durations across participants. The distribution showed a 25th percentile (Q1) of approximately 3 seconds, a 75th percentile (Q3) of approximately 6 seconds, and a small number of long-duration trials extending up to approximately 18 seconds (as shown in Figure 3). To address this variability, we applied a time-window selection strategy, resulting in three conditions: **(1) using the full recording**, **(2) truncating to the last 3 seconds**, and **(3) truncating to the last 11 seconds**. If a recording exceeded 3 or 11 seconds, we retained only the final segment of that length, while recordings shorter than these thresholds were kept intact. The choice of 3 seconds was guided by the first quartile (Q1), representing shorter but typical viewing times, while 11 seconds was derived from Tukey's upper fence ($Q3 + 1.5 \times IQR$) [99], which captures the upper bound before extreme outliers. This approach enabled us to minimise the impact of unusually long recordings without discarding valid data.

5.2.3 Results. In this section, we first present the ML classification performance for both personalised (within-participant) and generalised (between-participant) models. We then report a post-hoc evaluation conducted to better understand the effects of personalisation, feature selection, and time-window selection.

Machine Learning Experiments. The generalised models achieved an average balanced accuracy of $\approx$ 57.00%, as shown in Table 4. This represents a small improvement over our initial generalised experiments (shown in Table 2), but performance still remained below 60.00%.

In contrast, the personalised models reached an average balanced accuracy of 68.40%, outperforming both the generalised models and the personalised models from our initial experiments. As summarised in Table 5, these personalised results aggregate outcomes across feature subsets and time windows. The best performance was obtained with a fully personalised setup, where, for each participant, both the feature subset and time window were selected individually. Under this setup, one participant achieved an average balanced accuracy of 59.30%, while the remaining 29 participants all exceeded 60.00%, with one reaching 90.00%, resulting in the overall average balanced accuracy of 68.40%. The detailed results for each participant are shown in Table A.1 in the Appendix. The personalised models also demonstrated a stronger ability to identify disagreement than the generalised model, achieving high F1-scores

Table 4: The best generalised results using the group-based 5-fold cross-validation for the dataset as a whole, Group A separately, and Group B separately. The method column contains the time window and the feature subset, where $\mathcal{F}_A$ contains features that were significant for at least one participant in Group A; $\mathcal{F}_B$ contains features that were significant for at least one participant in Group B; and $\mathcal{F}_{Pool}$ contains features that were significant across the dataset as a whole.

Data	Method	BA (%)	F1 (%)	Recall (%)	Precision (%)	AUC (%)
Whole Dataset	11 Seconds & $\mathcal{F}_A$	57.00	60.40	64.73	57.00	57.70
Group A	11 Seconds & $\mathcal{F}_B$	57.30	56.80	59.30	57.70	58.30
Group B	Full Recording & $\mathcal{F}_B$	57.40	47.90	42.30	59.80	60.60

Table 5: Personalised results averaged across participants for each feature subset and temporal condition combination. The personalised combinations consist of the unique feature subset and temporal information for each participant that produced their best performance. The method column contains the time window and the feature subset, where $\mathcal{F}_A$ contains features that were significant for at least one participant in Group A; $\mathcal{F}_B$ contains features that were significant for at least one participant in Group B; and $\mathcal{F}_{Pool}$ contains features that were significant across the dataset as a whole.

Method	BA (%)	F1 (%)	Recall (%)	Precision (%)	AUC (%)
Personalised Combinations	**68.40 ± 5.70**	**68.80 ± 8.30**	70.10 ± 10.80	**70.50 ± 7.70**	**69.10 ± 6.10**
11 Seconds & $\mathcal{F}_A$	62.10 ± 5.30	63.50 ± 8.20	65.20 ± 11.80	64.70 ± 6.70	61.80 ± 7.80
11 Seconds & $\mathcal{F}_B$	62.00 ± 6.20	62.80 ± 10.80	64.90 ± 12.90	63.50 ± 9.80	63.60 ± 8.60
11 Seconds & $\mathcal{F}_{Pool}$	62.00 ± 6.50	62.80 ± 9.30	64.50 ± 12.60	63.60 ± 7.30	62.30 ± 8.60
3 Seconds & $\mathcal{F}_A$	61.60 ± 6.00	63.00 ± 11.90	66.90 ± 16.50	62.30 ± 10.90	60.40 ± 8.70
3 Seconds & $\mathcal{F}_B$	63.10 ± 5.70	64.70 ± 9.70	68.10 ± 14.20	64.80 ± 8.70	62.90 ± 7.90
3 Seconds & $\mathcal{F}_{Pool}$	61.80 ± 5.10	65.50 ± 8.00	**72.40 ± 13.10**	62.80 ± 6.30	63.20 ± 7.20
Full Recording & $\mathcal{F}_A$	61.60 ± 6.30	63.80 ± 8.90	67.70 ± 13.00	63.10 ± 9.80	62.50 ± 7.20
Full Recording & $\mathcal{F}_B$	62.50 ± 7.20	63.80 ± 8.10	65.90 ± 10.50	65.00 ± 8.50	62.50 ± 8.00
Full Recording & $\mathcal{F}_{Pool}$	61.60 ± 6.10	63.00 ± 8.80	66.40 ± 11.80	63.10 ± 7.30	62.90 ± 9.10

(68.80%) and Recall (70.10%). Their high Precision (70.50%) indicates that most predicted disagreement labels were correct, with few false positives. In contrast, the generalised model, despite using the entire dataset and achieving an F1-score of 60.40%, had a Precision of only 57.00%, suggesting it struggled to identify disagreement accurately and produced many false positives.

Post-hoc Statistical Analysis. To determine whether the observed improvement in model performance was statistically significant and whether it was influenced by specific conditions (feature subset or time-window selection), we conducted several post-hoc tests. We applied a Bonferroni correction[7] at $\alpha = 0.05$ to adjust p-values for multiple comparisons, using the conservative approach to control the risk of false positives.

To assess the effects of feature subsets and temporal information, we performed a two-way repeated-measures Analysis of Variance (ANOVA)[8]. This analysis examined the impact of time-window selection (i.e., full recording, last 11 seconds, last 3 seconds) and feature subset ($\mathcal{F}_A, \mathcal{F}_B, \mathcal{F}_{Pool}$), as well as their interaction, on balanced accuracy. Each participant contributed data to all factor combinations, allowing us to account for within-subject variability. The results showed no significant main effects of gaze duration ($F(2,58) = 0.0602$, $p = 0.9416$) or feature subset ($F(2,58) = 0.6636$, $p = 0.5189$). The interaction between gaze duration and feature subset was also not significant ($F(4,116) = 0.2977$, $p = 0.8790$). These findings indicate that neither factor, alone or combined, explained a meaningful proportion of the variance in balanced accuracy. In other words, there was no single condition that consistently outperformed others; individual differences in the best-performing condition for each participant masked any overall effect.

In order to evaluate whether each participant's identified best-performing condition (feature subset and time window combination) represented a genuine and consistent advantage rather than random chance, we compared it with the participant-wise average across the remaining conditions and with the participant's second-best condition. The Shapiro–Wilk test [87] indicated that the distribution of paired differences deviated significantly from normality ($W = 0.7937$, $p = 0.0001$). Therefore, we used a Wilcoxon signed-rank test [103], which showed that the best condition produced significantly higher accuracies (mean increase $\approx 7.183\%$, median increase $\approx 6.953\%$) across all participants ($W = 465.0$, $p < 0.0001$). The effect size was maximal, reflecting a consistent directional difference. A second Wilcoxon test comparing the best condition with the second-best condition also revealed a significant advantage for the best condition (mean increase $\approx 2.425\%$, median increase $\approx 1.735\%$; $W = 465.0$, $p < 0.0001$), again with a maximal effect size.

[7] https://www.statsmodels.org/dev/generated/statsmodels.stats.multitest.multipletests.html (Accessed August 18, 2025)
[8] https://www.statsmodels.org/dev/generated/statsmodels.stats.anova.AnovaRM.html (Accessed August 18, 2025)

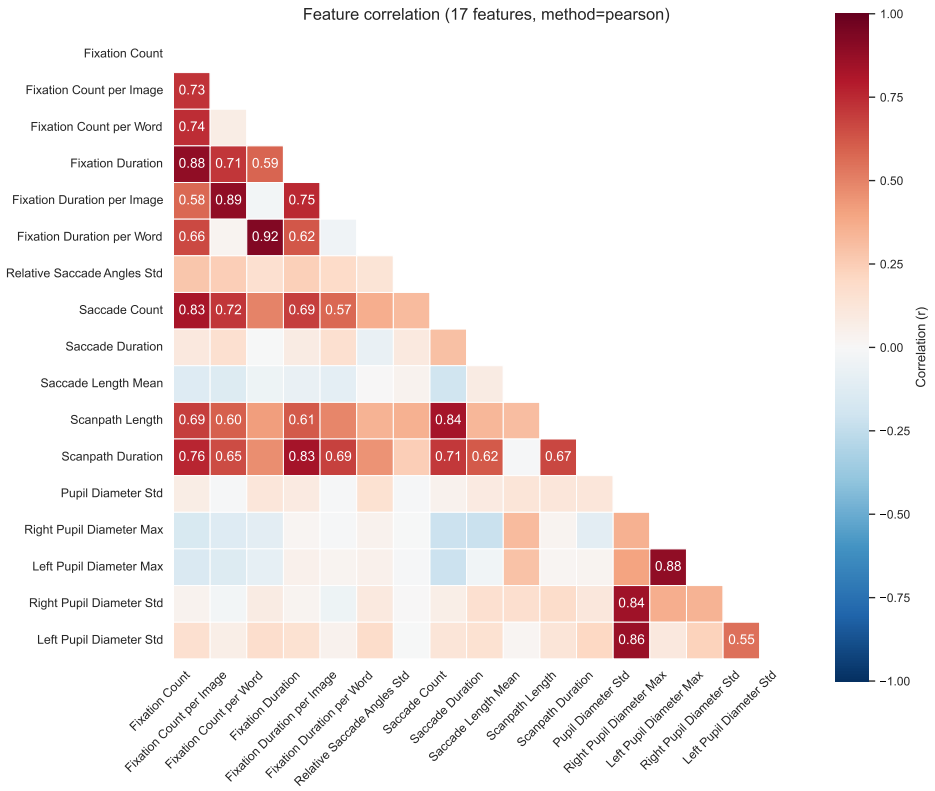


Figure 4: The feature correlation matrix for $\mathcal{F}_B$, with values above 0.5 being shown.

