UNIVERSITÀ
DI PAVIA

UNIVERSITY OF PAVIA
FACULTY OF ENGINEERING
DEPARTMENT OF ELECTRICAL, COMPUTER AND
BIOMEDICAL ENGINEERING

MASTER'S DEGREE IN COMPUTER ENGINEERING

MASTER THESIS

Detecting Behavioral Randomness in Multiple-Choice Tests Through Gaze Dynamics

Rilevamento della casualità comportamentale nei test a scelta multipla tramite l'analisi delle dinamiche dello sguardo

Candidate: **Behnaz Amiri**

Supervisor: Prof. Marco Porta

Co-supervisors: Prof. Piercarlo Dondi and Mr. Hoàng Nam Lê

A.Y. 2024/2025

Abstract

The increasing use of computer-based assessment in e-learning environments has highlighted the importance of process-oriented data for understanding test-taking behavior beyond response correctness. Among available process measures, eye-tracking data provide detailed information about attention allocation and visual engagement during task solving. Despite this potential, the detection of disengaged or random-like responding in digital assessments is often based on response time heuristics or accuracy-dependent criteria. Such criteria may not fully capture the behavioral structure underlying response processes.

This thesis develops a structured and reproducible framework for identifying random-like responding based on gaze-derived temporal engagement markers and response timing, independently of answer correctness. A three-stage analytical pipeline is proposed. First, raw eye-tracking data are preprocessed and mapped to predefined Areas of Interest (AOIs) representing relevant task regions. Second, an operational attention threshold t_a is introduced to distinguish minimally engaged responses from more sustained engagement. Third, gaze-based behavioral features are derived to characterize response patterns independently of answer correctness. Random-like responding is thus defined as insufficient or structurally unorganized visual engagement with task-relevant regions rather than as incorrect responding.

The framework is evaluated using eye-tracking data collected in an e-learning assessment setting. Behavioral and comparative analyses examine differences between engagement conditions, and sequence-based deep learning models are employed to explore temporal gaze patterns associated with varying levels of behavioral engagement. The results suggest that the proposed attention threshold may help differentiate systematic from random-like response behavior and that temporal models can capture structured gaze dynamics without relying on correctness information. Among the evaluated architectures, the LSTM model achieved the best performance under the reported hold-out evaluation setup, reaching an accuracy of 76.47% and an F1-score of 82.86%. The thesis contributes a methodologically transparent approach to behavioral validity analysis in digital assessment contexts.

Sommario

Il crescente impiego di sistemi di valutazione basati sul computer nel contesto dell'e-learning ha evidenziato la rilevanza dei dati orientati ai processi per l'analisi delle risposte degli utenti nell'interazione con i sistemi di valutazione digitali, andando oltre la sola valutazione della correttezza delle risposte di per sé. In tale contesto, i dati di eye tracking costituiscono una fonte informativa rilevante per lo studio dell'attenzione visiva e del livello di coinvolgimento comportamentale durante la risposta a domande a scelta multipla. Nonostante questo potenziale, l'individuazione di risposte fornite casualmente, senza impegno, è spesso basata su euristiche legate solo ai tempi di risposta o su criteri di accuratezza, che possono non riflettere adeguatamente il comportamento dell'utente.
La presente tesi propone un framework strutturato e riproducibile per l'identificazione di comportamenti di risposta casuali, basato su indicatori ottenuti tramite eye tracking e sui tempi di risposta, indipendentemente dalla correttezza delle risposte fornite. Il framework è articolato in una pipeline analitica composta da tre fasi principali. In primo luogo, i dati grezzi di eye tracking vengono preelaborati e associati a specifiche Aree di Interesse (Areas of Interest, AOI) che rappresentano le regioni rilevanti del quesito. In secondo luogo, viene introdotta una soglia operativa di attenzione t_a, finalizzata a distinguere risposte caratterizzate da un coinvolgimento visivo minimo da risposte associate a un'attenzione più sostenuta. In terzo luogo, vengono estratte caratteristiche comportamentali basate sullo sguardo, progettate per descrivere i pattern di risposta in modo indipendente dall'accuratezza. In questo approccio, il comportamento di risposta di tipo casuale è definito come un coinvolgimento visivo insufficiente o strutturalmente disorganizzato sulle AOI rilevanti, anziché come una semplice risposta errata.
Il framework proposto viene valutato utilizzando dati di eye tracking raccolti in un contesto di valutazione in ambiente e-learning. Analisi comportamentali e comparative consentono di esaminare le differenze tra diverse condizioni di coinvolgimento, mentre modelli di deep learning basati su sequenze temporali vengono impiegati per analizzare i pattern dinamici dello sguardo associati ai diversi livelli di engagement comportamentale. I risultati sperimentali indicano che la soglia di attenzione proposta contribuisce alla distinzione tra comportamenti di risposta sistematici e comportamenti di tipo random-like e che i modelli temporali sono in grado di catturare dinamiche strutturate dello sguardo senza utilizzare informazioni sulla correttezza delle risposte. Tra le architetture valutate, il modello LSTM ha ottenuto le migliori prestazioni nella configurazione di valutazione hold-out adottata, raggiungendo un'accuratezza del 76,47% e un F1-score pari all'82,86%.

La tesi fornisce un contributo metodologicamente solido e trasparente all'analisi della validità comportamentale nei contesti di valutazione digitale, con potenziali appli-

cazioni nello sviluppo di sistemi di assessment intelligenti e adattivi.

Acknowledgements

I would like to express my profound gratitude to my supervisor, Prof. Marco Porta, for his outstanding guidance, remarkable patience, and unwavering support throughout this research. His intellectual rigor, clarity of vision, and steadfast commitment to academic excellence have been decisive in shaping both the direction and the quality of this thesis. Without his continuous encouragement and thoughtful supervision, the completion of this work in its present form would not have been possible.

I am also grateful to my co-supervisors, Prof. Piercarlo Dondi and Mr. Hoàng Nam Lê, for their constructive feedback and technical expertise, which meaningfully contributed to the development of this work.

Finally, I wish to express my deepest gratitude to my spouse for his unwavering support, encouragement, and for the confidence he continually inspired in me throughout this journey. His constant belief in me and his reassuring presence were a profound source of strength and hope, making it possible to see this work through to its completion.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

The rapid expansion of digital learning and computer-based assessment platforms has led to the increasing availability of fine-grained behavioral process data that extend beyond final response accuracy. Within this landscape, eye-tracking technology provides temporally precise measurements of visual attention allocation during task interaction [1, 2]. By recording gaze coordinates at high sampling rates, eye tracking enables the reconstruction of sequential behavioral phases such as reading, option inspection, comparison, and response commitment.

In multiple-choice assessment environments, conventional scoring mechanisms rely exclusively on the selected response option. However, identical outcomes may arise from fundamentally different cognitive and behavioral processes. A selected answer may reflect systematic inspection and structured comparison of alternatives, or it may result from minimal engagement with the answer area, particularly under temporal constraints [3, 4]. From the perspective of behavioral interpretability and assessment validity, these processes should not be treated as equivalent.

Distinguishing engagement-driven responding from random-like behavior is therefore essential for advancing process-based validity in digital assessment systems and for enabling more reliable modeling of learner interaction dynamics. In this thesis, the notion of **RANDOM** responding is treated as an operational behavioral label derived from gaze-grounded temporal criteria, rather than as a direct measure of stochastic guessing in a strict probabilistic sense.

1.2 Problem Statement

Response time has traditionally served as a proxy indicator for disengaged or rapid responding [5]. While extremely short response times may signal low effort, purely time-based thresholds cannot verify whether a participant visually engaged with answer alternatives prior to submitting a response. Conversely, many eye-tracking studies concentrate on predicting correctness, estimating cognitive load, or modeling latent knowledge states [6], rather than explicitly formalizing random versus non-random responding as an independent behavioral construct.

This reveals a methodological gap: to the best of our knowledge, there is no clearly established, transparent, and reproducible framework that integrates gaze-derived engagement markers with response timing to operationalize random-like responding independently of correctness within the scope of the literature reviewed here. Without such a framework, behavioral interpretation remains either accuracy-dependent or reliant on indirect temporal heuristics.

Addressing this gap requires a principled approach that:

- enforces explicit data quality control for noisy eye-tracking streams,
- derives temporally grounded markers of answer-related visual engagement,
- implements condition-aware behavioral labeling rules aligned with the experimental design.

Accordingly, the central research question of this thesis is formulated as follows:

> Can gaze-derived temporal onset markers and response timing be integrated into a principled, transparent, and reproducible framework for distinguishing random-like from non-random responding in a controlled e-learning assessment environment?

1.3 Research Objectives

To address this research question, the thesis pursues four interrelated objectives:

1. Design and implement a controlled experimental protocol consisting of 15 mathematical multiple-choice questions administered under multiple response conditions. The final item set was selected from standardized-style mathematical reasoning questions reviewed from SAT, GRE, and GMAT sources.Candidate items were screened to retain questions that were text-based, did not require diagrams, tables, or calculators, were neither excessively short nor excessively long, and followed a broadly comparable structural format. A preliminary pool of approximately 20 such items was then difficulty-screened by a separate validation group, from which the final 15 questions were selected for the experiment. The assessment was administered through a custom experimental interface.

2. Collect eye-tracking recordings (150 Hz) capturing complete trial-level gaze dynamics from question onset to response submission.

3. Develop a deterministic three-stage behavioral processing pipeline that transforms raw gaze streams into interpretable trial-level engagement markers while explicitly controlling for data quality.

4. Formalize a condition-aware operational definition of **RANDOM** and **NOT_RANDOM** responding based on post-engagement commitment time, independently of response correctness, thereby enabling downstream behavioral analysis and sequence based deep learning modeling.

1.4 Research Questions

The study is guided by the following research questions:

- **RQ1:** How do temporally constrained and non-constrained assessment conditions differ in the distribution of gaze-grounded engagement markers and derived behavioral labels during multiple-choice responding?

- **RQ2:** Can a gaze-grounded temporal onset marker (t_a) provide a consistent and empirically supported operational boundary between engagement-driven and random-like responding?

- **RQ3:** To what extent can data-driven sequence models learn discriminative patterns from gaze trajectories for classifying **RANDOM** and **NOT_RANDOM** labels under leakage-aware evaluation constraints?

