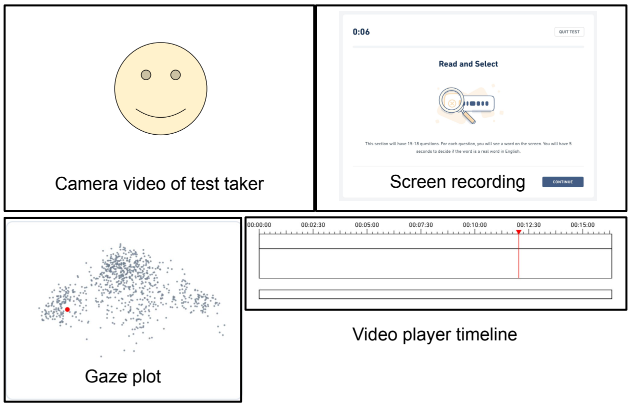
AI-assisted Gaze Detection for Proctoring Online Exams

Yong-Siang Shih[1], **Zach Zhao**[1], **Chenhao Niu**[1], **Bruce Iberg**[2], **James Sharpnack**[1],
Mirza Basim Baig[1]

Duolingo, Inc.
[1]{yongsiang,zach,chenhao,james.sharpnack,basim}@duolingo.com, [2]bruceiberg@duolingocontractors.com

Abstract

For high-stakes online exams, it is important to detect potential rule violations to ensure the security of the test. In this study, we investigate the task of detecting whether test takers are looking away from the screen, as such behavior could be an indication that the test taker is consulting external resources. For asynchronous proctoring, the exam videos are recorded and reviewed by the proctors. However, when the length of the exam is long, it could be tedious for proctors to watch entire exam videos to determine the exact moments when test takers look away. We present an AI-assisted gaze detection system, which allows proctors to navigate between different video frames and discover video frames where the test taker is looking in similar directions. The system enables proctors to work more effectively to identify suspicious moments in videos. An evaluation framework is proposed to evaluate the system against human-only and ML-only proctoring, and a user study is conducted to gather feedback from proctors, aiming to demonstrate the effectiveness of the system.



This section will have 15-18 questions. For each question, you will see a word on the screen. You will have 5 seconds to decide if the word is a real word in English.

Figure 1: The user interface of the gaze system allows proctors to navigate to different video frames using the video player timeline. The points on the gaze plot represent the gaze direction of each frame. When proctors select regions on the gaze plot, the corresponding frames on the timeline would be highlighted. The gaze direction of the current frame is colored in red in the gaze plot.

Introduction

The adoption of online proctoring systems has grown in recent years (Nigam et al. 2021). Online tests offer greater flexibility because test takers can take the test remotely without going to a specific test center. However, the problem of cheating is a threat to the validity of the test results (Bilen and Matros 2021). Therefore, security measures need to be built to detect and prevent cheating behaviors.

Online proctoring comes in various forms, including live synchronous proctoring where the proctor watches the test taker remotely during the test session, and asynchronous proctoring where video recordings of the test sessions are recorded and reviewed by proctors. In this study, we focus on the application of online proctoring in the Duolingo English Test (DET) (Cardwell et al. 2024), which is an online, high-stakes English assessment test where a test taker's video is recorded with the test taker's webcam. The test taker's video, the screen recording, the responses, and other relevant information are collected, and proctors review each test session asynchronously.

We focus our study on the task of detecting if test takers are looking away from the screen suspiciously, as such a behavior could indicate that test takers are consulting external resources. There are two challenges proctors face when examining such behaviors. Firstly, when the exam video length is long, it could be tedious for the proctors to watch the entire video to find all moments where the test taker is looking away. In addition, as pointed out by Belzak, Lockwood, and Attali (2024), a test taker could also naturally look at arbitrary spots as part of their cognitive processing. With the limited amount of information available in the exam videos, different proctors' decisions could be less consistent for this task compared to other tasks such as detecting plagiarism.