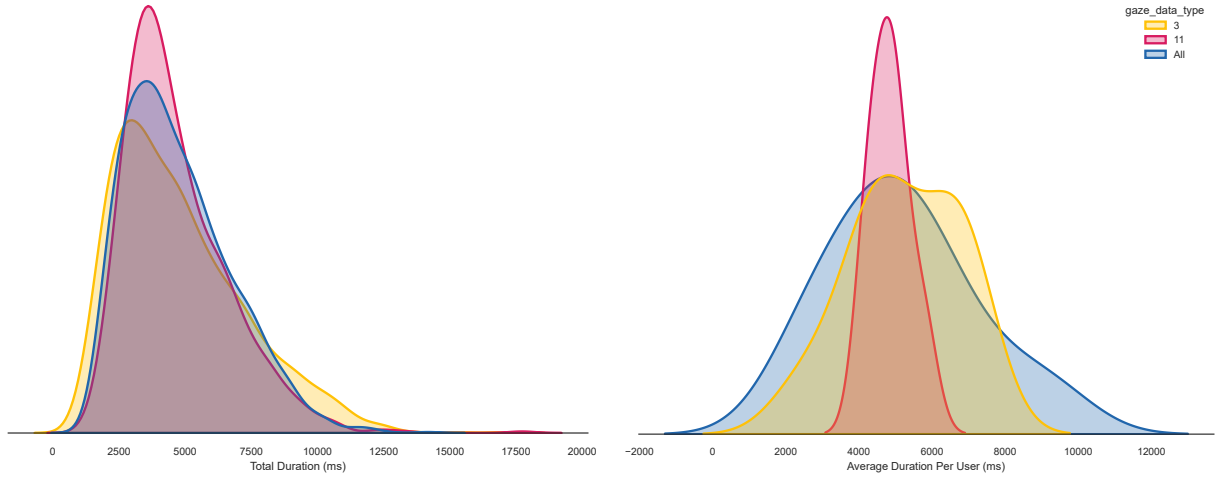
Finally, we conducted permutation-based paired comparisons[9] to confirm these findings without relying on distributional assumptions. For the comparison between the best condition and the average of the remaining conditions, the observed mean difference was ≈ 7.18%, with a permutation test indicating a highly significant effect ($p < 0.0001$, $r = 1.486$). Similarly, comparing the best condition with the second-best condition yielded a mean difference of ≈ 2.43%, with a similar permutation result ($p < 0.0001$, $r = 1.754$). These results showcase that the best-performing condition for each participant was significantly better than both the average and the second-best condition, confirming the reliability and consistency of these maximum accuracies.

Post-hoc Feature Analysis. To understand why manual feature selection based on statistical analysis outperformed the automatic feature selection used in our initial experiments, we examined the resulting feature subsets in more detail. We had three different feature subsets, i.e., $\mathcal{F}_A$, $\mathcal{F}_B$, and $\mathcal{F}_{Pool}$, as explained in Section 5.2.1. Figure 4 shows the correlation matrix for $\mathcal{F}_B$ ($\mathcal{F}_A$ and $\mathcal{F}_{Pool}$ are shown in Figure A.1 in the Appendix).

Across the different feature types (i.e., fixation-based, saccade-based, scanpath-based, AOI transition-based, and pupil-based), we observe strong internal correlations. For example, in $\mathcal{F}_B$, fixation counts and durations correlate highly with each other (r ranging from 0.580 to 0.920), and scanpath duration is strongly linked to fixation metrics (r ranging from 0.650 to 0.830). Similarly, pupil-based features form a tightly correlated group, with left and right pupil diameter measures reaching $r \approx 0.960$. In contrast, saccade-based features, such as relative saccade angles and transitions between AOIs, show weaker correlations, suggesting they may provide complementary information. Importantly, correlations between pupil and oculomotor features (i.e., fixation, saccade, and scanpath) remain low, indicating that pupil-based features contribute distinct information. $\mathcal{F}_A$ further reinforces these observations. Fixation counts and durations correlate strongly, while scanpath duration aligns closely with fixation counts ($r \approx 0.930$). Pupil features again cluster tightly (r ranging from 0.920 to 0.950), but show little correlation with fixation-based and saccade-based features. $\mathcal{F}_{Pool}$, a compact set of four duration and scanpath-based features, also exhibits high internal correlation (fixation duration total and scanpath duration total, $r = 0.860$), confirming that even minimal subsets capture overlapping dynamics.

Post-hoc Time-window Analysis. We examined three time-window selection strategies: the full recording, the last 11 seconds, and the last 3 seconds of each trial, as explained in Section 5.2.2. Among the best-performing personalised models, 13 participants achieved their highest balanced accuracy with the full recording, 6 participants with the 11-second window, and 11 participants with the 3-second window, as shown in Table A.1 in the Appendix.

[9] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.permutation_test.html (Accessed August 18, 2025)



(a) Total stimuli durations

(b) Mean stimuli duration per participant

Figure 5: Density plots for the total durations across users, and mean duration per user, coloured based on the temporal condition of their best performing combination. Participants within the 3-second window group tend to have longer total durations, suggesting discriminative cues occur near the end; those best within the 11-second window group show tighter average durations, indicating that trimming the initial seconds within longer trials improves the signal quality; those within the full recording had the widest duration range, implying a heterogeneous or cumulative decision process.

The density plots shown in Figure 5 indicate that participants who performed best with 3-second windows tended to have longer overall stimulus durations. In contrast, those with 11-second windows exhibited more consistent durations with an average of five seconds and a narrower total duration range. Participants who performed best with full recordings displayed the widest range of durations.

6 Discussion

In this study, we explored implicit disagreement detection using a multimodal gaze and facial dataset from 30 participants, with a more in-depth analysis of the gaze modality. We conducted two main experiments: the first involved using the full stimulus recording with automatic feature selection (Section 5.1) in order to understand the multimodal nature of disagreement, and evaluate personalised (within-participant) versus generalised (between-participant) models; the second focused more on the gaze data, exploring the effects of personalising the time-window selection (i.e., either the full recording, the last 11 seconds, or the last 3 seconds) and selecting feature subsets based on statistical analysis (Section 5.2). All experiments were carried out using both generalised processing with group-based 5-fold cross-validation, and personalised processing using stratified 5-fold cross-validation on each participant's data separately. In this section, we interpret our findings with respect to each research question. In the next section (i.e., Section 7), we critically reflect on the study's limitations and discuss their implications for future work in detail.

RQ1: Can perceived disagreement with image–caption outputs be reliably detected from passive gaze and facial signals collected during a simulated human–AI interaction task?

As established in Section 2.1, disagreement is a complex phenomenon to model. It can be seen as a cognitive-affective state [85, 86] that typically manifests as a communicative act [24, 25, 46, 94] and is inherently multimodal [13, 46]. Facial expressions [13, 23, 32, 46] and gaze behaviour [45, 52, 79, 94] are both thought to play an important role in implicitly conveying or expressing disagreement. However, our results did not fully align with this multimodal expectation, suggesting that our current computational approach did not effectively capture multimodal disagreement signals.

In the initial experiments (Section 5.1), gaze-only models achieved slightly higher balanced accuracies (personalised $\approx$ 57.00%, generalised $\approx$ 55.00%) than the multimodal models that combined gaze and facial features (personalised $\approx$ 54.00%, generalised $\approx$ 51.00%). Facial data on its own led to chance-level performance in both personalised and generalised settings ($\approx$ 50.00%). This pattern suggests that, in our setup, including facial features did not provide useful information and resulted in deteriorated performance relative to gaze-only models. Building on these observations, the additional experiments focusing on gaze-based models (Section 5.2) showed that using a fully personalised pipeline, i.e., optimising both feature subsets and time windows per participant, achieved improved performance ($\approx$ 68%) than any of our initial configurations.

To better understand our results, we manually inspected various facial video samples for each participant. For most participants, facial expressions remained *neutral* throughout the experiment, and no clear differences were visible between disagreement and

agreement trials. For two of the three participants with personalised facial models achieving above 60% balanced accuracy (A08 and A16), we observed subtle cues such as *eyebrow movements*, *frowning*, or *lip movements* during some disagreement trials, with one participant (A08) occasionally showing *smirks*. However, the third participant, with above 60% accuracy (B10), did not exhibit similar patterns. These observations suggest that the captions did not consistently elicit distinctive facial expressions, which likely limited the predictive power of facial features in our task. When these facial features were concatenated with the more informative gaze features in our multimodal setup, they may have introduced additional variance and diluted the useful gaze signal, which may help explain why the approach underperformed compared to gaze-only models. Therefore, we view the limited contribution of the facial data as a limitation of our experimental setup, likely due to our processing choices or the subtle nature of our stimuli with single-word caption errors, rather than evidence against the multimodal nature of disagreement, which is conceptually grounded and supported in the literature.

Overall, our fully personalised, gaze-only pipeline achieved a mean balanced accuracy of approximately 68.00%, which remains a modest improvement and does not yet support fully automated disagreement detection in real-world interactive systems. More generally, what constitutes *sufficient* performance is application-dependent, as it depends on both the downstream action triggered by the detector and the relative costs of false positives (unnecessary interventions) versus false negatives (missed disagreements). If disagreement predictions trigger disruptive system behaviour (e.g., interruptive alerts or frequent feedback requests), false positives become particularly costly. Prior work on workplace notifications shows that users acknowledge alerts as disruptive, yet still tolerate them when the perceived value is high, implying that any practical deployment must carefully control when and how the system interrupts the user [40]. In contrast, if the detector is used more conservatively as a *human-in-the-loop* signal that enables lightweight, user-controllable interventions (e.g., offering an alternative output or requesting feedback only under high confidence), then moderate reliability may still provide practical value as a feasibility component rather than a decisive automation module. In Section 7, we discuss more concrete future directions to help transition from our lab-based experiments towards fully deployed, interactive human–AI systems. We conclude that while implicit disagreement can be detected above chance level using passive gaze signals, reliable detection suitable for real-world applications remains an open challenge. It is also important to note that any practical deployment of continuous gaze and facial monitoring raises significant ethical concerns about privacy and user comfort and would require careful data minimisation, transparency, and user control.

RQ2: How do personalised and generalised disagreement-detection models compare in terms of accuracy, reliability, and robustness to inter-participant variability?

Our findings reveal a clear distinction between personalised and generalised modelling approaches. When focusing on the results of our additional experiments, the generalised models achieved only modest performance (BA $\approx$ 57.00%, F1 $\approx$ 60.00%, as shown in Table 4), and produced many false positives and near-chance behaviour; the personalised models, on the other hand, achieved substantially improved performance (mean BA = 68.40% $\pm$ 5.70%, F1 = 68.80% $\pm$ 8.30%, Recall = 70.10% $\pm$ 10.80%, Precision = 70.50% $\pm$ 7.70%). This improvement was statistically significant when using Wilcoxon signed-rank and permutation tests, which confirmed that each participant's best-performing condition was reliably superior to both the average and the second-best condition (mean increase $\approx$ 7.18%, median $\approx$ 6.95%; $p < 0.0001$).

The two-way repeated-measures ANOVA found no significant main effects of feature subset or time window, nor any interaction, indicating that no single combination consistently outperformed others across participants. Instead, the optimal configuration was highly personalised: some participants benefited from short temporal windows, others from full recordings, and the most effective feature subset also varied between participants. This heterogeneity explains why generalised comparisons among participants across fixed conditions yielded non-significant results, whereas personalised comparisons showed strong effects. This pattern aligns with eye-tracking research showing that individuals differ systematically in how they distribute their gaze [57, 66, 74, 108], and with different studies emphasising that emotions and cognitive states are inherently subjective [23, 86, 98]. Together, these perspectives help explain why gaze-based disagreement detection was more successful with personalised models.

These findings highlight the importance of personalisation for robust disagreement detection. The variability among participants means that fixed, one-size-fits-all models are unlikely to generalise well. At the same time, a personalised approach introduces additional computational and design considerations such as per-user calibration. From a practical perspective, the per-user calibration, such as selecting the best feature–temporal configuration from a small validation set, still offers the best compromise between accuracy and system complexity.

RQ3-A: Which gaze-based features are most predictive of perceived disagreement, and how consistent are these predictive features across participants?