1.5 Contributions

The primary contributions of this thesis are:

- A controlled eye-tracking dataset collected during a multi-condition mathematical multiple-choice assessment with explicitly defined temporal and response-related manipulations, based on a curated and difficulty-screened set of mathematical reasoning questions selected from SAT-, GRE-, and GMAT-style sources and administered through a custom-built interface.
- A fully specified three-stage behavioral processing pipeline integrating rule-based quality filtering, AOI-grounded stable engagement detection, and condition-aware timing thresholds.
- A transparent operational definition of **RANDOM** and **NOT_RANDOM** responding based on post-engagement commitment time,

$$t_{\text{answer}} = t_{\text{end}} - t_a,$$

  explicitly decoupled from correctness.
- A unified modeling framework supporting both behavioral analysis and sequence-based deep learning classification under leakage-aware evaluation protocols.

1.6 Thesis Organization

The remainder of this thesis is organized as follows.

Chapter 2 introduces the methodological and conceptual foundations underlying the proposed behavioral framework.

Chapter 3 reviews prior research on gaze-based assessment analytics and rapid responding detection, positioning the present work within the literature.

Chapter 4 describes the experimental design and data acquisition protocol.

Chapter 5 details the three-stage behavioral processing pipeline used to derive engagement-based behavioral markers.

Chapter 6 presents the experimental results of the behavioral pipeline and integrates them with the modeling findings.

Chapter 7 presents the deep learning modeling framework and evaluation strategy used to predict behavioral randomness labels from eye-tracking sequences.

Chapter 8 reports the results of the deep learning experiments and compares the evaluated model families.

Finally, Chapter 9 summarizes the contributions of the thesis and outlines directions for future research.

Chapter 2

Conceptual and Methodological Foundations

This chapter establishes the conceptual and methodological framework that underlies the behavioral pipeline proposed in this thesis. Grounded in a process-oriented perspective on assessment [7, 8], the chapter formalizes the principles that guide the transformation of raw gaze streams into interpretable, trial-level behavioral constructs. Rather than reviewing prior studies (addressed in Chapter 3), the focus here is on defining the theoretical assumptions and operational decisions that structure the proposed framework.

The framework rests on three methodological pillars:

1. Explicit multi-level data quality control,
2. AOI-driven temporal onset detection,
3. Distribution-aware timing thresholds for operationalizing random-like responding.

Together, these pillars define a principled and reproducible pathway from low-level eye-tracking signals to high-level behavioral labels.

2.1 Eye Tracking as Response-Process Evidence

In computer-based assessment environments, response-process evidence refers to temporally structured signals describing how a participant interacts with an item, beyond

the final selected option [7]. From a measurement perspective, such evidence contributes to the interpretability of response behavior by revealing intermediate engagement phases that remain latent in accuracy-only scoring models.

Eye tracking provides such evidence by continuously recording gaze coordinates during task execution [1]. From a process-oriented viewpoint, a multiple-choice trial can be decomposed into sequential behavioral phases:

- Initial reading of the question stem,
- Visual transition toward answer alternatives,
- Inspection and comparison of options,
- Decision commitment and response submission.

These phases are not directly observable from response accuracy alone. Eye tracking enables reconstruction of their temporal ordering and duration, thereby supporting behavioral interpretation grounded in measurable attention allocation rather than inferred outcome states.

Within this thesis, gaze signals are treated not as proxies for correctness, but as structured temporal indicators of engagement with task-relevant regions, particularly the answer options.

2.2 Formalizing Data Quality in Gaze Streams

Eye-tracking recordings inherently contain noise due to tracking loss, blinks, head movement, or off-screen gaze [1]. Without systematic filtering, such artifacts can distort derived temporal markers and compromise interpretability.

Data quality control is therefore conceptualized as a foundational transformation stage rather than as a peripheral preprocessing step. Let a trial consist of N gaze samples. Each sample must be evaluated for reliability before any higher-level behavioral metric is computed.

Quality control operates at two hierarchical levels:

1. **Sample-level validation:** Each gaze sample is classified as valid or invalid based on device-provided reliability indicators.

2. **Trial-level reliability:** Trials exceeding a predefined proportion of invalid samples are excluded from further analysis.

This hierarchical filtering ensures that downstream temporal markers are derived from stable and interpretable data segments rather than from noise-contaminated streams.

2.3 AOI-Driven Temporal Onset Detection

Areas of Interest (AOIs) provide a semantic partition of the task interface. When interface geometry is fixed and known, each gaze sample can be deterministically mapped to a functional region (e.g., Question, individual Answer AOIs, Timer, Submit) [9].

Within the present framework, a critical behavioral event is defined as the onset of stable attention to the answer options, denoted t_a. Importantly, this onset is defined with respect to sustained engagement within the *same answer AOI*, rather than merely entering a general answer area.

Rather than relying on instantaneous AOI transitions, stability is operationalized using a fixed-duration sliding window. Let W denote a temporal window length and θ a coverage threshold. A stable onset is detected when the proportion of samples within the window that fall within the same answer AOI exceeds θ.

This definition provides three desirable properties:

- Robustness to brief excursions outside the AOI,
- Reduced sensitivity to transient noise,
- Interpretability as a transition from reading-oriented to answer-oriented processing.

The resulting marker t_a establishes a gaze-grounded temporal boundary separating pre-answer engagement from post-engagement commitment.

2.4 Randomness as an Operational Post-Engagement Construct

Random responding is frequently inferred from short overall response times [5]. However, total response time does not distinguish between rapid commitment following systematic inspection and minimal engagement prior to selection.

In this thesis, **RANDOM** responding is treated as an *operational behavioral label* derived from temporal engagement patterns, rather than as a direct measure of stochastic guessing. Specifically, randomness is defined relative to post-engagement commitment duration.

Post-engagement commitment time is defined as:

$$t_{\text{answer}} = t_{\text{end}} - t_a$$

where t_{end} denotes total trial duration and t_a represents the onset of stable attention to a specific answer option.

Under this formulation, random-like responding corresponds to relatively short commitment intervals following answer engagement. This definition remains independent of correctness outcomes and instead relies on temporally grounded behavioral markers derived from gaze data and response timing.

Operational thresholds are computed condition-wise to avoid mixing distributions across experimentally distinct timing regimes, thereby preserving interpretability and reducing the risk of structural confounding [10, 11].

2.5 Conceptual Integration

The three pillars outlined above jointly define a deterministic mapping:

$$\begin{aligned} \text{Raw Gaze Stream} &\longrightarrow \text{Quality-Controlled Trial} \\ &\longrightarrow t_a \\ &\longrightarrow \left(t_{\text{end}},\, t_{\text{answer}}\right) \\ &\longrightarrow \text{Operational Randomness Label} \end{aligned}$$

This structured transformation ensures that behavioral labels are explicitly grounded in measurable temporal events, decoupled from correctness, and reproducible under fixed configuration parameters. By formalizing randomness as an operational and transparent construct, the framework provides a stable conceptual bridge between low-level gaze signals and high-level behavioral interpretation.

The next chapter positions this framework within the broader literature, and Chapter 5 formalizes its algorithmic implementation.

Chapter 3

Related Work

This chapter situates the present thesis within prior research at the intersection of process data analytics, eye-tracking–based behavioral modeling, and rapid responding detection in computer-based assessment. Rather than merely summarizing existing findings, this chapter critically examines how prior approaches conceptualize engagement, timing, and behavioral labeling, and clarifies the methodological gap addressed by the present framework.

The literature relevant to this work can be organized into four strands: (i) eye tracking as response-process data, (ii) AOI-driven behavioral modeling, (iii) rapid responding and disengagement detection, and (iv) gaze-based predictive modeling.

3.1 Eye Tracking as Response-Process Data in Assessment

Research in educational measurement has consistently emphasized that response accuracy alone provides an incomplete representation of examinee behavior [3, 7]. From a validity perspective, process data—capturing temporally structured interaction traces—offer valuable insight into intermediate cognitive and behavioral states that remain unobservable in outcome-based scoring models [12, 13].

Beyond aggregate response times, process data provide temporally fine-grained information about how examinees interact with assessment items. In technology-based environments, such data include sequences of actions, navigation traces, and visual attention signals [7]. Among these, eye tracking offers a relatively direct measurement

of attention allocation, enabling the observation of moment-to-moment interaction dynamics at high temporal resolution [1, 14].

Within this paradigm, eye tracking has been used to capture visual attention dynamics during task interaction. Empirical studies have examined gaze transitions between question stems and answer options to characterize reading strategies, search patterns, and decision-making processes [2, 15]. These findings are broadly consistent with foundational research on eye movements and cognitive processing, which establishes a strong association between fixation behavior and information processing [16, 17].

In multiple-answer contexts, systematic gaze transitions between alternatives have been associated with more structured comparison strategies, whereas fragmented patterns may indicate more superficial or heuristic responding. Such observations suggest that gaze allocation patterns can reflect differences in cognitive effort, depth of processing, and strategy selection beyond correctness outcomes.

More recent research in digital learning environments further supports the use of eye movement data as an informative signal for studying learner engagement and interaction patterns [18, 8]. In addition, survey and review papers highlight the growing role of eye tracking in educational analytics and machine learning applications [19, 20, 21].

However, most existing studies adopt a performance-centered perspective, focusing on predicting accuracy or inferring latent cognitive states. As a result, behavioral constructs such as disengaged or random-like responding are typically addressed indirectly or remain intertwined with performance outcomes. Moreover, gaze data are often used as auxiliary predictors rather than as the primary basis for defining behavioral constructs.

In contrast, the present work adopts a process-centered perspective, in which gaze signals are treated as primary evidence for defining behavioral engagement. In particular, random-like responding is operationalized as a temporal property of gaze behavior, rather than inferred from correctness or global response time alone.

3.2 AOI-Based Behavioral Modeling

Area-of-Interest (AOI) analysis is a standard methodological approach in eye-tracking research, particularly when interface geometry is fixed and well-defined [1]. By partitioning the visual interface into semantically meaningful regions, AOIs enable deterministic mapping of gaze samples to functional components, supporting interpretable feature extraction such as dwell time, fixation counts, and transition patterns.

Such mappings allow the transformation of continuous gaze streams into semantically interpretable sequences, facilitating the identification of interaction phases and behavioral transitions. In structured interfaces, AOI definitions can be made fully deterministic, enabling reproducible analysis across participants and experimental conditions.

Prior studies have employed AOI-based metrics to model decision dynamics in multiple-choice tasks and learning environments [8, 6]. Additionally, fixation detection algorithms—such as dispersion-based or velocity-based methods—are commonly used to identify stable attention episodes within gaze streams [9].

Recent work has further explored how gaze transitions can be structured into temporally meaningful sequences, enabling more advanced behavioral modeling approaches [22, 23]. These approaches highlight the importance of temporal structure in interpreting gaze behavior.