In this study, we present an AI-assisted gaze detection system, where the predicted gaze direction of the test taker in each frame is shown on a scatter plot. The user interface is shown in Figure 1. Proctors can select regions on the gaze plot, and the related timestamps on the video player timeline will be highlighted. This allows proctors to navigate to relevant video frames more efficiently, which improves the proctoring experience. In addition, because the system enables a consistent view on the test taker's gaze directions, the quality of the proctoring could also be improved.

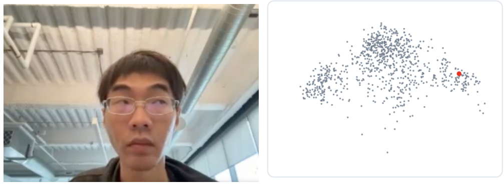


Figure 2: Each frame of the test session is shown as a point in the gaze plot, and the position of each point represents the gaze direction in each frame. The current frame's location is colored in red.

We propose to evaluate our AI-assisted gaze detection system against (1) the human-only system, and (2) the ML-only system in an end-to-end fashion. We choose to evaluate the end-to-end performance so that we could capture the effects of the biases that arise when proctors interact with the AI system (Cummings 2017; Selten, Robeer, and Grimmelikhuijsen 2023; Bashkirova and Krpan 2024). Our framework allows us to properly determine if the proposed system could have a positive impact when deployed into production.

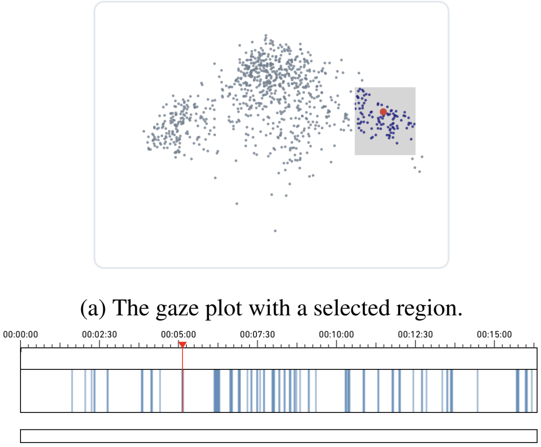
The remaining part of this paper is organized into four sections. Firstly, we describe how the proposed gaze detection system works, including the user interface and how proctors would interact with the system. Secondly, we describe our proposed evaluation framework, including a concrete definition of the task being evaluated. Thirdly, we present the results of a user study where we let proctors try out the system. Finally, we conclude the paper with a discussion on the limitations and future works for our study.

System Overview

Our system is designed for asynchronous proctoring of online exams. When a test taker takes a test, a video is recorded for the entire test session, and once the video is uploaded, a gaze detection model can be run on each frame of the video to predict the gaze direction in each frame. In practice, a suitable frame rate would need to be selected for inference according to the resource constraints.

The gaze direction predictions will be displayed to proctors as a scatter plot as shown in Figure 2. In particular, the gaze angles predicted by the model can be represented as unit directional vectors originating from the origin, and these vectors are projected onto the 2D plane, with each point on the plot representing a frame and its associated gaze direction. Currently, our gaze plot only represents the gaze directions (i.e., the angles of the gazes), and not the exact location on the screen where the test taker is looking at. However, a similar plot can also be used for models that predict the exact screen location.

Proctors can select regions on the eye gaze plot and the corresponding frames will be highlighted on the video player timeline. This allows proctors to navigate to frames with similar gaze directions within the selected region. For instance, if a proctor observes a suspicious moment when the test taker is looking away from the screen, the proctor can consult the gaze plot to find the current video frame's location, and select a region around it. The system will highlight all other relevant timestamps, allowing the proctor to navigate to those timestamps and confirm whether the test taker is also exhibiting suspicious behaviors at those moments.



(a) The gaze plot with a selected region.

(b) The video player timeline with frames highligted.

Figure 3: The video player timeline allows proctors to navigate to specific moments by clicking on the desired timestamp. Timestamps where gaze predictions fall within the selected region of the gaze plot are highlighted in blue. The space on top of the timeline can be used to show other notable events. The white bar below is used to select a specific time interval to zoom into.