Our additional experiments produced three feature subsets, $\mathcal{F}_A$, $\mathcal{F}_B$, and $\mathcal{F}_{Pool}$, as described in Section 5.2.1. To identify which gaze-based features are most predictive of disagreement, we combined participant-wise significance testing with an analysis of the internal correlations within these subsets (see Section 5.2.3). Across subsets, our findings suggest that fixation and saccade features, along with pupil-diameter variability, are the primary gaze-based signals for detecting perceived disagreement. In particular, fixation counts and durations, saccade and scanpath characteristics (e.g., saccade length, saccade count/duration, total scanpath duration), and the standard deviations of pupil diameter appear in at least two of the three subsets, indicating that they are recurrently selected as informative features.

However, the contribution of individual features is not consistent across participants. Fully personalised models achieve higher performance (average balanced accuracy = 68.40%) compared to

the best generalised models (balanced accuracy ≈ 57.00%), yet the two-way repeated-measures ANOVA found no significant main effects between the feature subsets. At the same time, optimising the feature subset and time window on a per-participant basis significantly improved performance (Wilcoxon and permutation tests, $p < 0.0001$). Together, these results suggest that there is no single, universally optimal feature set; instead, systems should favour participant-specific feature selection or dimensionality reduction strategies rather than relying on a fixed, global feature subset.

The high correlations among many gaze features likely explain why the automated feature selection methods used in our initial experiments underperformed relative to the manually derived subsets. Standard automated procedures tend to remove highly correlated features, whereas our results show that, despite redundancy, combining them remains informative. Moreover, the absence of pupil-based features in $\mathcal{F}_{Pool}$ suggests that for some participants, oculomotor features alone are sufficient to capture disagreement-related behaviour, while for others, pupil-based measures provide additional discriminative value. This further reinforces the need for adaptive, participant-specific modelling rather than enforcing a single feature configuration across all users.

RQ3-B: Which time window prior to a participant's decision best discriminates perceived disagreement using passive gaze signals, and how does the duration of this window impact classification performance?

We examined three time-window selection strategies as described in Section 5.2.2: the full recording, the last 11 seconds, and the last 3 seconds of each trial. As reported in Section 5.2.3, the optimal window length varied across individuals, indicating that the most informative temporal segment for disagreement detection is participant-dependent rather than globally fixed. Our motivation for varying the time window was grounded in the temporal nature of cognitive–affective processes. Gaze behaviour unfolds in stages, with early fixations reflecting an orienting phase and later fixations associated with deeper analysis and evaluation [98, 109]. Neurocognitive models similarly propose distinct temporal stages of decision-making, from preference assessment to action selection and outcome evaluation, each with characteristic neural dynamics [26, 50, 75]. Our results align with this view: different users concentrate discriminative information in distinct temporal segments, suggesting that disagreement-related gaze behaviour can either accumulate over the entire trial or appear in brief, late-stage bursts.

When aggregating performance by each participant's selected window, models trained on **full recordings** provided the most robust overall performance (balanced accuracy = 70.20%, AUC = 70.30%). This suggests that, at the group level, including the entire trial tends to capture a broad mixture of orienting, exploratory, and evaluative gaze behaviour, providing a stable and reliable basis for distinguishing agreement from disagreement. However, for some participants, truncated windows improved performance relative to full recordings, suggesting that earlier parts of the trial may be dominated by broader orienting or exploratory viewing, whereas gaze in later segments more directly reflects the evaluative and decision-related processes [26, 75, 98, 109]. The **11-second window** generally produced intermediate performance (balanced accuracy = 67.50%, AUC = 70.10%), but with lower recall (64.80%). Its relatively high AUC indicates that this window preserves useful ranking information, even as sensitivity decreases, implying that it retains much of the discriminative structure of the full recording while filtering out some early, potentially less informative gaze behaviour. However, the modest improvement suggests that the signal still includes some noise that affects performance, especially for detecting the positive class (i.e., disagreement). In contrast, the **3-second window** slightly reduced overall discrimination (balanced accuracy = 66.90%, AUC = 67.20%), but delivered the highest recall (72.50%) and F1-score (70.10%), particularly enhancing sensitivity to disagreement trials. This pattern is consistent with the idea that, for many participants, decisive disagreement-related gaze patterns occur in brief, concentrated episodes close to the moment of response, mirroring late-stage decision dynamics reported in both eye-tracking and neurocognitive studies [26, 75].

Statistically, the two-way repeated-measures ANOVA did not reveal a significant main effect of window length on balanced accuracy. At the same time, per-participant optimisation of the time window (in combination with a feature subset) yielded significant gains in accuracy (Wilcoxon and permutation tests, $p < 0.0001$). These results indicate that performance improvements stem from the interaction between time windows and feature subsets, rather than from window length alone. This is in line with work showing that individual eye-movement profiles evolve idiosyncratically over the course of a trial [66, 74], and with evidence that emotional and cognitive states have distinct temporal trajectories [50].

From a modelling perspective, these results have two key implications. First, if a single fixed policy is required, using full recordings is the safest choice, as it offers the most balanced and reliable performance. Second, when lightweight per-user adaptation is possible, selecting the time window (and feature subset) based on a small validation set can substantially improve performance with limited additional complexity. For example, a short calibration step can be added for new users, similar to eye-tracking calibration, where we collect a small number of labelled trials, evaluate candidate time windows and feature subsets on a held-out split, and use a lightweight classifier (e.g., LDA) to select the configuration that maximises balanced accuracy. This provides a per-user default with minimal overhead and can be retrained periodically.

7 Limitations & Future Work

Detecting implicit disagreement from gaze data is inherently challenging due to its subtle and individualised nature. Our findings confirm that personalisation is essential, as no single feature–temporal configuration generalised well across participants. Given the relatively small size of our dataset, these results should be interpreted as exploratory and require validation on additional datasets before stronger conclusions can be drawn. Importantly, we did not implement or evaluate a full interactive human–AI system that responds to detected disagreement in real time. Our contribution is best viewed as an empirical feasibility study that motivates and informs future end-to-end interactive prototypes. Below, we outline several limitations we identified and propose directions for future research.

7.1 Stimuli

Our findings suggest that a key limitation of this study was our choice to use the FOIL-COCO dataset. The foil captions differed by a single word, typically drawn from the same semantic category, as the correct counterpart [88]. While this design ensures tightly controlled, grammatically plausible errors, it likely made the agreement and disagreement signals in our data overly homogeneous, reducing the variability needed for robust modelling. Moreover, the dataset does not systematically vary foil difficulty or semantic plausibility; most substitutions are subtle (e.g., "dog" → "cow" or "banana" → "orange") without being highly atypical, which limits their potential to induce strong surprise or disagreement. As a result, the perceived difference between correct and foiled captions was likely too small to consistently elicit full disagreement. This subtlety contributes to the near-chance performance observed in generalised models and explains why improvements in personalised models remained limited. This interpretation aligns with our observations in Section 6 (RQ1), where facial reactions were largely neutral, and differences between agree and disagree trials were often subtle; this suggests that the stimuli primarily induced low-intensity appraisal (e.g., mild uncertainty or confusion) rather than reliably evoking strong, overt disagreement.

More broadly, this reflects a mismatch between how we measured disagreement and how it is experienced. A binary agree/disagree response can hide differences in disagreement intensity, ranging from a barely noticeable mismatch to a clear error. Because many foils change only one semantically similar word, they might have produced weaker disagreement signals, which are harder to distinguish using passive measures. Despite these challenges, we release our full dataset to enable future work to re-examine these signals. Researchers can leverage this data to explore alternative feature extraction methods or relaxed processing choices that might uncover patterns our current pipeline missed.

Future work should incorporate datasets with controlled variation in foil similarity and contextual plausibility, enabling more detailed analysis of disagreement strength. Additionally, integrating computational measures such as semantic distance between correct and foil words and image–caption alignment could help predict and calibrate disagreement difficulty, thereby improving both experimental design and model generalisability. To better align the labels with participants' subjective experience, future studies could collect graded responses (e.g., Likert-scale disagreement intensity and/or confidence) rather than relying solely on a binary judgement, to explore meaningful variation. Finally, disagreement manipulations could be broadened beyond single-word substitutions to include qualitatively different error types (e.g., object, attribute, relation, or context violations), allowing a more systematic test of when passive signals shift from subtle mismatch detection to clearer, stronger disagreement.

7.2 Modalities

As highlighted in Section 2.1, disagreement is often regarded as an inherently multimodal signal, involving a coordinated combination of facial, gaze, gestures, and other behavioural cues [13, 46]. In our study, the facial stream contributed little predictive signal after processing. We attribute this primarily to the interaction between our experimental design, which elicited subtle reactions, and our computational pipeline, which may not have been sensitive enough to capture these faint cues. In particular, facial expressions associated with low-intensity appraisals can be subtle, highly individual, and sensitive to recording conditions and feature-extraction choices [23, 86]. Under our task conditions, many participants remained facially neutral, suggesting that the AU-based summaries we extracted were too coarse to capture the relevant temporal dynamics and cross-signal dependencies emphasised in multimodal accounts of dispreferred response [46, 71]. Therefore, these results likely reflect the methodological challenge of computationally capturing implicit disagreement signals, rather than a conceptual challenge to the multimodal nature of disagreement.

Future work should continue to explore the multimodal nature of disagreement, potentially by incorporating richer facial, vocal, and other behavioural signals to capture complementary cues more effectively. Overall, multimodal modelling of disagreement remains a promising direction for future research in computational human–computer interaction.

7.3 Methods & Generalisability

Our results indicate that no fixed feature set generalises across participants, and high feature correlations likely limited the effectiveness of traditional feature-selection methods. Future research should investigate automatic feature extraction and representation learning using state-of-the-art deep neural networks, which may capture more discriminative patterns. However, this requires larger, more diverse datasets to avoid overfitting and support the development of generalisable models, which reflects a recurring challenge in learning-based gaze research [108].

In preliminary experiments, we evaluated deep learning architectures and transfer learning with pre-trained models, but given the current dataset constraints, these approaches tended to overfit and failed to generalise reliably, reinforcing the need for greater data scale and diversity. We also explored simple multimodal fusion strategies, including early and late fusion of gaze and facial features, but observed only limited benefits. Future work should therefore examine more recent intermediate or hybrid fusion techniques that can model cross-modal interactions more effectively.

We did not compare against prior baseline disagreement-detection models because, to the best of our knowledge, there is no established, directly comparable benchmark for implicit (dis)agreement detection from gaze and facial signals in a similar setup. Related studies using gaze/facial cues typically target different constructs (e.g., confusion, workload, affect) or different interaction contexts. Therefore, our work serves as a first attempt to establish such a benchmark for implicit disagreement detection, providing a baseline against which future models and feature-extraction techniques can be compared.

Beyond data scale and model choice, our study was conducted under controlled laboratory conditions with a relatively small sample, which is a known limitation in the eye-tracking community. This may limit generalisability to more diverse users and real-world interaction settings. Future work should expand on this by examining disagreement detection in more naturalistic contexts and across broader populations to assess robustness under realistic variability.

Incorporating adaptive or online personalisation strategies could further help bridge the gap between lab-based performance and practical deployment, especially when user-specific patterns are expected to evolve over time.