Despite these advances, most AOI-based approaches remain either descriptive (e.g., summarizing attention patterns) or predictive (e.g., classifying outcomes). A recurring limitation is that the transition between different phases of the response process is often not explicitly formalized. In particular, the shift from reading-oriented processing to answer-related engagement is rarely defined using a deterministic and reproducible temporal criterion.

The present framework addresses this limitation by introducing a window-based stability criterion to define a gaze-grounded onset marker (t_a) based on sustained attention within a specific answer option. This marker provides a principled temporal boundary within each trial, enabling a structured decomposition of the response process into pre-engagement and post-engagement phases.

3.3 Rapid Responding and Disengagement Detection

Rapid responding has been widely studied as an indicator of low effort or disengagement in computer-based assessment [5, 4]. A common methodological approach involves defining response-time thresholds based on empirical distributions, often at the item or test-section level.

While such methods are effective for large-scale screening, they have an inherent limitation: total response time aggregates multiple behavioral phases, including reading, option inspection, and final selection [10]. Consequently, short response times do not necessarily imply lack of engagement, as examinees may still visually process relevant information before responding.

Empirical evidence further suggests that response time is associated with engagement and effort, with shorter times often reflecting reduced cognitive investment [3, 24]. More recent work emphasizes the importance of effort-based validation frameworks for interpreting response behavior [25].

More advanced statistical approaches, such as mixture models and response-time–based latent variable models, have been proposed to better capture disengagement behavior [11]. However, these methods remain fundamentally time-based and typically do not incorporate direct evidence of visual attention.

While response-time thresholds provide a practical approximation of disengagement, they lack specificity with respect to underlying behavioral mechanisms. In particular, they cannot distinguish between rapid but visually engaged processing and genuinely superficial responding.

The present thesis extends this line of research by grounding the timing construct in gaze-derived evidence. By defining post-engagement commitment time ($t_{\text{answer}} = t_{\text{end}} - t_a$), the proposed framework separates reading duration from decision commitment, providing a more behaviorally interpretable indicator of random-like responding.

3.4 Gaze-Based Predictive Modeling

Recent advances in learning analytics and human-computer interaction have leveraged eye-tracking data within predictive modeling frameworks [7, 26]. These approaches aim to infer cognitive states, predict performance, or classify user strategies based on gaze features.

In particular, sequence-based deep learning architectures—including recurrent neural networks, convolutional temporal models, and attention-based methods—have been applied to gaze trajectories [27, 28]. More recent studies suggest that deep learning models can capture complex temporal dependencies in eye movement patterns and support classification tasks in interactive systems [29, 30].

Related work in sequence modeling and educational data mining further highlights the effectiveness of deep architectures in capturing temporal dependencies in behavioral data [31]. These approaches provide a methodological foundation for modeling gaze trajectories as structured sequences.

However, the effectiveness of such models critically depends on the quality and validity of the target labels used during training. When labels are weakly defined or confounded with performance outcomes, predictive models may capture superficial correlations rather than meaningful behavioral patterns.

This highlights the importance of constructing labels that are both theoretically grounded and operationally transparent. In this context, the present work emphasizes the role of deterministic, gaze-based labeling as a prerequisite for reliable predictive modeling.

The present work addresses this issue by explicitly separating label construction from predictive modeling. A deterministic preprocessing pipeline is first used to generate interpretable behavioral labels based on gaze-derived temporal markers. These labels are then used as targets in downstream modeling, helping ensure that the predictive task remains aligned with a well-defined behavioral construct.

3.5 Comparative Positioning and Research Gap

The reviewed literature indicates that eye tracking provides rich process-level evidence beyond correctness, AOI-based methods support interpretable feature extraction, rapid

responding can serve as a proxy for disengagement, and gaze trajectories can be modeled using advanced machine learning techniques.

Nevertheless, existing approaches tend to address these components in a fragmented manner. In particular, prior work does not typically integrate, within a single framework:

- explicit multi-level data quality control,
- a deterministic gaze-grounded onset marker for answer-related processing,
- a principled separation between reading and post-engagement commitment phases,
- condition-aware timing thresholds aligned with experimental design,
- and correctness-independent behavioral labeling.

This fragmentation can limit interpretability and reproducibility, as behavioral constructs are often implicitly defined or partially confounded with performance metrics.

The present thesis addresses this gap by proposing a deterministic and reproducible behavioral pipeline that integrates these components into a coherent framework. By grounding randomness in observable gaze dynamics and explicitly separating preprocessing from predictive modeling, the approach aims to provide a transparent and methodologically consistent basis for analyzing engagement behavior in digital assessment environments.

In this sense, the contribution of the present work is not only methodological but also conceptual, in that it clarifies how gaze-derived signals can be used to define and model engagement-related constructs in an operationally explicit and reproducible manner.

Chapter 4

Experimental Framework

This chapter describes the experimental framework used to collect the eye-tracking data analyzed in this thesis. Its primary purpose is to document the study design, participant characteristics, assessment materials, interface structure, and data acquisition procedure in sufficient detail to support reproducibility and interpretation of the subsequent analyses [1].

Unlike the deterministic behavioral pipeline introduced in Chapter 5 and the deep learning framework presented later in Chapter 7, the present chapter focuses exclusively on how the data were generated. In particular, it clarifies the experimental conditions under which gaze signals were recorded and the constraints imposed by the assessment environment.

The chapter is organized as follows. First, the participants and stimuli are described. Next, the full experimental procedure is presented, including calibration, participant instructions, and trial flow. The chapter then describes the assessment interface and the resulting data acquisition setup. Finally, a brief summary of the exported eye-tracking data is provided in order to connect the experimental protocol with the later behavioral and modeling analyses.

4.1 Participants

A total of 30 participants took part in the main experiment. Participants were recruited from the student population of the University of Pavia. One participant was excluded during Stage 1 data quality filtering, resulting in 29 valid datasets for subsequent

analysis.

All participants had normal or corrected-to-normal vision and were familiar with standard computer-based testing environments. Participation was voluntary, and informed consent was obtained prior to the experiment. The informed consent procedure was documented in two identical copies signed by both the participant and the researcher. One copy, containing the study information and the researcher's contact details, was retained by the participant for personal records.

Participants completed the task individually in a controlled setting intended to minimize external distractions and support stable eye-tracking acquisition [1].

Basic demographic information available for the participant sample is summarized below:

- age: range 19–59 years (mean = 28.8, SD = 8.8)
- gender representation: 66.7% male ($n = 20$) and 33.3% female ($n = 10$)
- population: university students from quantitatively oriented academic backgrounds

These characteristics indicate a relatively homogeneous sample in terms of educational background and familiarity with analytical problem-solving tasks, which is appropriate for the type of stimuli used in the experiment.

These demographic variables were not explicitly modeled in the analysis, but they provide useful contextual information for interpreting the observed behavioral patterns.

4.2 Stimuli

The experimental material consisted of 15 multiple-choice mathematical questions. The questions were selected for this experiment in order to require analytical reasoning rather than simple recall while remaining compatible with a controlled digital presentation format.

The item-selection process drew directly on mathematical multiple-choice questions reviewed from SAT-, GRE-, and GMAT-type preparation materials and item collections include references. During the preparation phase, a relatively large set of candidate questions from these sources was examined in order to identify items suitable for the aims of the present study.

More specifically, the screening of candidate questions emphasized several methodological constraints that were important for this experiment. Candidate items were required to be compatible with a text-based digital interface, to avoid dependence on diagrams, figures, or tables, and to be solvable without calculators or external tools. In addition, the wording and length of the questions had to remain manageable for on-screen reading under both timed and non-timed conditions, meaning that excessively short items and overly long items were avoided. A further consideration was structural comparability: the selected questions were intended to follow a broadly similar multiple-choice reasoning format so that differences in gaze behavior would not be dominated by large variations in item presentation style.

Based on this review, an intermediate pool of approximately 20 candidate questions was selected for further screening. The final experimental set of 15 items was drawn from this pool. Thus, the final item set should be understood as a curated selection of mathematical reasoning questions taken from SAT-, GRE-, and GMAT-style sources and chosen for methodological compatibility with the eye-tracking experiment, rather than as a set of questions written from scratch for the study.

Each item included four answer options displayed simultaneously on the screen (Figure 4.1). For the "standard" conditions, the selected questions were presented in their retained multiple-choice structure. For the condition that we will call **Timer-No-Correct**, the correct answer option was intentionally removed and replaced with a newly constructed alternative designed to remain coherent with the style and difficulty of the remaining options while ensuring that no correct answer was available in that condition.

The questions were formatted to maintain a consistent layout, enabling reliable mapping between gaze coordinates and predefined Areas of Interest (AOIs) [1]. The design of the items therefore served not only a content-related purpose, but also a methodological one, since later stages of the analysis depend on stable spatial correspondence between gaze samples and task-relevant interface regions.

The main properties of the stimuli are summarized below:

- source of the questions: selected from SAT-, GRE-, and GMAT-style mathematical reasoning question sets reviewed for this study
- selection constraints: text-based, calculator-free, diagram-free, table-free, and

structurally comparable items suitable for controlled digital presentation

- intended difficulty level: undergraduate-level mathematical reasoning
- condition-specific modification: in the **Timer-No-Correct** condition, the original correct option was removed and replaced with a newly constructed distractor
- layout consistency: identical visual layout across all 15 questions
- item order: randomized across participants through condition assignment and trial-sequence randomization; answer options randomized per trial

4.2.1 Question Validation and Difficulty Screening

Before the main data collection, a separate validation group consisting of ten volunteers reviewed the intermediate pool of 20 candidate questions in order to assess their perceived difficulty. In this preliminary screening stage, the candidate items were rated using three qualitative categories: *easy*, *medium*, and *hard*.

Out of these 20 candidate questions, exactly 15 were consistently judged to have *medium* difficulty and were thus retained for the final experimental set. This screening procedure was adopted to reduce excessive variance caused by items that were either too easy or too difficult and to limit the risk of floor or ceiling effects. More broadly, selecting items with comparable perceived difficulty was intended to encourage relatively stable task engagement across trials while preserving the analytical character of the assessment.

Accordingly, although no separate formal pilot study of the stimuli was conducted beyond this screening step, the final item set was not selected arbitrarily. Rather, it was constructed from a screened pool of candidate SAT-, GRE-, and GMAT-style questions that had undergone an initial difficulty-validation phase prior to the main experiment.

4.3 Experimental Conditions

The experiment was designed to compare response behavior under different assessment conditions. In particular, the study included non-timed and timed settings, as well as a timed condition in which no correct answer was available.