Evaluation Framework

To evaluate the effectiveness of the system in human-based asynchronous proctoring, we apply the concept of *human-ML complementarity* (Rastogi et al. 2023) to define the evaluation goals and propose our experiment plans.

Human-ML Complementarity

In a hybrid decision-making system like AI-assisted proctoring, human-ML complementarity is the condition where the hybrid system outperforms both humans and ML models. Following the notation used by Rastogi et al. (2023), denote $\mathcal{X}$ as the set of all available features of a given test session, including video recording, responses, scores, etc. Denote the action space as $\mathcal{A}$, where for a test session with T frames, $\mathbf{a} \in \mathcal{A}$ is a binary sequence with length T, and $\mathbf{a}_t$ indicates whether the t -th frame is labeled positive (i.e. looking away from the screen) or not. Then a decision-making system for labeling gaze direction in a test session can be written as a mapping $\pi : \mathcal{X} \to \mathcal{A}$. Denote Π as the set of all possible π.

In this work, there are three systems of interest: (1) the human-only system π_H, where human proctors label the test session mainly by watching the video; (2) the ML-only system π_M, where binary predictions are made by thresholding predicted gaze directions on each frame; and (3) the hybrid system π_{H+M}, where human proctors label the test session with additional access to predicted gaze directions.

Empirically, it is less likely that a π_M is better than π_H, as the gaze detection system only has access to a frame to make each prediction, while human proctors have more context from the whole test session than a frame. However, it is possible that π_{H+M} is a better system than π_H and π_M through human-ML complementarity. That is, with an evaluation function $F : \Pi \to \mathbb{R}$, we want to verify human-ML complementarity: $F(\pi_{H+M}) > \max\{F(\pi_H), F(\pi_M)\}$.

Proposed Experiments

On a dataset with N test sessions, we define the evaluation function F as:

$$F(\pi) = \frac{1}{N} \sum_{i=1}^{N} s(\mathbf{X}^{(i)}, \pi(\mathbf{X}^{(i)}))$$

Where $s : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is a scoring function of a labeling result for a given test session, regardless of where the labeling result comes from. Without access to ground truth labels, we use a labeling process with multiple proctors to generate high-precision labels to define s.

Specifically, given the i-th test session, $\mathbf{a}_H^{(i)} = \pi_H(\mathbf{X}^{(i)})$ is the labeling result from a proctor without using the gaze plot, $\mathbf{a}_M^{(i)} = \pi_M(\mathbf{X}^{(i)})$ is the labeling result made by thresholding the predicted gaze directions in each frame, and $\mathbf{a}_{H+M}^{(i)} = \pi_{H+M}(\mathbf{X}^{(i)})$ is the labeling result from a proctor using the gaze plot. For the three binary vectors $\mathbf{a}_H^{(i)}$, $\mathbf{a}_M^{(i)}$, and $\mathbf{a}_{H+M}^{(i)}$, we collect all the positive intervals, and present the intervals for a group of K proctors to label (without gaze plot), and take the majority opinion $\mathbf{a}^{*(i)}$ as the reference for comparison.

Note that we ensure high precision for $\mathbf{a}^{*(i)}$ by using multiple proctors to reduce variance and selecting only positive intervals instead of the entire video to reduce tediousness. However, this also means that if an interval is labeled as negative by all three systems, it will not be labeled differently in this process.

Comparing $\mathbf{a}_H^{(i)}$, $\mathbf{a}_M^{(i)}$, and $\mathbf{a}_{H+M}^{(i)}$ with $\mathbf{a}^{*(i)}$, we can calculate the average precision and (upper-bounded) recall of each system as $F(\pi_{H+M})$, $F(\pi_H)$, and $F(\pi_M)$. Conducting this experiment is the next step in this project.

User Study

We also conducted a user study, where we recruited 11 DET proctors to try out the AI-assisted gaze detection system on 300 test sessions sampled from DET. The proctoring results were not used for the official certification, but we collected the feedback from the proctors regarding the gaze detection system with a survey form.