7.4 Feature Subset Differences

In our post-hoc feature analysis, we observed a difference in the number of statistically significant gaze features between the two counterbalanced groups, with $|\mathcal{F}_A| = 10$ and $|\mathcal{F}_B| = 20$. Although age and gender were balanced and stimuli were counterbalanced, this difference likely came from how these subsets were constructed: $\mathcal{F}_A$ and $\mathcal{F}_B$ represent the union of significant features for any single participant in that group, and since only three participants in Group A and four in Group B yielded significant features, small sampling variations and strong inter-feature correlations could easily amplify subset differences even under counterbalancing. In addition, upon further examination, Group A had a higher proportion of participants with prior eye-tracking experience ($n = 12$) compared to Group B ($n = 7$); this familiarity may have led to more stable, task-oriented viewing strategies in Group A, thereby reducing the variance required for certain features to reach significance compared to Group B. Finally, although the base images were the same, the specific image–caption pairs and foil words differed between groups, and subtle variations in foil difficulty or semantic distance may have interacted with disagreement responses. Therefore, future work should use larger, more balanced samples to further assess the robustness of these feature-level patterns.

7.5 Ethical & Privacy Concerns

Passive monitoring raises important ethical considerations, including informed consent, data privacy, and transparency regarding the signals collected and their use. The collection of continuous behavioural or physiological data (e.g., gaze and facial video) can elicit discomfort due to feelings of constant monitoring and impose a burden on participants who must carry and safeguard sensing devices, particularly in medical or high-stakes contexts [61]. These concerns have also been highlighted specifically for passive gaze-based interaction, where gaze traces can support inferences beyond the immediate interaction context; the increasing feasibility of webcam-based gaze sensing further raises the stakes by lowering the barrier to large-scale deployment outside controlled research settings [67]. In this context, our results suggest that, at least with our current stimuli and pipeline, the practical benefits of continuous gaze/facial monitoring are limited and are unlikely to justify deployment in real-world systems where the cost to privacy and user comfort is high. Rather than treating these concerns and negative results as a reason to abandon this research direction altogether, we interpret them as strengthening the case for *responsible constraints*. Future work should adopt a proportionality principle in which data collection is minimised, and the expected benefit clearly outweighs the privacy burden. Concretely, this implies privacy-preserving designs (e.g., opt-in consent, purpose limitation, and local/on-device processing of derived features instead of storing raw video), transparency about what is sensed and why, and user control mechanisms such as easy opt-out and episodic, rather than continuous, sensing.

8 Conclusion

In this study, we investigated the feasibility of detecting perceived disagreement from passive gaze signals during a simulated AI interaction task. We conducted a 30-participant user study in which eye-tracking and facial data were collected while participants judged whether image–caption pairs were correct. Our initial experiments using both modalities with automatic feature selection and generalised processing achieved near-chance-level performance, with models that combined gaze and facial features performing worse than gaze-only models. This prompted us to conduct further experiments on the unimodal gaze data using three time windows (i.e., full recording, last 11 seconds, last 3 seconds), and three different feature subsets based on statistically significant differences between **agree** and **disagree** trials. Our results show that disagreement is highly personalised; personalised models achieved an average balanced accuracy of 68.40%, compared to 57.00% for generalised models. Further post-hoc statistical analysis revealed no single feature subset or time window consistently outperformed others, underscoring the need for adaptive strategies. While gaze-based signals encode useful information, their variability suggests that multimodal fusion and advanced feature learning are essential for robust performance. Additionally, we presented directions for future work, such as using richer stimuli with graded disagreement difficulty to improve our performance, and highlighted ethical and privacy considerations that should guide any future development and evaluation of passive disagreement detection. Overall, this work represents an initial step towards understanding and modelling implicit disagreement to enhance adaptive human–AI systems.

Acknowledgments

This work was funded, in part, by the European Union under grant number 101093079 (MASTER); the German Federal Ministry of Research, Technology and Space (BMFTR) under grant number 01IW23002 (No-IDLE) and grant number 16IW24006 (NoIDLEChatGPT); the Lower Saxony Ministry of Science and Culture (MWK) through the zukunft.niedersachsen program; and the Endowed Chair of AAI at the University of Oldenburg. We also thank **Ádám Fodor** for his assistance with facial data processing.

References

[1] Ahmed Abdou, Ekta Sood, Philipp Müller, and Andreas Bulling. 2022. Gaze-enhanced Crossmodal Embeddings for Emotion Recognition. *Proc. ACM Hum.-Comput. Interact.* 6, ETRA (May 2022), 138:1–138:18. doi:10.1145/3530879

[2] Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *J. Hum.-Robot Interact.* 6, 1 (May 2017), 25–63. doi:10.5898/JHRI.6.1.Admoni

[3] Rawan Alghofaili, Yasuhito Sawahata, Haikun Huang, Hsueh-Cheng Wang, Takaaki Shiratori, and Lap-Fai Yu. 2019. Lost in Style: Gaze-Driven Adaptive Aid for VR Navigation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300578

[4] 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. doi:10.1609/aimag.v35i4.2513

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300233

[6] Richard Andersson, Marcus Nyström, Kenneth Holmqvist, Richard Andersson, Marcus Nyström, and Kenneth Holmqvist. 2010. Sampling Frequency and

Eye-Tracking Measures: How Speed Affects Durations, Latencies, and More. *Journal of Eye Movement Research* 3, 3 (Sept. 2010), 1–12. doi:10.16910/jemr.3.3.6 Company: Bern Open Publishing Distributor: Bern Open Publishing Institution: Bern Open Publishing Label: Bern Open Publishing Publisher: publisher.

[7] Jo Angouri and Miriam A. Locher. 2012. Theorising disagreement. *Journal of Pragmatics* 44, 12 (Sept. 2012), 1549–1553. doi:10.1016/j.pragma.2012.06.011

[8] Michael Barz, Omair Shahzad Bhatti, and Daniel Sonntag. 2022. Implicit Estimation of Paragraph Relevance From Eye Movements. *Frontiers in Computer Science* 3 (Jan. 2022). doi:10.3389/fcomp.2021.808507 Publisher: Frontiers.

[9] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (Jan. 1995), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

[10] 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 (ETRA '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3379157.3391302

[11] Nilavra Bhattacharya, Somnath Rakshit, Jacek Gwizdka, and Paul Kogut. 2020. Relevance Prediction from Eye-Movements Using Semi-Interpretable Convolutional Neural Networks. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval (CHIIR '20)*. Association for Computing Machinery, New York, NY, USA, 223–233. doi:10.1145/3343413.3377960 event-place: Vancouver BC, Canada.

[12] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl. 2017. Visualization of Eye Tracking Data: A Taxonomy and Survey: Visualization of Eye Tracking Data. *Computer Graphics Forum* 36, 8 (Dec. 2017), 260–284. doi:10.1111/cgf.13079

[13] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. 2013. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing* 31, 2 (Feb. 2013), 203–221. doi:10.1016/j.imavis.2012.07.003

[14] Babette Bühler, Efe Bozkir, Hannah Deininger, Patricia Goldberg, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2024. Detecting Aware and Unaware Mind Wandering During Lecture Viewing: A Multimodal Machine Learning Approach Using Eye Tracking, Facial Videos and Physiological Data. In *Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 244–253. doi:10.1145/3678957.3685710

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

[16] Cambridge University Press & Assessment. 2025. *Disagreement*. Cambridge Dictionary. https://dictionary.cambridge.org/dictionary/english/disagreement Accessed: 2025-11-12.

[17] Benjamin T. Carter and Steven G. Luke. 2020. Best practices in eye tracking research. *International Journal of Psychophysiology* 155 (Sept. 2020), 49–62. doi:10.1016/j.ijpsycho.2020.05.010

[18] José J. Cañas. 2022. AI and Ethics When Human Beings Collaborate With AI Agents. *Frontiers in Psychology* 13 (March 2022). doi:10.3389/fpsyg.2022.836650 Publisher: Frontiers.

[19] Laurence Chaby, Isabelle Hupont, Marie Avril, Viviane Luherne-du Boullay, and Mohamed Chetouani. 2017. Gaze Behavior Consistency among Older and Younger Adults When Looking at Emotional Faces. *Frontiers in Psychology* 8 (2017). doi:10.3389/fpsyg.2017.00548

[20] Jin Hyun Cheong, Eshin Jolly, Tiankang Xie, Sophie Byrne, Matthew Kenney, and Luke J. Chang. 2023. Py-Feat: Python Facial Expression Analysis Toolbox. *Affective Science* 4, 4 (Dec. 2023), 781–796. doi:10.1007/s42761-023-00191-4

[21] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4302–4310.

[22] Yucel Cimtay, Erhan Ekmekcioglu, and Seyma Caglar-Ozhan. 2020. Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion. *IEEE Access* 8 (2020), 168865–168878. doi:10.1109/ACCESS.2020.3023871

[23] Elizabeth A. Clark, J'Nai Kessinger, Susan E. Duncan, Martha Ann Bell, Jacob Lahne, Daniel L. Gallagher, and Sean F. O'Keefe. 2020. The Facial Action Coding System for Characterization of Human Affective Response to Consumer Product-Based Stimuli: A Systematic Review. *Frontiers in Psychology* 11 (May 2020). doi:10.3389/fpsyg.2020.00920 Publisher: Frontiers.

[24] Eve V. Clark. 2015. Common Ground. In *The Handbook of Language Emergence*. John Wiley & Sons, Ltd, 328–353. doi:10.1002/9781118346136.ch15 Section: 15 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118346136.ch15.

[25] Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge. doi:10.1017/CBO9780511620539

[26] Pedro Manuel Cortes, Juan Pablo García-Hernández, Fabiola Alejandra Iribe-Burgos, Marisela Hernández-González, Carolina Sotelo-Tapia, and Miguel Angel Guevara. 2021. Temporal division of the decision-making process: An EEG study. *Brain Research* 1769 (Oct. 2021), 147592. doi:10.1016/j.brainres.2021.147592

[27] Yuchen Cui, Qiping Zhang, Brad Knox, Alessandro Allievi, Peter Stone, and Scott Niekum. 2021. The EMPATHIC Framework for Task Learning from Implicit Human Feedback. In *Proceedings of the 2020 Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 155)*, Jens Kober, Fabio Ramos, and Claire Tomlin (Eds.). PMLR, 604–626. doi:10.48550/arXiv.2009.13649

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. doi:10.1109/CVPR.2009.5206848 ISSN: 1063-6919.

[29] Nathaniel Dennler, Stefanos Nikolaidis, and Maja Matarić. 2025. Contrastive Learning from Exploratory Actions: Leveraging Natural Interactions for Preference Elicitation. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI '25)*. IEEE Press, Melbourne, Australia, 778–788.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. doi:10.48550/arXiv.2010.11929

[31] Anca D. Dragan, Kenton C.T. Lee, and Siddhartha S. Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 301–308. doi:10.1109/HRI.2013.6483603 ISSN: 2167-2148.

[32] Paul Ekman. 1999. Basic Emotions. In *Handbook of Cognition and Emotion*. John Wiley & Sons, Ltd, 45–60. doi:10.1002/0470013494.ch3 Section: 3 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470013494.ch3.

[33] 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, Klaus Scherer, Masatoshi Tomita, and Athanase Tzavaras. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology* 53, 4 (1987), 712–717. doi:10.1037/0022-3514.53.4.712 Place: US Publisher: American Psychological Association.