At the level of trial labeling and subsequent behavioral analysis, the dataset is organized into three experimental conditions:

- **No-Timer**: participants answered without time pressure and one option was correct;
- **Timer-Correct**: participants answered under time pressure and one option was correct;
- **Timer-No-Correct**: participants answered under time pressure and no correct option was provided.

These three conditions were implemented through a two-part protocol:

- **Part 1 (No-Timer):** 5 questions without time constraints;
- **Part 2 (Timer-Constrained):** 10 questions completed under time constraints, consisting of 5 **Timer-Correct** items and 5 **Timer-No-Correct** items.

Accordingly, the timed portion of the assessment was constrained at the individual-item level rather than at the part level. Specifically, each question in Part 2 had a maximum time limit of 90 seconds. A visible countdown timer was presented for each item and reset at the beginning of every trial (as shown in Figure 4.1).

A central manipulation of the experiment concerned the **Timer-No-Correct** condition. In this condition, the correct answer was intentionally removed and replaced, so that no objectively correct option was available. Importantly, participants were not informed that some timed items lacked a correct answer. From the participant's perspective, all trials appeared to be standard multiple-choice questions. This design was intended to create a situation in which a participant could inspect the item and options in good faith, yet still be forced to make a final selection under time pressure despite the absence of a valid solution. In methodological terms, this condition was designed to elicit a response regime closer to low-certainty or forced commitment than ordinary accuracy-driven answering.

4.3.1 Randomization Strategy

Although the experimental conditions are conceptually grouped into two parts for descriptive clarity, the actual presentation order of trials was not fixed. Instead, a randomized assignment procedure was implemented in order to reduce ordering effects and potential learning or fatigue biases.

More specifically, for each participant:

- the assignment of questions to experimental conditions was randomized, meaning that a given question could appear under different conditions (e.g., No-Timer, Timer-Correct, or Timer-No-Correct) across participants;
- the overall sequence of trials was randomized, so that the order in which conditions appeared was not fixed and could vary between participants;
- within each trial, the spatial ordering of the answer options was randomized, ensuring that the position of each option differed across participants and trials.

This design ensures that the dataset does not encode systematic positional or ordering biases at either the question level or the answer-option level. In particular, it prevents participants from relying on positional heuristics (e.g., consistently selecting a specific option location) and reduces confounding effects related to fixed condition ordering.

The use of randomization at multiple levels (question assignment, trial sequence, and option ordering) was therefore an important component of the experimental design, contributing to the internal validity of the collected data and to the robustness of subsequent behavioral analyses.

The implemented experimental design is summarized as follows:

- number of items per condition: 5 (**No-Timer**), 5 (**Timer-Correct**), 5 (**Timer-No-Correct**)
- time constraint in timed conditions: 90 seconds per question
- condition order: randomized across participants
- trial sequence across participants: randomized

- answer option order: randomized per trial

These conditions were later used in the behavioral pipeline to define condition-aware timing thresholds and interpret differences in post-engagement commitment behavior.

4.4 Experimental Procedure

Each participant completed the experiment individually in a controlled environment (the Computer Vision and Multimedia Laboratory of the Department of Electrical, Computer and Biomedical Engineering of the University of Pavia).

At the beginning of each session, the participant, who was seated in front of the experimental setup, received a short verbal introduction to the task. The participant was informed that they would answer multiple-choice mathematical questions presented on the computer screen while their eye movements were being recorded.

Before the task began, a calibration procedure was performed using the eye-tracking device to ensure accurate gaze recording. If calibration quality was not satisfactory, the procedure was repeated until acceptable tracking accuracy was obtained. A standard 9-point calibration procedure was used, consistent with common experimental practice [1].

Participants were instructed to solve each question as accurately as possible and to interact with the interface using the mouse. They selected an answer by clicking on one of the available options and then explicitly confirmed the selection using the `Submit` button. Participants were not informed that some items in the timed portion could contain no correct answer; instead, all trials were presented as standard multiple-choice tasks in order to preserve the intended behavioral manipulation.

Each trial followed a fixed sequence:

1. presentation of the question and answer options;
2. visual inspection of the question and alternatives;
3. answer selection via mouse click;
4. confirmation through the `Submit` button;
5. transition to the next trial.

In timed conditions, a visible countdown timer was displayed for each item and reset at the start of every trial (see Figure 4.1). In non-timed conditions, the timer was absent. This manipulation allowed the study to compare gaze behavior and response commitment under different temporal constraints.

The total duration of the experimental session varied per participant, ranging from approximately 20 to 30 minutes. This included the informed consent procedure, setup, calibration, and task execution.

Additional general procedural details for the main experiment are summarized below:

- calibration: standard 9-point eye-tracker calibration
- recalibration: repeated if initial calibration quality was insufficient
- practice trials: no separate formal practice block was included
- breaks: the session was conducted continuously without scheduled breaks

4.4.1 Pilot Testing

In addition to the 30 participants included in the main experiment, three participants took part in a preliminary pilot phase. This pilot was not intended for substantive behavioral analysis; rather, it was conducted to verify the technical stability of the experimental setup before the main data collection.

The pilot phase was used to confirm the correct functioning of the custom software interface, the stability of eye-tracking acquisition, the integrity of data storage, and the reliability of calibration under the intended recording conditions. It also served to verify that the system produced the expected per-trial data exports and that response-event logs remained synchronized with the gaze recordings.

More specifically, the pilot confirmed that each experimental session generated the expected set of trial-level CSV files, that the recorded gaze streams were continuous and interpretable, that gaze coordinates remained within the monitor bounds under normal operation, and that the temporal alignment between gaze data and response logs was sufficiently stable for later preprocessing and analysis.

4.5 Assessment Interface and AOI Structure

An important methodological component of this study was the design and implementation of the computer-based testing interface used during the eye-tracking experiment. The assessment environment was not treated merely as a display system, but as a controlled interaction space enabling reliable measurement of gaze behavior.

Because later stages of the analysis rely on identifying gaze behavior within specific screen regions—particularly the answer options—the spatial layout of the interface was designed to ensure clear and stable AOI boundaries across trials [1].

4.5.1 Interface Design Process

The interface was implemented specifically for this study and refined iteratively before the final data collection. Multiple layout configurations were considered and adjusted in order to ensure both usability for participants and methodological suitability for gaze-based analysis.

Two primary requirements guided the design:

- the interface had to be intuitive and similar to standard digital assessment systems;
- the answer regions had to be spatially stable and clearly separable to ensure reliable AOI mapping.

These design constraints were especially important because the later behavioral analysis depends on deterministic mapping between gaze coordinates and semantic screen regions. The iterative refinement of the layout therefore served a dual purpose: it improved visual clarity for participants and also supported precise separation of the answer-option AOIs for downstream processing.

4.5.2 Final Response Mechanism

Participants selected an answer by clicking on one of the options and confirmed their choice using a dedicated `Submit` button.

The separation between option selection and final submission was an intentional design choice. It provides a clearer behavioral signal for analyzing response timing

and commitment, and helps reduce ambiguity in identifying the moment at which a participant finalized a decision.

4.5.3 Rationale for the Final Interaction Design

Alternative response modalities were considered during the design phase, including spoken responses and handwritten answering. However, these options were not adopted in the final system because they would have introduced additional sources of movement and temporal uncertainty during recording. In particular, spoken or handwritten responses could have interfered with stable eye-tracking acquisition and made the temporal identification of final decision commitment less precise.

For this reason, the final interaction design relied on a click-based selection mechanism followed by explicit confirmation through the `Submit` button. This configuration was judged to provide the best balance between participant usability, recording stability, and interpretability of the resulting behavioral signals.

4.5.4 Assessment Layout

Figure 4.1 illustrates the interface used during data collection.

The layout consists of:

- a question region (top);
- four answer options (center);
- a `Submit` button (right);
- a timer (in timed conditions).

This structured layout is important for ensuring consistent AOI mapping, which directly affects the reliability of gaze-based feature extraction.

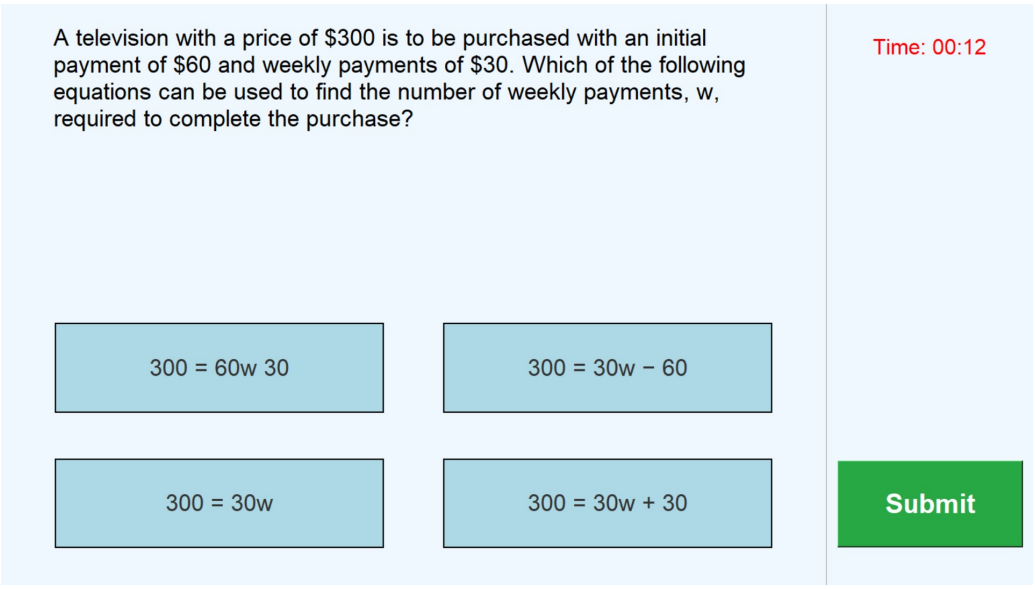

Figure 4.1: Assessment interface used during data collection. The screen includes a question region, four answer options, a timer in timed conditions, and a `Submit` button.

4.5.5 AOI Definition

For the purposes of subsequent analysis, the interface was partitioned into a set of predefined Areas of Interest (AOIs). These included the question region, four individual answer-option AOIs, the timer region, the submit region, and an "other" category for gaze points falling outside the relevant task areas.

In addition to the four individual answer-option AOIs, an aggregate `Answer_Area` can be understood as the larger screen region containing all answer boxes. However, the onset logic used later in the behavioral pipeline is based on sustained gaze within the *same answer-option AOI*, rather than merely entering the broader answer area.

AOI coordinates were defined programmatically using normalized screen-space proportions (range 0–1), ensuring consistency across participants and display configurations.