The survey is based on a scale of 1-5, where 1 represents "absolutely disagree" and 5 represents "absolutely agree". Here we show the survey questions and the final averaged scores in Figure 4. Positive responses were received in the user study.

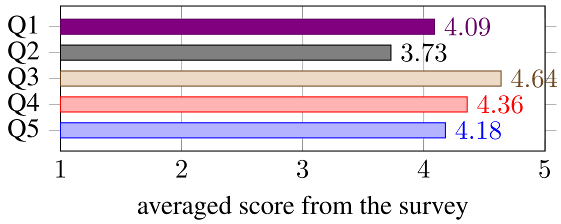


Figure 4: Survey questions[1]: (Q1) I felt comfortable utilizing the tool, (Q2) I felt confident that the tool was providing me with correct information, (Q3) I felt the documentation/videos provided allowed me to easily understand how to use the tool, (Q4) I didn't have difficulty interpreting or understanding any visual elements of the tool, (Q5) I found it easy to incorporate the tool in my normal proctoring processes.

Conclusion

This paper presents an AI-assisted gaze detection system, which enables proctors to work effectively in finding the moments where a test taker is looking away from the screen. For the demo, we plan to show the gaze detection system on a laptop with an example test session, and the audience would be able to play with the system and give us feedback.

Limitations

We acknowledge that our system still has limitations, and future work will be needed to further improve the design. Firstly, the gaze plot only shows the gaze directions of the test takers, it doesn't show where on the screen the test taker is actually looking at. Therefore, the gaze plot should not directly be used alone to determine if the test taker is looking away. We expect the ML-only system to perform poorly because calibration will be needed to determine the exact relative positional relationships between the screen, the camera, and the test taker. Secondly, our system currently only works in an asynchronous proctoring environment, where the exam video is recorded. If synchronous proctoring is required, real-time prediction would be needed and the predictions need to be gradually added into the gaze plot. Finally, our proposed evaluation is still based on proctor decisions, and therefore is limited by the information that could be derived from the recorded information. To further improve the accuracy of the evaluation, we could have test takers taking the exams in a controlled environment where the camera and the screen are carefully calibrated. This will allow us to gather accurate measurements for test takers' eye gazes.

Acknowledgments

We thank the colleagues who had tested the system, provided feedback, reviewed the code, or reviewed the paper.

[1] Questions that are related to the details of the internal proctoring process are omitted, and the expression of Q4 and its score were reversed to make the interpretation of the scores more consistent. The original Q4 asked if proctors had difficulty.

References

Bashkirova, A.; and Krpan, D. 2024. Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. *Computers in Human Behavior: Artificial Humans*, 2(1): 100066.

Belzak, W.; Lockwood, J.; and Attali, Y. 2024. Measuring Variability in Proctor Decision Making on High-Stakes Assessments: Improving Test Security in the Digital Age. *Educational Measurement: Issues and Practice*, 43(1): 52–65.

Bilen, E.; and Matros, A. 2021. Online cheating amid COVID-19. *Journal of Economic Behavior & Organization*, 182: 196–211.

Cardwell, R.; Naismith, B.; LaFlair, G. T.; and Nydick, S. 2024. Duolingo english test: technical manual. https://go.duolingo.com/dettechnicalmanual.

Cummings, M. L. 2017. Automation bias in intelligent time critical decision support systems. In *Decision making in aviation*, 289–294. Routledge.

Nigam, A.; Pasricha, R.; Singh, T.; and Churi, P. 2021. A systematic review on AI-based proctoring systems: Past, present and future. *Education and Information Technologies*, 26(5): 6421–6445.

Rastogi, C.; Leqi, L.; Holstein, K.; and Heidari, H. 2023. A Taxonomy of Human and ML Strengths in Decision-Making to Investigate Human-ML Complementarity. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, 127–139.

Selten, F.; Robeer, M.; and Grimmelikhuijsen, S. 2023. ‘Just like I thought’: Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review*, 83(2): 263–278.