[34] Andrew Emerson, Wookhee Min, Jonathan Rowe, Roger Azevedo, and James Lester. 2023. Multimodal Predictive Student Modeling with Multi-Task Transfer Learning. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK2023)*. Association for Computing Machinery, New York, NY, USA, 333–344. doi:10.1145/3576050.3576101

[35] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces* (Miami, Florida, USA) *(IUI '03)*. Association for Computing Machinery, New York, NY, USA, 39–45. doi:10.1145/604045.604056

[36] Frank O. Flemisch, Marie-Pierre Pacaux-Lemoine, Frederic Vanderhaegen, Makoto Itoh, Yuichi Saito, Nicolas Herzberger, Joscha Wasser, Emmanuelle Grislin, and Marcel Baltzer. 2020. Conflicts in Human-Machine Systems as an Intersection of Bio- and Technosphere: Cooperation and Interaction Patterns for Human and Machine Interference and Conflict Resolution. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. Institute of Electrical and Electronics Engineers, 1–6. doi:10.1109/ICHMS49158.2020.9209517

[37] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature* 637, 8045 (Jan. 2025), 319–326. doi:10.1038/s41586-024-08328-6 Publisher: Nature Publishing Group.

[38] Kenneth Holmqvist, Marcus Nystrom, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, Weijer, and Joost van de. 2011. *Eye Tracking: A comprehensive guide to methods and measures*. Oxford University Press, Oxford, New York.

[39] 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 (Oct. 2020), 63–72. doi:10.1609/hcomp.v8i1.7464

[40] Shamsi T. Iqbal and Eric Horvitz. 2010. Notifications and awareness: a field study of alert usage and preferences. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work (CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 27–30. doi:10.1145/1718918.1718926

[41] Tingting Jiang, Zhumo Sun, Shiting Fu, and Yan Lv. 2024. Human-AI interaction research agenda: A user-centered perspective. *Data and Information Management* 8, 4 (Dec. 2024), 100078. doi:10.1016/j.dim.2024.100078

[42] Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. 2023. Multimodal Adaptive Emotion Transformer with Flexible Modality Inputs on A Novel Dataset with Continuous Labels. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. Association for Computing Machinery, New York, NY, USA, 5975–5984. doi:10.1145/3581783.3613797

[43] Christopher Kanan, Dina N.F. Bseiso, Nicholas A. Ray, Janet H. Hsiao, and Garrison W. Cottrell. 2015. Humans have idiosyncratic and task-specific scanpaths

for judging faces. *Vision Research* 108 (March 2015), 67–76. doi:10.1016/j.visres.2015.01.013

[44] Adam Kendon. 2002. Some uses of the head shake. *Gesture* 2, 2 (Jan. 2002), 147–182. doi:10.1075/gest.2.2.03ken Publisher: John Benjamins.

[45] Kobin H. Kendrick and Judith Holler. 2017. Gaze Direction Signals Response Preference in Conversation. *Research on Language and Social Interaction* 50, 1 (Jan. 2017), 12–32. doi:10.1080/08351813.2017.1262120 Publisher: Routledge _eprint: https://doi.org/10.1080/08351813.2017.1262120.

[46] Kobin H. Kendrick, Judith Holler, and Stephen C. Levinson. 2023. Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences* 378, 1875 (March 2023), 20210473. doi:10.1098/rstb.2021.0473 Publisher: Royal Society.

[47] Suzie Kim, Hye-Bin Shin, and Seong-Whan Lee. 2025. Aligning Humans and Robots via Reinforcement Learning from Implicit Human Feedback. doi:10.48550/arXiv.2507.13171 arXiv:2507.13171 [cs].

[48] Se Young Kim, Hahyeon Park, Hongbum Kim, Joon Kim, and Kyoungwon Seo. 2022. Technostress causes cognitive overload in high-stress people: Eye tracking analysis in a virtual kiosk test. *Information Processing & Management* 59, 6 (Nov. 2022), 103093. doi:10.1016/j.ipm.2022.103093

[49] W. Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: the TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture (K-CAP '09)*. Association for Computing Machinery, New York, NY, USA, 9–16. doi:10.1145/1597735.1597738

[50] Philip A. Kragel, Ahmad R. Hariri, and Kevin S. LaBar. 2022. The Temporal Dynamics of Spontaneous Emotional Brain States and Their Implications for Mental Health. *Journal of Cognitive Neuroscience* 34, 5 (March 2022), 715–728. doi:10.1162/jocn_a_01787

[51] 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*, Jose M. Alonso and Alejandro Catala (Eds.). Association for Computational Linguistics, Dublin, Ireland, 55–60. https://aclanthology.org/2020.nl4xai-1.12/

[52] Maximilian Krug. 2025. Chapter 6. Gaze aversion as a marker of disalignment in interactions. In *Mobile Eye Tracking: New avenues for the study of gaze in social interaction*, Elisabeth Zima and Anja Stukenbrock (Eds.). John Benjamins Publishing Company, 165–187. doi:10.1075/pbns.351.06kru

[53] Emmanouil Ktistakis, Vasileios Skaramagkas, Dimitris Manousos, Nikolaos S. Tachos, Evanthia Tripoliti, Dimitrios I. Fotiadis, and Manolis Tsiknakis. 2022. COLET: A dataset for COgnitive workLoad estimation based on eye-tracking. *Computer Methods and Programs in Biomedicine* 224 (Sept. 2022), 106989. doi:10.1016/j.cmpb.2022.106989

[54] 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 (IJCAI'16)*. AAAI Press, New York, New York, USA, 2529–2535.

[55] Antonio Lanatà, Gaetano Valenza, and Enzo Pasquale Scilingo. 2013. Eye gaze patterns in emotional pictures. *Journal of Ambient Intelligence and Humanized Computing* 4, 6 (Dec. 2013), 705–715. doi:10.1007/s12652-012-0147-6

[56] Alexander Leube, Katharina Rifai, Siegfried Wahl, Alexander Leube, Katharina Rifai, and Siegfried Wahl. 2017. Sampling Rate Influences Saccade Detection in Mobile Eye tracking of a Reading Task. *Journal of Eye Movement Research* 10, 3 (June 2017), 1–11. doi:10.16910/jemr.10.3.3 Company: Bern Open Publishing Distributor: Bern Open Publishing Institution: Bern Open Publishing Label: Bern Open Publishing Publisher: publisher.

[57] Joe Li and Peter Washington. 2024. A Comparison of Personalized and Generalized Approaches to Emotion Recognition Using Consumer Wearable Devices: Machine Learning Study. *JMIR AI* 3, 1 (May 2024), e52171. doi:10.2196/52171 Company: JMIR AI Distributor: JMIR AI Institution: JMIR AI Label: JMIR AI Publisher: JMIR Publications Inc., Toronto, Canada.

[58] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755. doi:10.1007/978-3-319-10602-1_48

[59] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. 2022. Learning Multi-dimensional Edge Feature-based AU Relation Graph for Facial Action Unit Recognition, Vol. 2. 1239–1246. doi:10.24963/ijcai.2022/173 ISSN: 1045-0823.

[60] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. 2017. Interactive learning from policy-dependent human feedback. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17)*. JMLR.org, Sydney, NSW, Australia, 2285–2294.

[61] Nicole A. Maher, Joeky T. Senders, Alexander F. C. Hulsbergen, Nayan Lamba, Michael Parker, Jukka-Pekka Onnela, Annelien L. Bredenoord, Timothy R. Smith, and Marike L. D. Broekman. 2019. Passive data collection and use in healthcare: A systematic review of ethical issues. *International Journal of Medical Informatics* 129 (Sept. 2019), 242–247. doi:10.1016/j.ijmedinf.2019.06.015

[62] Päivi Majaranta and Andreas Bulling. 2014. *Eye Tracking and Eye-Based Human–Computer Interaction*. Springer, London, 39–65. doi:10.1007/978-1-4471-6392-3_3

[63] Andreas Mallas, Michalis Xenos, and Christos Katsanos. 2022. A Descriptive Model of Passive and Natural Passive Human-Computer Interaction. In *Human-Computer Interaction. Theoretical Approaches and Design Methods*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 104–116. doi:10.1007/978-3-031-05311-5_7

[64] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. 2025. Artificial Intelligence Index Report 2025. doi:10.48550/arXiv.2504.07139 arXiv:2504.07139 [cs].

[65] Jack McGuire, David De Cremer, and Tim Van de Cruys. 2024. Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence. *Scientific Reports* 14, 1 (Aug. 2024), 18525. doi:10.1038/s41598-024-69423-2 Publisher: Nature Publishing Group.

[66] Eyal Mehoudar, Joseph Arizpe, Chris I. Baker, and Galit Yovel. 2014. Faces in the eye of the beholder: Unique and stable eye scanning patterns of individual observers. *Journal of Vision* 14, 7 (June 2014), 6–6. doi:10.1167/14.7.6 Publisher: The Association for Research in Vision and Ophthalmology.

[67] 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 (June 2024). doi:10.3389/frai.2024.1391745 Publisher: Frontiers.

[68] Abdulrahman Mohamed Selim, Omair Shahzad Bhatti, Michael Barz, and Daniel Sonntag. 2024. Perceived Text Relevance Estimation Using Scanpaths and GNNs. In *Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 418–427. doi:10.1145/3678957.3685736

[69] Douglas Oard and Jinmook Kim. 1998. Implicit Feedback for Recommender System. *Proceedings of the AAAI Workshop on Recommender Systems* 83 (1998), 81–83.

[70] Oxford University Press. 2025. *Disagreement*. Oxford Learner's Dictionaries. https://www.oxfordlearnersdictionaries.com/definition/english/disagreement Accessed: 2025-11-12.

[71] Simona Pekarek Doehler, Hilla Polak-Yitzhaki, Xiaoting Li, Ioana Maria Stoenica, Martin Havlík, and Leelo Keevallik. 2021. Multimodal Assemblies for Prefacing a Dispreferred Response: A Cross-Linguistic Analysis. *Frontiers in Psychology* 12 (Sept. 2021). doi:10.3389/fpsyg.2021.689275 Publisher: Frontiers.

[72] Isabella Poggi, Francesca D'Errico, and Laura Vincze. 2011. Agreement and its Multimodal Communication in Debates: A Qualitative Analysis. *Cognitive Computation* 3, 3 (Sept. 2011), 466–479. doi:10.1007/s12559-010-9068-x

[73] 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? doi:10.48550/arXiv.2202.09634 arXiv:2202.09634.

[74] William Poynter, Megan Barber, Jason Inman, and Coral Wiggins. 2013. Individuals exhibit idiosyncratic eye-movement behavior profiles across tasks. *Vision Research* 89 (Aug. 2013), 32–38. doi:10.1016/j.visres.2013.07.002

[75] Philip Pärnamets, Petter Johansson, Lars Hall, Christian Balkenius, Michael J. Spivey, and Daniel C. Richardson. 2015. Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences* 112, 13 (March 2015), 4170–4175. doi:10.1073/pnas.1415250112 Publisher: Proceedings of the National Academy of Sciences.

[76] Philip Pärnamets, Roger Johansson, Kerstin Gidlöf, and Annika Wallin. 2016. How Information Availability Interacts with Visual Attention during Judgment and Decision Tasks. *Journal of Behavioral Decision Making* 29, 2-3 (2016), 218–231. doi:10.1002/bdm.1902 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.1902.