This AOI structure plays a central role in the present thesis, since later stages of the behavioral pipeline rely on stable gaze engagement with a specific answer option to define the onset of answer-related processing.

4.6 Apparatus and Data Acquisition

Eye-tracking data were recorded during task execution using a *GazePoint GP3 HD* eye tracker[1] operating at 150 Hz.

The experiment was conducted on a desktop computer with a 23-inch Full HD monitor (1920×1080 resolution). Participants were seated at an approximate viewing distance of 50–60 cm from the screen.

No head restraint (e.g., chin rest) was used, allowing natural head movement during the task.

The experiment took place in the Computer Vision and Multimedia Laboratory of the Department of Electrical, Computer and Biomedical Engineering of the University of Pavia. Environmental conditions such as lighting and distractions were not strictly controlled or quantitatively recorded.

The task was implemented using a software interface specifically developed for this thesis and responsible for stimulus presentation, response logging, and gaze data acquisition.

4.7 Exported Data Representation

Eye-tracking data were exported as trial-level gaze sequences sampled at 150 Hz. For each trial, multiple gaze- and pupil-related measurements were recorded.

The downstream modeling pipeline uses 13 feature channels:

- BPOGX
- BPOGY
- FPOGD
- FPOGX
- FPOGY
- LPCX
- LPCY

[1] https://www.gazept.com/blog/introducing-the-gazepoint-gp3-hd/?v=0d149b90e739

- LPD
- LPUPILD
- RPCX
- RPCY
- RPD
- RPUPILD

These variables include binocular point-of-gaze coordinates, fixation-related coordinates and duration, left and right pupil-center coordinates, and pupil-diameter measurements. They were extracted from the raw recordings and later used for behavioral processing and supervised modeling.

A more detailed description of the modeling-oriented tensor construction, preprocessing choices, and architecture-specific inputs is deferred to Chapter 7, in order to keep the present chapter focused on experimental design and data acquisition.

4.8 Dataset Summary for Subsequent Analysis

After data collection, the recordings were organized at the trial level for subsequent preprocessing and analysis. The full raw dataset contains 30 participants answering 15 questions each, for a total of 450 trials prior to filtering.

Later stages of the thesis apply quality control, AOI-based processing, behavioral labeling, and configuration-specific filtering. As a result, the number of trials available for downstream modeling differs from the total number of collected trials. In particular, the behavioral pipeline yields 385 valid labeled trials, while specific deep learning configurations may use smaller subsets depending on AOI filtering, temporal windowing, and condition selection.

This distinction is important for avoiding confusion between:

- the total number of trials collected experimentally;
- the number of trials retained after behavioral preprocessing;
- the number of trials used in a specific modeling configuration.

4.9 Chapter Summary

This chapter described the experimental framework used to collect the eye-tracking data analyzed in this thesis. It documented the participants, stimuli, experimental conditions, task procedure, interface structure, AOI design, and data acquisition setup.

By making the data collection process explicit, the chapter provides the methodological foundation required to interpret the behavioral pipeline introduced in Chapter 5. Details of the deep learning architectures, training strategy, class imbalance handling, and evaluation protocol are presented separately in Chapter 7, where they can be discussed in direct relation to the modeling experiments.

Chapter 5

Behavioral Data Processing Pipeline

This chapter presents a fully deterministic and reproducible behavioral processing pipeline that transforms raw eye-tracking recordings into trial-level behavioral labels. The sole objective of the pipeline is to classify each trial as **RANDOM**, **NOT_RANDOM**, or **INVALID**. Importantly, the labeling procedure does not use observed answer correctness as an input variable. Rather than modeling performance accuracy, the framework evaluates whether the observed response process reflects rapid, low-engagement commitment or sustained, structured interaction with the answer alternatives.

The pipeline is intentionally rule-based. No statistical learning, adaptive fitting, or model-dependent optimization procedures are involved in label generation. All operations are deterministic functions of recorded gaze signals, response timing, and predefined configuration parameters. This design choice supports interpretability, traceability, and reproducibility throughout the transformation from raw trial recordings to final behavioral labels.

5.1 Formal Problem Definition

Let a trial i correspond to one participant answering one item. For each trial, the eye tracker produces a time-ordered sequence of gaze samples:

$$G_i = \{(t_k, x_k, y_k, v_k)\}_{k=1}^{N_i},$$

where:

- t_k is the timestamp,
- (x_k, y_k) are normalized screen coordinates,
- v_k contains device-level validity indicators.

In addition to the gaze stream, each trial includes response-event information, in particular the final trial end time t_{end}. The goal of the pipeline is to compute a deterministic mapping:

$$f : (G_i, t_{\text{end}}) \rightarrow y_i,$$

where $y_i \in \{\text{RANDOM}, \text{NOT\_RANDOM}, \text{INVALID}\}$. The mapping f is implemented through three sequential stages described below.

5.2 Stage 1: Reliability-Constrained Quality Filtering

5.2.1 Rationale

Eye-tracking data are inherently noisy due to tracking loss, blinks, head movement, and acquisition artifacts. If unreliability is not explicitly controlled, downstream temporal markers may become unstable and compromise interpretability. Stage 1 therefore enforces conservative reliability constraints at both trial and participant levels to ensure that subsequent behavioral indicators are derived only from sufficiently stable gaze streams.

5.2.2 Sample-Level Validity

For each sample k, a Boolean validity indicator is defined as:

$$\texttt{is\_valid}_k = (\texttt{BPOGV}_k = 1) \vee (\texttt{BKID}_k \neq 0).$$

Here, `BPOGV` denotes the device-provided *best point-of-gaze validity flag*, and `BKID` denotes a *blink indicator*. This rule retains:

- valid gaze samples (`BPOGV` = 1),
- blink-marked samples, which are treated as structured physiological events rather than as purely arbitrary missing observations.

Blink-marked samples are retained for reliability accounting at the trial level. However, they do not positively contribute to AOI-based stable-engagement detection unless usable gaze coordinates are available. In this way, the pipeline preserves temporal integrity without allowing blinks to artificially strengthen evidence for stable AOI engagement.

5.2.3 Trial-Level Reliability Constraint

For trial i with N_i samples:

$$p_{\text{invalid}}^{(i)} = \frac{N_{\text{invalid}}^{(i)}}{N_i}.$$

A trial is excluded if:

$$p_{\text{invalid}}^{(i)} > \tau_{\text{trial}},$$

with $\tau_{\text{trial}} = 0.30$.

The selected threshold reflects a conservative reliability criterion intended to remove substantially corrupted trials while preserving the majority of recorded data.

5.2.4 Participant-Level Reliability Constraint

For participant j with T_j trials:

$$p_{\text{excluded}}^{(j)} = \frac{T_{\text{excluded}}^{(j)}}{T_j}.$$

A participant is excluded if:

$$p_{\text{excluded}}^{(j)} > \tau_{\text{participant}},$$

with $\tau_{\text{participant}} = 0.50$.

This hierarchical filtering rule prevents subject-level noise patterns from disproportionately affecting the analysis and helps ensure that later stages operate on interpretable gaze data.

5.2.5 Experimental Summary

Table 5.1: Stage 1 reliability filtering summary.

Stage 1 Summary	Value
Trial invalid threshold τ_{trial}	0.30
Participant exclusion threshold $\tau_{\text{participant}}$	0.50
Total trials evaluated	450
Kept trials	431
Excluded trials	19
Total participants evaluated	30
Kept participants	29
Excluded participants	1

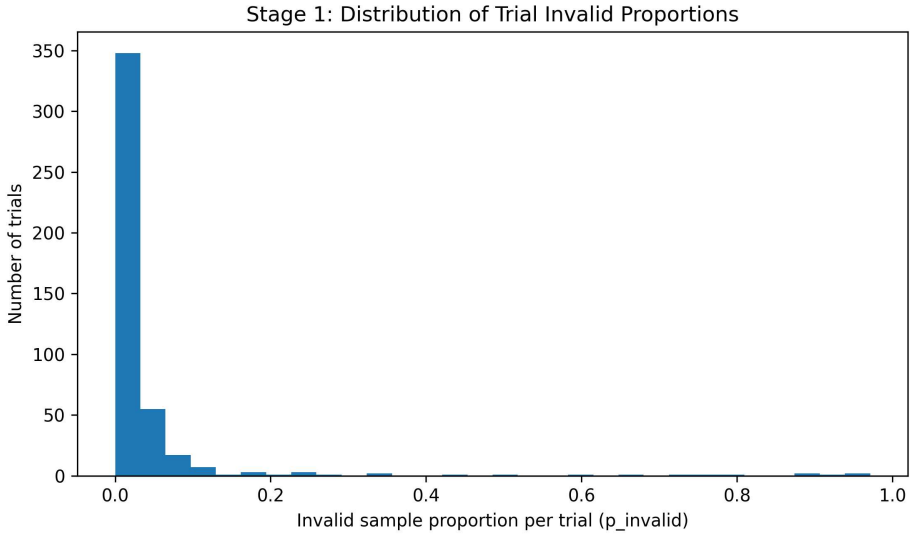

Figure 5.1: Distribution of trial-level invalid sample proportions.

Overall, Stage 1 ensures that all downstream behavioral labels are computed only from trials with acceptable signal reliability, thereby establishing the experimental foundation for the subsequent onset-detection and labeling stages.

5.3 Stage 2: AOI Mapping and Stable Engagement Onset

5.3.1 Deterministic AOI Mapping

Each gaze sample is deterministically mapped to a semantic Area of Interest (AOI):

$$\mathrm{AOI}_k = g(x_k, y_k),$$

where $g(\cdot)$ partitions the interface into **Question**, four individual **Answer-Option AOIs**, **Timer**, **Submit**, and **Other**.

Because the interface layout is fixed, AOI geometry can be specified explicitly in configuration files. This guarantees reproducible mapping between raw gaze coordinates and functional screen regions across all trials and runs.