[77] George E. Raptis, Christina Katsini, Marios Belk, Christos Fidas, George Samaras, and Nikolaos Avouris. 2017. Using Eye Gaze Data and Visual Activities to Infer Human Cognitive Styles: Method and Feasibility Studies. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. Association for Computing Machinery, New York, NY, USA, 164–173. doi:10.1145/3079628.3079690

[78] Jeba Rezwana and Mary Lou Maher. 2023. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Transactions on Computer-Human Interaction* 30, 5 (Oct. 2023), 1–28. doi:10.1145/3519026

[79] Federico Rossano. 2012. Gaze in Conversation. In *The Handbook of Conversation Analysis*. John Wiley & Sons, Ltd, 308–329. doi:10.1002/9781118325001.ch15 Section: 15 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118325001.ch15.

[80] Anup Kumar Roy, Md. Nadeem Akhtar, Manjunatha Mahadevappa, Rajlakshmi Guha, and Jayanta Mukherjee. 2020. A Novel Technique to Develop Cognitive

Models for Ambiguous Image Identification Using Eye Tracker. *IEEE Transactions on Affective Computing* 11, 1 (2020), 63–77. doi:10.1109/TAFFC.2017.2768026

[81] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (Dec. 2015), 211–252. doi:10.1007/s11263-015-0816-y

[82] 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. doi:10.1145/3176349.3176391

[83] 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. doi:10.1145/3361570.3361577

[84] 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 (ETRA '00)*. Association for Computing Machinery, New York, NY, USA, 71–78. doi:10.1145/355017.355028

[85] Klaus R. Scherer. 1982. Emotion as a process: Function, origin and regulation. *Social Science Information* 21, 4-5 (July 1982), 555–570. doi:10.1177/053901882021004004 Publisher: SAGE Publications Ltd.

[86] Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information* 44, 4 (Dec. 2005), 695–729. doi:10.1177/0539018405058216

[87] S. S. Shapiro and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples)†. *Biometrika* 52, 3-4 (Dec. 1965), 591–611. doi:10.1093/biomet/52.3-4.591 _eprint: https://academic.oup.com/biomet/article-pdf/52/3-4/591/962907/52-3-4-591.pdf.

[88] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie 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 (Volume 1: Long Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 255–265. doi:10.18653/v1/P17-1024

[89] Zhen-Feng Shi, Chang Zhou, Wei-Long Zheng, and Bao-Liang Lu. 2017. Attention evaluation with eye tracking glasses for EEG-based emotion recognition. In *2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*. 86–89. doi:10.1109/NER.2017.8008298

[90] 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 (ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 15–23. doi:10.1145/3382507.3418828

[91] Jim Smith, Phil Legg, Milos Matovic, and Kristofer Kinsey. 2018. Predicting User Confidence During Visual Decision Making. *ACM Trans. Interact. Intell. Syst.* 8, 2 (June 2018). doi:10.1145/3185524

[92] Tanya Stivers. 2008. Stance, Alignment, and Affiliation During Storytelling: When Nodding Is a Token of Affiliation. *Research on Language and Social Interaction* 41, 1 (March 2008), 31–57. doi:10.1080/08351810701691123 Publisher: Routledge _eprint: https://doi.org/10.1080/08351810701691123.

[93] Tanya Stivers and John Heritage. 2001. Breaking the sequential mold: Answering 'more than the question' during comprehensive history taking. *Text & Talk* 21, 1-2 (June 2001), 151–185. doi:10.1515/text.1.21.1-2.151 Publisher: De Gruyter Mouton.

[94] Tanya Stivers and Jeffrey D. Robinson. 2006. A Preference for Progressivity in Interaction. *Language in Society* 35, 3 (2006), 367–392. doi:10.1017/S0047404506060179 Publisher: Cambridge University Press.

[95] Theodore R. Sumers, Mark K. Ho, Robert D. Hawkins, Karthik Narasimhan, and Thomas L. Griffiths. 2021. Learning Rewards From Linguistic Feedback. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 7 (May 2021), 6002–6010. doi:10.1609/aaai.v35i7.16749

[96] Maurizio Talamo. 2023. The Digital Revolution and the Art of Co-creation. In *Technological Imagination in the Green and Digital Transition*, Eugenio Arbizzani, Eliana Cangelli, Carola Clemente, Fabrizio Cumo, Francesca Giofrè, Anna Maria Giovenale, Massimo Palme, and Spartaco Paris (Eds.). Springer International Publishing, Cham, 27–35. doi:10.1007/978-3-031-29515-7_4

[97] Stefano Teso and Kristian Kersting. 2019. Explanatory Interactive Machine Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 239–245. doi:10.1145/3306618.3314293

[98] Michael C. Trumbo, Mikaela L. Armenta, Michael J. Haass, Karin M. Butler, Aaron P. Jones, and Charles S. H. Robinson. 2016. Real Time Assessment of Cognitive State: Research and Implementation Challenges. In *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience*, Dylan D. Schmorrow and Cali M. Fidopiastis (Eds.). Springer International Publishing, Cham, 107–119. doi:10.1007/978-3-319-39952-2_12

[99] John W. Tukey. 1977. *Exploratory data analysis*. Addison-Wesley, Reading, Mass. [u.a.].

[100] Tim Verdonck, Bart Baesens, María Óskarsdóttir, and Seppe vanden Broucke. 2024. Special issue on feature engineering editorial. *Machine Learning* 113, 7 (July 2024), 3917–3928. doi:10.1007/s10994-021-06042-2

[101] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 2018. Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (April 2018). doi:10.1609/aaai.v32i1.11485

[102] Ryen W. White, Ian Ruthven, and Joemon M. Jose. 2002. The Use of Implicit Evidence for Relevance Feedback in Web Retrieval. In *Advances in Information Retrieval (Lecture Notes in Computer Science)*, Fabio Crestani, Mark Girolami, and Cornelis Joost van Rijsbergen (Eds.). Springer, Berlin, Heidelberg, 93–109. doi:10.1007/3-540-45886-7_7

[103] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. doi:10.2307/3001968 Publisher: [International Biometric Society, Wiley].

[104] Natnael A. Wondimu, Cédric Buche, and Ubbo Visser. 2022. Interactive Machine Learning: A State of the Art Review. doi:10.48550/arXiv.2207.06196 arXiv:2207.06196 [cs].

[105] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020. Visual Transformers: Token-based Image Representation and Processing for Computer Vision. doi:10.48550/arXiv.2006.03677 arXiv:2006.03677 [cs].

[106] Qun Wu, Nilanjan Dey, Fuqian Shi, Rubén González Crespo, and R. Simon Sherratt. 2021. Emotion classification on eye-tracking and electroencephalograph fused signals employing deep gradient neural networks. *Applied Soft Computing* 110 (Oct. 2021), 107752. doi:10.1016/j.asoc.2021.107752

[107] Duo Xu, Mohit Agarwal, Ekansh Gupta, Faramarz Fekri, and Raghupathy Sivakumar. 2021. Accelerating Reinforcement Learning using EEG-based implicit human feedback. *Neurocomputing* 460 (Oct. 2021), 139–153. doi:10.1016/j.neucom.2021.06.064

[108] Xucong Zhang, Seonwook Park, and Anna Maria Feit. 2021. Eye Gaze Estimation and Its Applications. In *Artificial Intelligence for Human Computer Interaction: A Modern Approach*, Yang Li and Otmar Hilliges (Eds.). Springer International Publishing, Cham, 99–130. doi:10.1007/978-3-030-82681-9_4

[109] Seren Zhu, Kaushik J Lakshminarasimhan, Nastaran Arfaei, and Dora E Angelaki. 2022. Eye movements reveal spatiotemporal dynamics of visually-informed planning in navigation. *eLife* 11 (May 2022), e73097. doi:10.7554/eLife.73097 Publisher: eLife Sciences Publications, Ltd.

A Supplementary Material

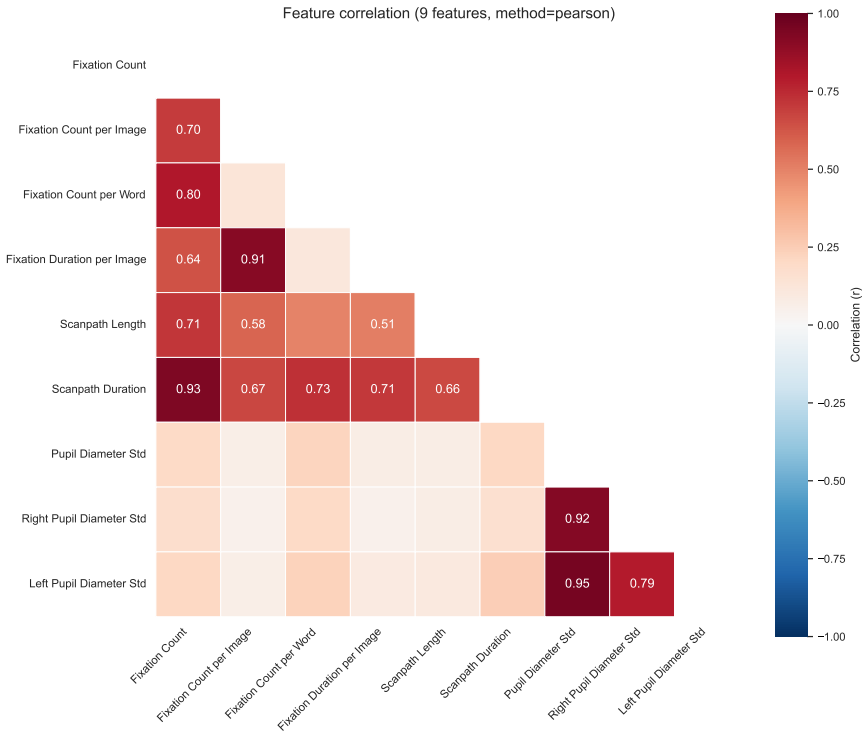
Algorithm A.1: Paired feature significance test logic.