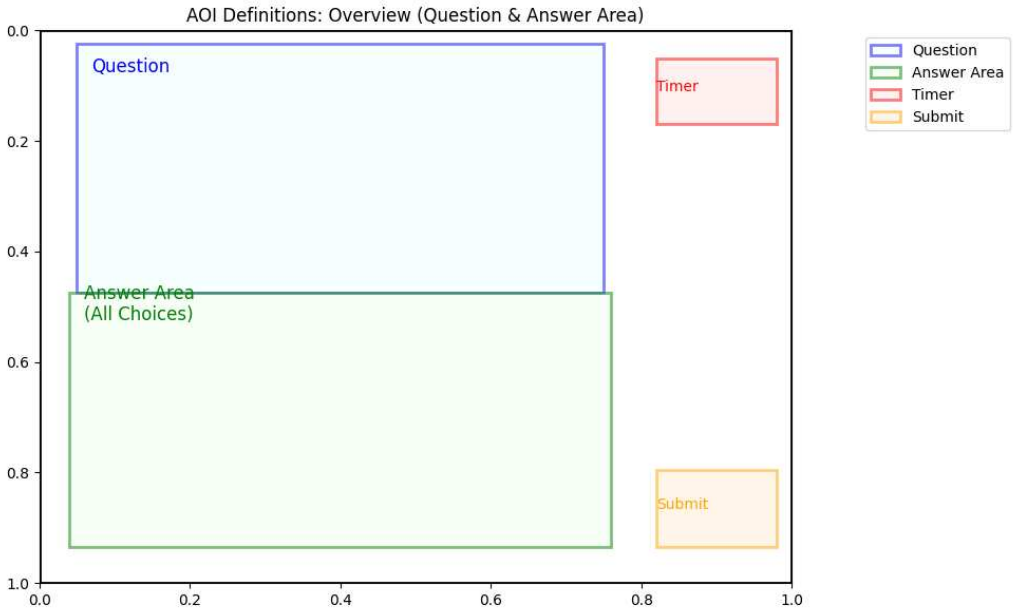

Figure 5.2: Overview of the AOI layout. The figure illustrates the aggregate spatial organization of the main screen regions, including the broader answer region. In the implemented onset-detection logic, however, stable engagement is defined using sustained gaze within the *same answer-option AOI*, not merely within the aggregate answer area.

The overview shown in Figure 5.2 is intended to clarify the global interface geometry rather than to depict the full granularity of the onset rule. In the implementation, the answer region is subdivided into four specific answer-option AOIs, and the onset marker is based on stability within one of these individual regions.

5.3.2 Stable Answer-Option Onset

The onset of stable engagement with the answer options is denoted by t_a.

This onset marker is intended to represent the transition from earlier processing phases (e.g., question reading or broad visual exploration) to focused inspection of a specific answer option. To reduce sensitivity to brief gaze fluctuations or isolated noisy samples, onset is not defined by a single instantaneous fixation but by sustained AOI coverage over a temporal window.

Let:

- f_s be the sampling rate (Hz),
- W be the stability window length (ms),
- θ be the coverage threshold.

In the reported implementation,

$$f_s = 150 \text{ Hz}, \qquad W = 1000 \text{ ms}, \qquad \theta = 0.85.$$

The window size in samples is therefore:

$$M = \left\lfloor \frac{W \cdot f_s}{1000} \right\rfloor.$$

Coverage within a candidate window is defined as:

$$c = \frac{\#\{\text{valid samples in the window assigned to the same answer-option AOI}\}}{M}.$$

A window is considered stable if:

$$c \geq \theta.$$

The earliest window satisfying this criterion defines:

$$t_a = t_{k_{\text{start}}}.$$

This definition is intentionally conservative: it requires not merely entering the broader answer area, but sustained engagement with the same answer option. Informal

inspection suggested that moderate perturbations of W and θ did not substantially alter the main ordering of detected onset times across trials, although no separate formal sensitivity study is reported here.

5.3.3 Diagnostic Indicators for Missing t_a

Not all Stage 1-retained trials yield a detectable t_a. For such cases, the pipeline does not merely record onset failure as a binary outcome; it also supports diagnostic inspection of the underlying gaze trace in order to better characterize why the stability criterion was not satisfied.

In particular, three diagnostic indicators are relevant for trials in which t_a is not detected:

- **Off-screen percentage:** the proportion of samples falling outside the effective screen or task-relevant gaze area, which helps identify trials dominated by tracking loss, off-screen gaze, or unstable viewing behavior.
- **Longest consecutive answer-option sequence:** the longest uninterrupted run of gaze samples assigned to an answer-related region, which indicates whether the participant approached sustained answer inspection without meeting the full onset criterion.
- **Best achieved window coverage:** the maximum observed coverage within any candidate window, even when that value remains below the required threshold $\theta = 0.85$. This quantity is informative for distinguishing near-miss cases from trials showing little or no stable answer-focused engagement.

These diagnostic quantities do not alter the deterministic labeling rule itself. Rather, they provide an additional interpretive layer for understanding onset-detection failures. In methodological terms, this helps distinguish between qualitatively different sources of missing t_a, such as fragmented answer inspection, substantial off-screen behavior, or windows that approached but did not reach the required stability threshold.

Table 5.2: Summary statistics of detected onset times (in seconds).

Statistic	Value
Count	387
Mean (s)	52.75
Std (s)	28.77
Min (s)	4.90
25th percentile (s)	31.91
Median (s)	48.67
75th percentile (s)	71.04
Max (s)	228.20

Table 5.2 reports the distribution of the detected onset times t_a. A total of 387 valid trials were included in this analysis.

The mean onset time is 52.75 seconds (SD = 28.77), indicating substantial variability in the duration of the pre-engagement phase across trials. The median value (48.67 seconds) is slightly lower than the mean, suggesting a moderately right-skewed distribution.

The interquartile range spans from 31.91 to 71.04 seconds, showing that the middle 50% of trials exhibit considerable dispersion in the timing of stable engagement with a specific answer option. This variability reflects differences in reading duration, initial exploration behavior, and individual response strategies.

Minimum and maximum values (4.90 s and 228.20 s, respectively) further indicate that some trials involve very rapid transitions to answer-focused processing, while others include extended pre-engagement phases. Because these onset statistics pool both timed and non-timed trials, larger values are expected in non-timed trials and do not contradict the timed structure of the constrained part of the experiment.

Overall, these results highlight that the onset of stable answer engagement is not temporally uniform across trials, supporting the need for a trial-specific, gaze-based definition of engagement rather than relying on fixed time thresholds.

5.4 Stage 3: Condition-Aware Randomness Labeling

5.4.1 Post-Engagement Commitment Time

After stable answer engagement has been detected, post-engagement commitment time is defined as:

$$t_{\text{answer}} = t_{\text{end}} - t_a.$$

This measure isolates the duration between stable visual engagement with a specific answer option and final response submission. It therefore serves as a more behaviorally grounded quantity than total response time, which conflates reading, inspection, and commitment phases.

5.4.2 Validity Constraints

A trial is labeled **INVALID** if:

- t_a is undefined,
- the trial was unanswered.

These criteria ensure that final labels are assigned only when the response process contains both a recoverable engagement onset and an observable response outcome.

5.4.3 Condition-Specific Thresholding

Trials belong to one of three experimental conditions: **No-Timer**, **Timer-Correct**, or **Timer-No-Correct**.

Trials in the **Timer-No-Correct** condition are operationally assigned the label **RANDOM** within the experimental design. Under this condition, time pressure is combined with the absence of a valid correct answer, and the resulting trials are treated as representing an intended low-commitment or forced-response regime rather than a gaze-only inferred class. Because participants were not informed that some items in this condition contained no correct answer, the manipulation was intended to capture situations in which a participant may inspect the available options but still be forced to finalize a choice under uncertainty.

For **No-Timer** and **Timer-Correct** conditions, randomness labeling is based on a within-condition percentile rule:

$$\text{RANDOM if } t_{\text{answer}} < Q_{25}^{(\text{condition})}.$$

That is, trials with unusually short post-engagement commitment times relative to their own condition are labeled **RANDOM**; all others are labeled **NOT_RANDOM**.

Thresholds are computed independently within each condition in order to avoid distributional mixing across experimentally distinct temporal regimes.

5.4.4 Formal Definition of RANDOM Behavior

Based on the pipeline described above, a trial is labeled as **RANDOM** if it satisfies at least one of the following conditions:

- The trial belongs to the **Timer-No-Correct** condition and is operationally assigned the **RANDOM** label by design.
- The trial belongs to the **No-Timer** or **Timer-Correct** condition and exhibits unusually short post-engagement commitment time, formally defined as:

$$t_{\text{answer}} < Q_{25}^{(\text{condition})}.$$

Conversely, a trial is labeled as **NOT_RANDOM** if it satisfies the validity constraints and does not meet the above criteria.

This definition operationalizes random-like responding as insufficient or prematurely terminated engagement with answer-related regions, rather than as incorrect responding. The labeling rule does not use observed response correctness as an input variable; however, one experimental condition (**Timer-No-Correct**) is assigned the **RANDOM** label by design. This improves clarity of the experimental construct within the study, but it also introduces a potential confound between condition structure and behavioral label assignment.

Table 5.3: Overall label distribution.

Label	Count
NOT_RANDOM	195
RANDOM	190
INVALID	46
Total	431

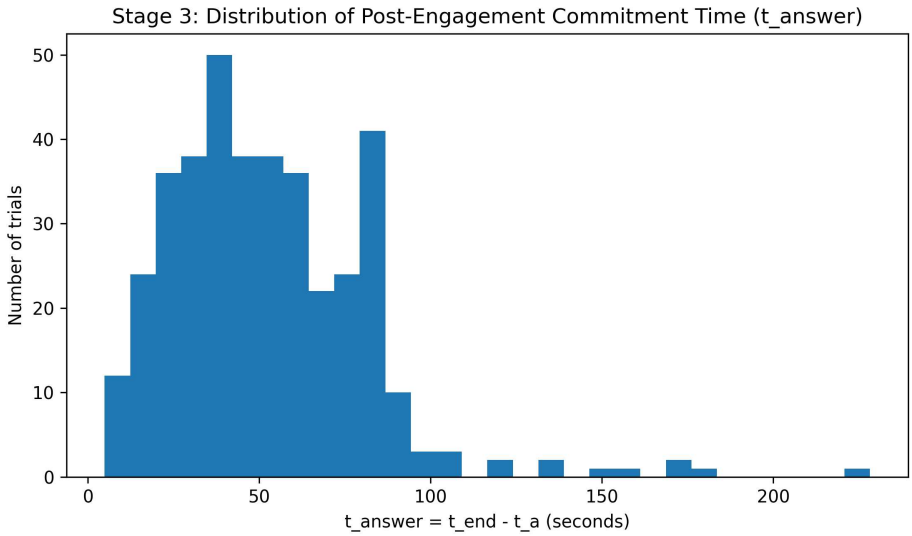

Figure 5.3: Distribution of post-engagement commitment time t_{answer}.

5.5 Determinism, Leakage Control, and Reproducibility

The pipeline is deterministic, correctness-independent in its direct inputs, and parameterized through fixed configuration files. Label construction is completed prior to any predictive modeling stage, and no supervised learning signal influences onset detection, threshold estimation, or trial exclusion.