Input: `Agree` and `Disagree` Data frames; feature set $\mathcal{F}$; boolean `same participant`; significance level α

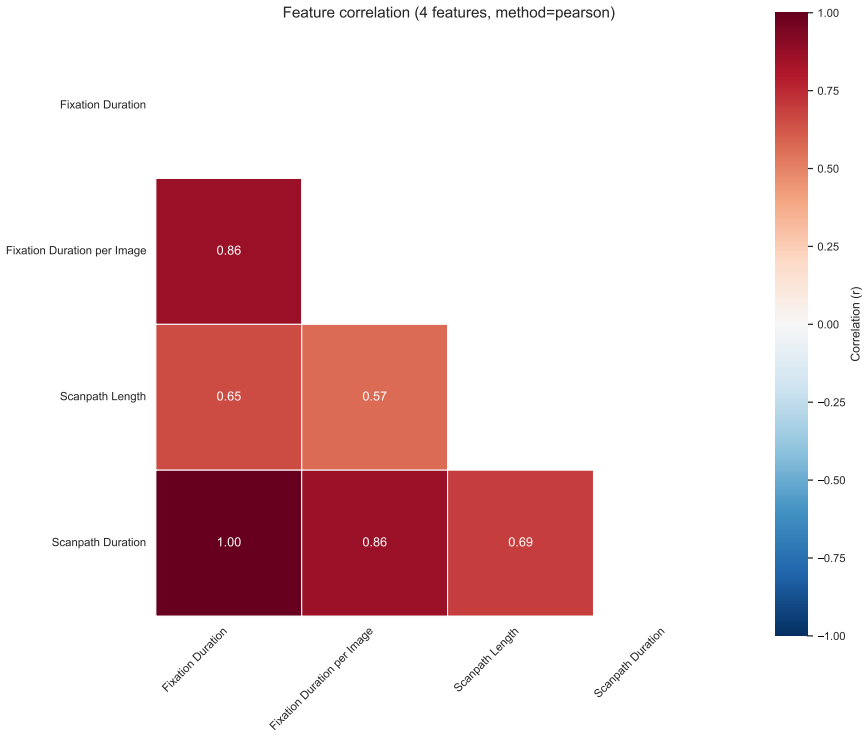
foreach *feature* $f \in \mathcal{F}$ **do**
  Build paired observations;
  if `same participant` = ***True*** **then**
    $x \leftarrow$ values of feature f in *Agree*; $y \leftarrow$ values of feature f in *Disagree*; `// aligned by participant id`
  else
    Aggregate within participant (e.g., participant mean across trials) in *Agree* and in *Disagree* to produce one value per participant: x, y;
  Remove pairs with missing values so x and y remain matched;
  $n_f \leftarrow$ number of remaining pairs for feature f;
  $\delta_f \leftarrow x - y$; `// within-pair differences`
  $\bar{\delta}_f \leftarrow \text{mean}(\delta_f)$; $s_{\delta_f} \leftarrow \text{sd}(\delta_f)$;
  Select test (Shapiro–Wilk decision);
  Perform Shapiro–Wilk on δ_f; obtain $p_{\text{SW},f}$;
  if $p_{\text{SW},f} \geq 0.05$ **then**
    Perform two-sided paired t-test on (x, y); store raw p as p_f;
    $\texttt{test}_f \leftarrow$ "Paired t-test";
  else
    Perform two-sided Wilcoxon signed-rank test on δ_f; store raw p as p_f;
    $\texttt{test}_f \leftarrow$ "Wilcoxon signed-rank";
  Effect size (paired Cohen's d);
  $d_f \leftarrow \bar{\delta}_f / s_{\delta_f}$; `// paired Cohen's d, using sample sd`
  Store record for feature f: $(n_f,\ \texttt{test}_f,\ p_f,\ d_f,\ \bar{\delta}_f,\ s_{\delta_f})$;
Benjamini-Hochberg adjustment (fixed method);
Let $m \leftarrow |\mathcal{F}|$ and collect raw p-values $\{p_f\}_{f \in \mathcal{F}}$;
Order p-values: $p_{(1)} \leq \cdots \leq p_{(m)}$ with corresponding features $(f_{(1)}, \ldots, f_{(m)})$;
Initialize $\tilde{p}_{(m+1)} := 1$;
for $i \leftarrow m$ ***down to*** 1 **do**
  $\tilde{p}_{(i)} \leftarrow \min\left(1,\ \frac{m}{i}\, p_{(i)},\ \tilde{p}_{(i+1)}\right)$;
Map each $\tilde{p}_{(i)}$ back to its feature $f_{(i)}$ and set $\texttt{adj\_p}_f \leftarrow \tilde{p}_f$;
Set $\texttt{significant}_f \leftarrow [\,\texttt{adj\_p}_f < \alpha\,]$;
return Table over $f \in \mathcal{F}$ with columns: `test`, `raw_p`, `adj_p`, d_f, `significant`, n_f, $\bar{\delta}_f$, s_{δ_f};

Table A.1: Per-user results, duration (Dur.) insights in seconds, and percentage of ground truth agreement (GTA).

User ID	Feature	Temporal Window	GTA	$Dur_{\cdot,Mean}$	$Dur_{\cdot,Min}$	$Dur_{\cdot,Max}$	$Dur_{\cdot,Median}$	Model	BA	AUC	Recall	Precision	F1
A01	$\mathcal{F}_B$	3 Seconds	0.857	6.836	2.188	19.616	6.152	MLP	0.659	0.660	0.696	0.718	0.694
A02	$\mathcal{F}_B$	11 Seconds	0.914	4.990	1.836	13.320	4.312	GB	0.593	0.615	0.500	0.614	0.538
A03	$\mathcal{F}_B$	3	0.913	7.036	2.660	24.784	5.932	DT	0.700	0.700	0.740	0.737	0.729
A04	$\mathcal{F}_{Pool}$	Full Recording	0.816	9.374	3.000	24.428	8.908	GB	0.691	0.784	0.643	0.592	0.602
A05	$\mathcal{F}_{Pool}$	3 Seconds	0.961	2.483	1.336	5.952	2.304	NB	0.648	0.634	0.818	0.639	0.714
A06	$\mathcal{F}_B$	11 Seconds	0.919	5.026	1.740	14.372	4.304	GP	0.670	0.730	0.648	0.670	0.654
A07	$\mathcal{F}_A$	3 Seconds	0.824	6.745	2.864	16.928	6.104	DT	0.611	0.611	0.575	0.700	0.625
A08	$\mathcal{F}_B$	3 Seconds	0.948	6.650	2.012	35.744	5.580	SVM	0.748	0.774	0.733	0.755	0.742
A09	$\mathcal{F}_A$	11 Seconds	0.856	4.771	1.292	19.904	4.216	MLP	0.666	0.638	0.686	0.697	0.670
A10	$\mathcal{F}_{Pool}$	3 Seconds	0.882	3.799	1.592	8.548	3.542	QDA	0.665	0.695	0.816	0.640	0.715
A11	$\mathcal{F}_{Pool}$	11 Seconds	0.974	4.362	1.580	12.836	3.784	RF	0.709	0.745	0.751	0.712	0.721
A12	$\mathcal{F}_A$	3 Seconds	0.895	4.574	1.452	16.108	4.152	SVM	0.718	0.693	0.933	0.733	0.819
A13	$\mathcal{F}_A$	Full Recording	0.967	3.271	1.112	9.492	2.792	GP	0.695	0.672	0.715	0.700	0.702
A14	$\mathcal{F}_A$	11 Seconds	0.857	4.233	1.476	8.372	3.858	MLP	0.728	0.751	0.707	0.709	0.698
A15	$\mathcal{F}_{Pool}$	Full Recording	0.948	5.037	2.052	11.132	4.632	SVM	0.706	0.707	0.760	0.671	0.708
A16	$\mathcal{F}_A$	3 Seconds	0.903	4.933	1.340	17.284	4.438	RF	0.717	0.717	0.733	0.750	0.726
B01	$\mathcal{F}_B$	Full Recording	0.455	2.715	0.944	12.544	2.148	DT	0.691	0.691	0.467	0.533	0.447
B03	$\mathcal{F}_A$	11 Seconds	0.867	5.755	1.248	18.460	5.564	NB	0.682	0.729	0.600	0.655	0.623
B04	$\mathcal{F}_{Pool}$	Full Recording	0.900	4.297	1.448	11.456	3.744	KNN	0.636	0.663	0.793	0.718	0.744
B05	$\mathcal{F}_{Pool}$	Full Recording	0.869	5.194	1.364	15.388	4.696	LGBM	0.719	0.722	0.808	0.741	0.765
B06	$\mathcal{F}_{Pool}$	Full Recording	0.864	7.965	2.512	34.324	6.900	GP	0.714	0.687	0.718	0.732	0.721
B07	$\mathcal{F}_B$	Full Recording	0.874	6.047	2.180	20.948	5.332	NB	0.645	0.602	0.544	0.746	0.620
B08	$\mathcal{F}_A$	3 Seconds	0.889	5.727	1.576	19.048	5.092	RF	0.656	0.680	0.524	0.744	0.571
B09	$\mathcal{F}_A$	Full Recording	0.948	6.571	1.840	20.480	6.330	RF	0.733	0.739	0.809	0.709	0.752
B10	$\mathcal{F}_B$	Full Recording	0.903	6.044	1.928	14.348	5.216	SVM	0.900	0.867	0.800	1.000	0.860
B11	$\mathcal{F}_{Pool}$	3 Seconds	0.926	4.297	1.588	10.212	3.752	SVM	0.616	0.622	0.656	0.648	0.646
B12	$\mathcal{F}_{Pool}$	Full Recording	0.935	2.318	1.044	6.408	2.188	MLP	0.686	0.716	0.789	0.741	0.759
B13	$\mathcal{F}_A$	Full Recording	0.877	4.486	1.576	13.616	4.104	QDA	0.632	0.625	0.593	0.678	0.621
B14	$\mathcal{F}_{Pool}$	3 Seconds	0.889	4.723	2.156	9.664	4.404	KNN	0.618	0.603	0.746	0.721	0.729
B15	$\mathcal{F}_B$	Full Recording	0.961	4.295	2.092	10.224	3.858	XGB	0.678	0.665	0.728	0.755	0.720



(a) Feature correlation for $\mathcal{F}_A$



(b) Feature correlation for $\mathcal{F}_{Pool}$

Figure A.1: The feature correlation matrices for $\mathcal{F}_A$ and $\mathcal{F}_{Pool}$.

B Facial Data Processing

In a survey paper by Bousmalis et al. [13], they defined disagreement as the belief that one holds an opinion opposite to that of an interlocutor, which they modified based on Poggi et al. [72] definition of agreement. This definition then led them to identify certain facial action units (AUs) that may be informative for detecting disagreement and agreement, as shown in Table B.1. They implied that these AUs are candidates rather than definitive markers, and robust modelling should emphasise their temporal dynamics and correlations. The descriptions of the AUs are guided by the work of Ekman [32, 33].

We re-encoded the facial data video recordings[10] to a uniform format and resampled them to 30 Hz to ensure temporal consistency. We ran basic quality checks, for example, to detect file corruption and to enforce a minimum frame length. We applied an automatic face detector to each frame and linked detections over time with a tracker to form continuous face tracks for each participant. From these tracks, we extracted facial action units frame by frame, producing a time series of facial features. For AU extraction, we used the functions in Exordium[11]. We initially tested Py-Feat[12] [20] but switched to Exordium because it uses OpenGraphAU[13] [59], which extracts a larger set of AUs and proved more suitable for our data, as shown in Table B.1. For the features, we computed common statistics (i.e., min, max, mean, median, std, skewness, and kurtosis) from the AUs as commonly found in the literature, e.g., [14].

Table B.1: Facial AUs potentially relevant to (dis)agreement, with brief descriptions and availability in Py-Feat [20] and OpenGraph [59]. All the AUs (except AU13) are used for disagreement, while agreement only uses AU1, AU2, AU12, and AU13.

AU	Associated States	Py-Feat	OpenGraph
AU01: Raising the inner eyebrows	Surprise, concern, and attentiveness.	✓	✓
AU02: Raising the outer eyebrows	Alertness and mild surprise.	✓	✓
AU04: Drawing the brows together	Frowning, effort, and anger.	✓	✓
AU05: Raising the upper eyelids	Alertness and surprise.	✓	✓
AU07: Tightening the eyelids	Concentration and determination.	✓	✓
AU09: Wrinkling the nose	Disgust and scepticism.	✓	✓
AU10: Raising the upper lip	Contempt and disgust.	✓	✓
AU11: Deepening the nasolabial folds	Tension and disapproval.	✓	✓
AU12: Smiling	Positive engagement and agreement.	✓	✓
AU13: Retracting the lips laterally	Nuanced agreement or mild amusement.	✗	✓
AU14: Dimpling near the mouth	Playful or nuanced smiling.	✓	✓
AU15: Depressing the lip corners	Sadness and displeasure.	✓	✓
AU17: Raising the chin	Determination and incredulity.	✓	✓
AU18: Puckering the lips	Disapproval and pre-speech preparation.	✗	✓
AU19: Subtle lower-lip or chin movement	Uncertainty or slight incredulity.	✗	✓
AU23: Tightening the lips	Restraint and disapproval.	✓	✓
AU24: Pressing the lips together	Suppression and disapproval.	✓	✓
AU25: Parting the lips	Neutral or transitional states.	✓	✓
AU26: Dropping the jaw	Shock and strong surprise.	✓	✓
AU32: Biting the lower lip	Tension and uncertainty.	✗	✓
AU38: Dilating the nostrils	Effort and heightened arousal.	✗	✓
AU43: Closing the eyes	Strong affect or relaxation.	✓	✗
AU44: Narrowing the eyes (squint)	Focused evaluation and scepticism.	✗	✗

C Multimodal Setup

We initially conducted a multimodal analysis combining gaze and facial data; however, preliminary results showed that including facial data did not improve performance. Consequently, we focus on gaze data in the main manuscript. This appendix outlines the original multimodal setup for completeness and to inform future work. The setup was designed to systematically evaluate how different data representations and fusion strategies could leverage the information from both modalities. We represented each modality using two data formats: engineered features and image-based visualisations. For gaze, the engineered features are detailed in subsection 4.2; visual representations included scanpaths and heatmaps (Figure C.1). For facial data, engineered features consisted of statistical summaries of AU activations, as described in Appendix B; the visual representation was aggregated AU heatmaps (Figure C.2), which show changes in facial expression over time[14].

[10] Our dataset is available online: https://github.com/DFKI-Interactive-Machine-Learning/Disagreement-Detection-Dataset-CHI-26
[11] https://github.com/fodorad/exordium (Accessed August 18, 2025)
[12] https://py-feat.org (Accessed August 18, 2025)
[13] https://github.com/CVI-SZU/ME-GraphAU (Accessed August 18, 2025)
[14] https://py-feat.org/basic_tutorials/03_plotting.html (Accessed August 18, 2025)

(a) Gaze Scanpath Image Visualisation (b) Gaze Heatmap Image Visualisation

Figure C.1: Examples showing the gaze data visual representations.

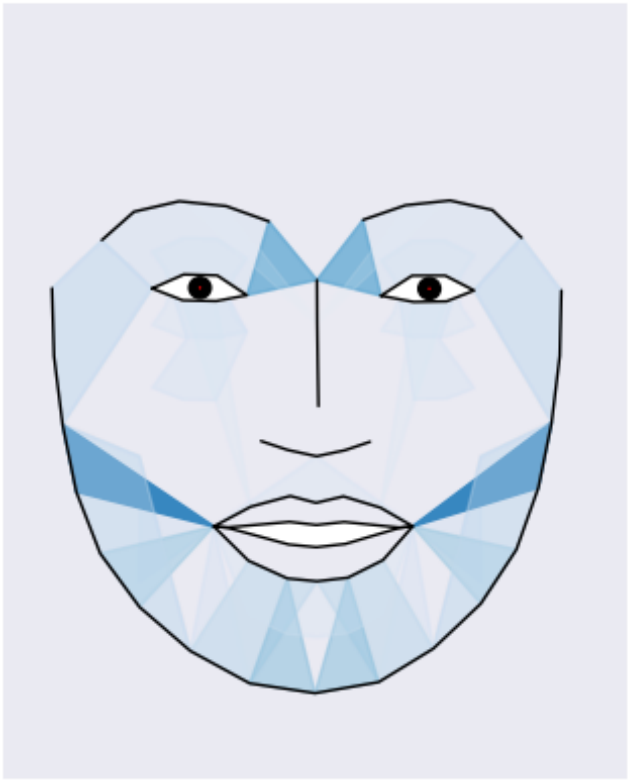


Figure C.2: An example of the facial AU-aggregated heatmap.

To establish performance baselines for each modality and data representation, we conducted an ablation study in addition to the fusion comparisons. For feature-based inputs, we evaluated XGBoost, Multilayer Perceptron (MLP), Support Vector Machines (SVM), and Random Forest (RF) from scikit-learn[15], as well as the Tabular Prior-data Fitted Network (TabPFN)[16] transformer-based model [37]. For image-based inputs, we used pre-trained deep learning architectures: VGG19, a 19-layer convolutional neural network pre-trained on ImageNet-1K [28, 81] via PyTorch[17], and the Vision Transformer (ViT) model (*google/vit-base-patch16-224*) [30, 105] from Hugging Face[18].

We compared two multimodal fusion approaches: early fusion (feature-level) and late fusion (decision-level). **Early Fusion (Feature-Level):** Data were combined at the input level before model training. For feature-based representations, we concatenated the feature vectors from both modalities into a single unified vector. For image-based representations, we combined the two images side-by-side after resizing and rescaling, then fed the result into a single network. **Late Fusion (Decision-Level):** Models were trained independently on each modality, and their outputs were combined. For feature-based data, we trained separate classifiers for each modality and derived the final prediction by averaging their prediction probabilities. For image-based data, we trained separate deep learning backbones for each modality, extracted feature vectors from each, concatenated them, and passed the combined vector through a dedicated classification layer.

D Initial Experiments Additional Results

This Appendix section contains detailed results for classical machine learning experiments reported in subsection 5.1 and Appendix C using statistical features and an RF classifier, which produced the best overall results across the three experimental setups. The multimodal results using both facial and gaze data are shown in Table D.1 and Table D.3; while the unimodal facial data results are shown in Table D.2 and Table D.4.

Table D.1: The best generalised multimodal results using the group-based 5-fold cross-validation. These results were from the early fusion feature concatenation strategy for gaze and facial statistical features.

BA (%)	F1 (%)	Recall (%)	Precision (%)	AUC (%)
51.32	51.35	49.00	53.94	53.85
51.01	37.50	28.07	56.45	52.01
53.18	56.73	73.93	46.02	52.58
51.85	57.97	66.30	51.50	52.58
51.38	57.53	59.27	55.89	50.86
51.75 ± 0.86	52.22 ± 8.65	55.31 ± 17.78	52.76 ± 4.24	52.38 ± 1.08

Table D.2: The best generalised facial data results using the group-based 5-fold cross-validation.

BA (%)	F1 (%)	Recall (%)	Precision (%)	AUC (%)
49.66	48.24	44.40	52.82	50.04
51.17	57.55	62.67	53.21	50.23
48.68	50.45	54.25	47.14	49.11
51.08	41.75	33.58	55.20	52.91
49.13	49.69	53.28	46.56	49.24
49.94 ± 1.13	49.54 ± 5.64	49.64 ± 11.06	50.99 ± 3.89	50.31 ± 1.53

[15]https://scikit-learn.org/stable/ (Accessed August 18, 2025)
[16]https://github.com/PriorLabs/tabpfn-client (Accessed August 18, 2025)
[17]https://pytorch.org/ (Accessed August 18, 2025)
[18]https://huggingface.co/google/vit-base-patch16-224 (Accessed August 18, 2025)

Table D.3: The best personalised results using multimodal data averaged across participants. These results were from the early fusion feature concatenation strategy for gaze and facial statistical features.

User ID	BA (%)	F1 (%)	Recall (%)	Precision (%)	AUC (%)
A01	45.59	54.59	58.25	52.08	50.11
A02	49.89	52.68	54.28	51.61	50.42
A03	55.30	57.50	57.83	61.54	58.04
A04	56.14	35.54	30.39	46.77	62.99
A05	55.93	60.37	63.58	58.73	61.28
A06	49.45	49.48	55.00	45.87	47.01
A07	57.49	67.52	73.25	63.87	53.55
A08	71.92	69.95	65.51	75.41	77.15
A09	60.49	58.30	61.32	57.23	60.80
A10	44.89	46.94	48.09	46.38	40.07
A11	53.10	59.13	64.87	54.51	57.19
A12	62.20	70.65	73.22	68.69	68.59
A13	46.35	53.68	59.44	49.08	46.56
A14	61.71	57.65	57.43	59.35	59.75
A15	60.32	60.21	65.51	56.25	58.76
A16	37.15	46.38	49.02	44.91	42.11
B01	50.50	01.60	01.00	04.00	50.53
B03	52.94	36.70	29.72	48.48	63.69
B04	51.28	66.09	72.87	60.93	52.66
B05	53.81	64.84	71.43	59.40	53.72
B06	56.49	51.89	50.49	56.18	58.65
B07	48.73	60.13	64.66	56.34	51.28
B08	42.02	34.54	32.76	37.07	43.88
B09	61.82	65.47	76.64	58.81	66.07
B10	55.75	76.01	84.17	71.44	63.18
B11	55.31	55.61	53.69	58.79	55.63
B12	55.62	70.45	79.81	63.29	60.64
B13	60.20	58.38	54.64	64.38	59.32
B14	52.91	70.27	79.30	63.65	51.75
B15	54.80	66.09	74.13	59.91	53.07
Mean ± SD	54.00 ± 6.93	55.95 ± 14.70	58.74 ± 17.67	55.16 ± 12.76	55.95 ± 8.13

Table D.4: The best personalised results using facial data averaged across participants.

User ID	BA (%)	F1 (%)	Recall (%)	Precision (%)	AUC (%)
A01	52.57	59.37	61.57	57.57	53.81
A02	52.56	57.79	64.11	53.54	48.66
A03	54.04	58.05	59.79	58.74	53.29
A04	43.60	10.69	08.57	15.42	41.27
A05	51.29	55.86	57.74	54.27	60.45
A06	50.85	50.01	48.76	51.88	50.96
A07	56.05	68.02	75.66	62.63	55.00
A08	67.32	65.91	63.38	68.66	68.89
A09	53.77	52.96	54.59	51.59	52.76
A10	41.99	41.66	40.45	43.67	37.72
A11	46.64	51.89	57.00	48.77	45.33
A12	55.02	65.42	67.73	63.35	57.48
A13	45.12	48.74	50.06	47.51	49.10
A14	50.31	37.82	33.67	45.06	56.96
A15	50.95	49.68	52.04	49.62	51.83
A16	66.38	71.70	73.85	70.61	73.29
B01	49.97	01.60	01.00	04.00	43.17
B03	52.19	45.10	46.47	45.38	53.99
B04	49.53	63.47	67.89	59.94	45.06
B05	52.87	66.54	75.67	59.67	55.89
B06	46.46	44.04	43.15	46.37	46.29
B07	49.09	59.55	64.10	55.81	50.03
B08	46.71	39.63	37.05	42.87	42.92
B09	47.60	47.29	47.87	47.24	43.72
B10	65.98	64.22	62.72	66.13	62.57
B11	54.31	56.18	55.68	57.12	52.75
B12	42.39	60.05	69.69	52.83	37.52
B13	55.77	53.31	50.88	57.51	54.92
B14	46.16	67.40	79.02	58.82	51.49
B15	55.31	59.70	61.31	58.88	54.18
Mean ± SD	51.76 ± 6.37	52.46 ± 15.47	54.38 ± 17.76	51.85 ± 13.66	51.71 ± 8.16