This strict separation between preprocessing and modeling removes direct circularity between label construction and prediction. It also reduces the risk of correctness-based or model-induced leakage within the label-generation stage. However, this does not by itself eliminate all possible sources of optimistic bias in downstream modeling, such as fixed-split model selection effects or condition-related confounding.

In addition, all intermediate artifacts are exported to support traceability, inspection, and practical reproducibility of the labeling process.

5.6 Summary

This chapter introduced a transparent and operational pipeline for transforming raw eye-tracking recordings into interpretable trial-level behavioral labels. By combining reliability filtering, AOI-grounded onset detection, and condition-aware thresholding, the framework produces correctness-independent labels that are both methodologically explicit and reproducible.

These labels provide the experimental basis for the descriptive analyses reported in Chapter 6 and the supervised modeling experiments reported in Chapter 8.

Chapter 6

Results and Discussion

This chapter reports the experimental outcomes of the three-stage behavioral pipeline and presents descriptive statistics for the derived randomness labels. The results are organized according to the pipeline structure: data quality filtering (Stage 1), AOI-based onset detection (Stage 2), and randomness labeling (Stage 3).

In addition to reporting numerical outcomes, this chapter provides an interpretation of the observed behavioral patterns and evaluates the effectiveness of the proposed pipeline in producing stable, interpretable, and reproducible labels under the operational definitions adopted in this thesis.

6.1 Dataset Overview

The raw dataset consists of 30 participants and 15 items, resulting in 450 trials prior to quality filtering. Eye-tracking data were sampled at 150 Hz.

Table 6.1: Dataset summary before quality filtering.

Number of participants	30
Number of items per participant	15
Total trials	450
Sampling rate (Hz)	150
Total gaze samples	3564145

6.2 Stage 1: Quality Filtering Results

Stage 1 identifies invalid gaze samples using the tracker validity flag and a blink indicator. Specifically, a sample is retained if (i) the tracker validity flag indicates a valid sample, or (ii) the sample is marked as a blink; otherwise it is marked invalid. Importantly, the pipeline preserves temporal integrity by *not* deleting individual samples; instead, it excludes entire trials when corruption exceeds a threshold.

6.2.1 Trial-level and Participant-level Exclusion Criteria

- **Trial-level exclusion:** if invalid samples exceed 30% (equivalently, valid samples fall below 70%), the trial is excluded.
- **Participant-level exclusion:** if a participant has more than 50% of trials excluded, the participant is removed entirely (including otherwise valid trials).

6.2.2 Stage 1 Outcomes

Table 6.2: Stage 1 filtering outcomes.

Metric	Count	Percentage
Total trials	450	100.0%
Trials excluded (invalid $>$ 30%)	14	3.1%
Participants excluded ($>$ 50% bad trials)	1	–
Good trials excluded due to participant removal	5	1.1%
Total excluded trials	19	4.2%
Kept trials (passed filters)	431	95.8%

In terms of sample counts, 211901 samples were excluded indirectly through trial- and participant-level removal, leaving 3352244 samples across 431 retained trials.

6.2.3 Interpretation

Overall, Stage 1 indicates high data quality: 95.8% of trials were retained. The majority of exclusions were driven by trials with extended tracking loss (invalid proportion $>$ 30%), while a smaller proportion resulted from the participant-level exclusion rule.

Importantly, the decision to exclude entire participants when more than half of their trials are corrupted is intentionally conservative. This prevents systematic subject-level noise from biasing the behavioral signals used in later stages, thereby increasing the reliability of the derived temporal markers.

6.3 Stage 2: AOI Assignment and Stable Answer-Option Onset

Stage 2 assigns gaze samples to Areas of Interest (AOIs) and detects the onset of stable engagement with a specific answer option, denoted t_a. The onset t_a is defined as the first timestamp at which gaze remains stably within the same answer-option AOI over a 1-second window.

6.3.1 Detection Rule

A candidate window spans 1000 ms ($\approx$ 150 samples at 150 Hz). A detection is accepted if at least 85% of samples within the window are both (i) valid and (ii) assigned to the same answer-option AOI. The first accepted window defines t_a.

6.3.2 Stage 2 Outcomes

Table 6.3: Stage 2 outcomes for t_a detection on Stage 1 retained trials.

Metric	Count	Percentage
Trials evaluated (after Stage 1)	431	100.0%
Trials with detected t_a	387	89.8%
Trials without detected t_a	44	10.2%

6.3.3 Interpretation

The high detection rate of t_a (89.8%) indicates that, in the majority of trials, participants exhibit a clear transition from initial exploration to focused inspection of a specific answer alternative.

Trials without a detected onset (10.2%) likely correspond to irregular or fragmented gaze behavior, such as rapid switching between regions or insufficient stabilization within a single answer-option AOI. These cases highlight the importance of defining engagement through a stability-based criterion rather than instantaneous fixations.

In addition to the binary distinction between detected and undetected t_a, the project also incorporated diagnostic inspection of onset-failure cases. In particular, missing-t_a trials can be characterized using indicators such as off-screen gaze percentage, the longest consecutive answer-option sequence, and the best achieved window coverage below the required threshold. These diagnostics do not change the labeling rule, but they are useful for distinguishing different types of failure cases—for example, trials dominated by off-screen behavior versus trials that approached stable answer inspection without satisfying the full 85% criterion.

This additional diagnostic perspective strengthens the interpretation of missing-t_a cases, since it shows that onset failures were not treated as a single undifferentiated error category. Rather, they were examined as potentially informative deviations from the required engagement pattern.

6.4 Stage 3: Randomness Labeling

Stage 3 assigns a trial-level behavioral label (**RANDOM** vs. **NOT_RANDOM**) based on the time elapsed from stable answer inspection onset to response completion:

$$t_{\text{answer}} = t_{\text{end}} - t_a.$$

6.4.1 Condition-aware Labeling Policy

- **Timer-No-Correct:** all trials in this condition are operationally labeled **RANDOM**. Under time pressure with no correct option available, the condition is intended to approximate a forced or low-commitment response regime.
- **No-Timer and Timer-Correct:** a data-driven threshold based on the 25th percentile (P25) of t_{answer} is computed *separately per condition*. Trials with $t_{\text{answer}} <$ P25 are labeled **RANDOM**, while all others are labeled **NOT_RANDOM**.

Table 6.4: Condition-specific percentile thresholds used for Stage 3 labeling.

Condition	Threshold value
No-Timer	$Q_{25} = 34.43$ s
Timer-Correct	$Q_{25} = 25.16$ s

Table 6.4 reports the empirical condition-wise thresholds used to distinguish unusually short post-engagement commitment times in the two threshold-based conditions. The **Timer-No-Correct** condition is excluded from this table because its trials are operationally assigned the **RANDOM** label by design rather than through percentile thresholding.

6.4.2 Overall Label Distribution

Table 6.5: Overall randomness label distribution on Stage 1 retained trials.

Label	Count	Percentage
NOT_RANDOM	195	45.2%
RANDOM	190	44.1%
INVALID	46	10.7%
Total trials	431	100.0%

Invalid trials arose primarily from missing t_a detection (44 trials) and a small number of unanswered trials (3 cases). One trial satisfied both criteria (1 case), resulting in a total of 46 invalid trials.

Consequently, 385 trials remained with valid behavioral labels and were used in subsequent modeling experiments.

6.4.3 Label Distribution by Condition

Table 6.6: Label distribution per experimental condition (Stage 3).

Condition	**NOT_RANDOM**	**RANDOM**	**INVALID**
No-Timer	98	33	14
Timer-Correct	97	32	14
Timer-No-Correct	0	125	18
Total	195	190	46

6.4.4 Interpretation

The observed label distribution is consistent with the experimental design. The Timer-No-Correct condition produces exclusively **RANDOM** labels by construction, serving as a reference regime for operationally induced low-commitment responding.

In contrast, the No-Timer and Timer-Correct conditions exhibit a mixture of **RANDOM** and **NOT_RANDOM** trials, reflecting natural variability in decision behavior. The use of condition-specific percentile thresholds ensures that labeling remains sensitive to contextual differences in response timing.

The resulting label distribution is relatively balanced between the two primary classes, which is advantageous for subsequent predictive modeling.

6.5 Discussion

The results suggest that integrating gaze-grounded onset detection with condition-aware timing thresholds yields interpretable and plausibly meaningful trial-level labels under the operational definitions adopted in this thesis.

A key strength of the proposed pipeline is its strict separation between label construction and predictive modeling. By defining randomness independently of observed correctness, the framework avoids direct circularity and helps ensure that behavioral labels reflect process-level dynamics rather than outcome-based performance.

Furthermore, the multi-stage structure—combining reliability filtering, AOI-based temporal segmentation, and condition-aware thresholding—provides a transparent and reproducible pathway from raw gaze data to final labels. The additional use of diagnos-

tic inspection for missing-t_a trials further strengthens this interpretation by showing that onset-detection failures were examined in a structured way rather than treated as opaque residual cases.

The resulting dataset contains 385 valid labeled trials (excluding invalid cases), which form the basis for the modeling framework described in Chapter 7 and the supervised learning experiments reported in Chapter 8.

6.6 Summary

This chapter presented the experimental results of the behavioral processing pipeline and demonstrated its ability to produce stable, interpretable, and reproducible labels from eye-tracking data under the adopted methodological assumptions. These findings support the methodological choices introduced in previous chapters and establish the empirical basis for the predictive modeling analysis reported separately in Chapter 8.

Chapter 7

Deep Learning Modeling Framework

This chapter presents the predictive modeling framework used to classify trial-level responding behavior as **RANDOM** or **NOT_RANDOM** based on eye-tracking signals. In contrast to the deterministic behavioral pipeline described in Chapter 5, the objective here is to evaluate whether data-driven models can learn discriminative temporal patterns from gaze-derived sequences under a controlled and reproducible evaluation setting [27, 26].

A core design constraint is that the modeling stage must remain faithful to the previously defined behavioral construct. The labels are deterministically computed prior to modeling and remain fixed throughout experimentation. The modeling phase therefore assesses *predictive learnability* of the constructed behavioral labels.

7.1 Problem Formulation

Each trial i is represented by:

$$X_i \in \mathbb{R}^{C \times T}.$$

The label is:

$$y_i \in \{0, 1\}.$$

In the binary formulation adopted in this thesis, $y_i = 1$ denotes **RANDOM** and $y_i = 0$ denotes **NOT_RANDOM**.

The objective is to learn:

$$\hat{y}_i = h(X_i; \theta),$$

where $h(\cdot)$ denotes a parameterized classifier and θ denotes the learnable model parameters.

7.2 Input Representation

Each trial is represented using 13 eye-tracking feature channels sampled at 150 Hz. The deep learning experiments do not rely on a single fixed dataset view; rather, multiple input configurations are explored later in Chapter 8, including different trial subsets, AOI filters, and temporal windows.

The main reported sequence-model setting uses signals restricted to the answer-related portion of the interface and a late-trial temporal window intended to emphasize answer inspection and commitment behavior. This emphasis is also behaviorally motivated by the experimental design described in Chapter 4. In particular, in the **Timer-No-Correct** condition, participants were not informed that some items had no correct answer. Under such trials, a participant may inspect the options for some time but still be forced to finalize a choice under uncertainty, especially near the end of the allowed response interval. For this reason, late-trial windows were considered especially relevant candidates for modeling, since random or forced final commitment is plausibly expressed most clearly in the final seconds before response submission.

At the same time, the modeling framework does not assume in advance that a specific late window is optimal. Instead, different temporal windows were explored empirically in order to test which segment of the gaze sequence provides the strongest predictive signal. The final reported setting reflects the best-performing configuration among the tested alternatives rather than a fixed theoretical assumption imposed before experimentation.

In that setting, the model input is represented with a fixed sequence length:

$$T = 300.$$

This corresponds to the final tensor length used by the model after configuration-

specific truncation and padding. Although the raw time window in the best reported setting spans the last 9 seconds of each trial, which at 150 Hz would contain approximately 1350 samples before filtering, truncation, and padding, the final tensor presented to the network is fixed at 300 time steps.

The main reported configuration is summarized below:

- AOI filtering: answer-option-focused input setting
- Time window: last 9 seconds
- Sampling rate: 150 Hz
- Final tensor length: $T = 300$
- Train/test split: 80/20 (stratified)
- Random seed: 42

This representation was chosen to preserve temporal structure while emphasizing the portion of the signal most directly related to answer inspection and commitment. Other AOI settings, temporal windows, and dataset views are explored systematically in Chapter 8.

7.3 Model Architectures

This section defines the neural architectures considered in the thesis. The configurations reported below define the reference architectural scales and design choices used in the experiments, while the selection of best-performing settings is discussed in Chapter 8. The comparative performance of these models is reported separately in Chapter 8.

7.3.1 Bidirectional LSTM

The LSTM model captures temporal dependencies in gaze sequences using a multi-layer bidirectional architecture. A schematic representation is shown in Figure 7.1.

$$\text{Input} \rightarrow \text{BiLSTM} \rightarrow \text{Pooling (mean + max)} \rightarrow \text{FC} \rightarrow \text{ReLU} \rightarrow \text{FC}$$

The implemented LSTM architecture uses the following reference configuration, which defines the architectural scale used in this study; model selection and performance comparisons are reported in Chapter 8.

- Hidden size: 384
- Layers: 3
- Bidirectional: Yes
- Dropout: 0.15
- FC dimension: 192

$$\text{Output dim} = \text{hidden} \times 2 \times 2 = 1536.$$

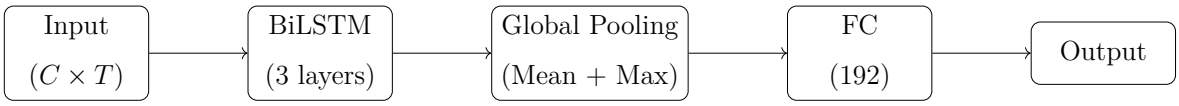

Figure 7.1: Architecture of the bidirectional LSTM model.

The architectures described in this section provide the formal modeling definitions used throughout the thesis. To avoid redundancy, these models are not redefined in Chapter 8; instead, that chapter refers back to the present section when discussing experimental results.

7.3.2 1D CNN

The CNN model extracts local temporal patterns through hierarchical convolutional layers. The architecture is illustrated in Figure 7.2.

$$\text{Input} \rightarrow \text{Conv1D} \times 4 \rightarrow \text{Pooling} \rightarrow \text{FC} \rightarrow \text{Output}$$

The implemented CNN architecture uses the following configuration as a reference design for the experiments:

- Channels: [96, 192, 384, 384]

- Kernel sizes: $7 \rightarrow 5 \rightarrow 3$
- Dropout: 0.15
- FC dimension: 192

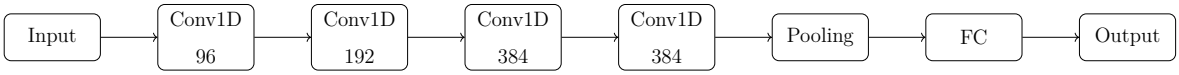

Figure 7.2: 1D CNN architecture.

7.3.3 Hybrid CNN–LSTM

This architecture combines local feature extraction and temporal modeling (Figure 7.3).

$$\text{Input} \rightarrow \text{CNN} \rightarrow \text{BiLSTM} \rightarrow \text{Pooling} \rightarrow \text{FC}$$

The implemented hybrid architecture uses the following configuration:

- CNN channels: 128
- LSTM hidden size: 256
- Layers: 2
- Dropout: 0.15

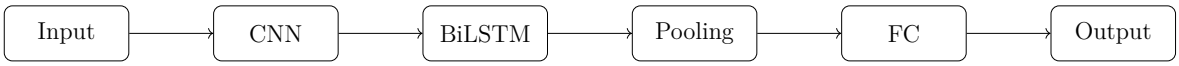

Figure 7.3: Hybrid CNN–LSTM architecture.

7.3.4 Transformer Encoder

The Transformer model uses self-attention to capture global dependencies. See Figure 7.4.

$$\text{Input} \rightarrow \text{Projection} \rightarrow \text{Positional Encoding} \rightarrow \text{Transformer} \rightarrow \text{Pooling} \rightarrow \text{FC}$$

The implemented Transformer architecture uses the following configuration:

- d_{model}: 192
- Heads: 12
- Layers: 5
- Feedforward: 768
- Dropout: 0.15
- Activation: GELU

Figure 7.4: Transformer encoder architecture.

7.3.5 MLP

The MLP baseline operates on flattened fixed-length sequence representations (Figure 7.5).

The implemented MLP architecture uses the following configuration:

- Layers: [256, 128, 64, 32]
- Dropout: 0.15
- BatchNorm: Yes

Figure 7.5: MLP architecture.

7.4 Training Configuration

All neural models were trained using the Adam optimizer. Final model-family-specific settings are summarized in Chapter 8; the common training configuration adopted across the reported experiments is summarized below.

- Optimizer: Adam

- Learning rate: 0.0004 (LSTM/Hybrid/Transformer/MLP), 0.0008 (CNN)
- Batch size: 16
- Weight decay: 5×10^{-5}
- Maximum epochs in the reported runs: 400
- Early stopping patience in the reported best configurations: 30

7.4.1 Loss Functions

- Sequence models: BCEWithLogitsLoss
- MLP: CrossEntropyLoss

The sequence models were implemented in a binary-output setting, making `BCEWithLogitsLoss` a natural choice. The MLP baseline, by contrast, was implemented with a two-logit output layer and therefore used `CrossEntropyLoss` in a standard vector-classification formulation.

7.4.2 Regularization

- Dropout: 0.15
- Gradient clipping: max_norm = 1.0
- LR scheduler: ReduceLROnPlateau

7.5 Reproducibility

- Random seed: 42
- Deterministic CUDA settings enabled where supported
- Configuration snapshots stored
- Experiment logs saved

7.6 Summary

This chapter presented a reproducible deep learning framework integrating multiple architectures and systematic experimentation. The modeling setup was designed to remain aligned with the behavioral interpretation developed in earlier chapters, including the expectation that final commitment behavior may be especially informative in trials involving time pressure and unresolved answer uncertainty. Chapter 8 then evaluates these models under different input configurations and reports their comparative predictive performance.

Chapter 8

Deep Learning Results

This chapter presents the experimental results obtained from deep learning models trained to classify trial-level behavioral labels (**RANDOM** vs. **NOT_RANDOM**) based on eye-tracking sequences.

The objective of this stage is to evaluate whether temporal gaze dynamics contain predictive signals that allow automatic discrimination between engagement-driven and random-like responding behavior. In addition to reporting final performance, the chapter also examines how model architecture, input construction, and hyperparameter choices affect predictive performance in a small-data setting.

8.1 Dataset for Modeling

After applying the behavioral pipeline described in Chapter 5, trials labeled as **INVALID** were excluded from modeling. At the level of the behavioral pipeline, this produced 385 valid labeled trials. However, the deep learning experiments were not restricted to a single dataset view. Instead, multiple modeling subsets were constructed under different experimental configurations, including AOI filtering, trial-part selection, and time-window extraction.

Accordingly, the number of trials used for modeling depends on the specific input configuration. The final dataset corresponding to the best reported deep learning configuration contained:

- 253 labeled trials
- 13 eye-tracking feature channels

- sequence length of 300 time steps (in the best reported configuration)
- sampling rate of 150 Hz

Class distribution in this best-performing modeling subset was:

- **NOT_RANDOM**: 97 trials
- **RANDOM**: 156 trials

Each trial is represented as a tensor:

$$X \in \mathbb{R}^{C \times T},$$

where $C = 13$ feature channels and $T = 300$ time steps.

The dataset was divided using a stratified 80/20 split:

- Training set: 202 trials (class0 = 77, class1 = 125)
- Test set: 51 trials (class0 = 20, class1 = 31)

Thus, while the behavioral pipeline provided the full pool of valid labeled trials, the deep learning stage further explored multiple filtered and windowed subsets in order to identify the most informative input configuration for each model family. Accordingly, the best reported performance reflects a configuration-specific subset rather than the full pool of valid labeled trials.

8.2 Input-Configuration Search Space (Filters and Time Windows)

A key part of this study was not only training multiple architectures, but also systematically exploring how *input construction choices* affect performance. In particular, for each model family, multiple dataset views were generated by varying the following factors:

- **Trial parts:** different subsets of trials were used for modeling, including (i) timer trials, (ii) No-Timer combined with Timer-No-Correct, and (iii) all parts.

- **Area of Interest (AOI):** sequences were extracted either from answer-focused AOI settings or from broader AOI settings (when enabled).
- **Time windowing:** instead of always using the full trial, the pipeline supports extracting the last W seconds of each trial. In the experiments, W was swept over multiple values (from 4 to 12 seconds) to test whether shorter or longer windows yield more discriminative gaze dynamics.
- **Feature selection:** experiments compared a compact, robust set of 13 selected channels against configurations that include a broader set of eye-tracking features (when enabled).

Given the sampling rate of 150 Hz, using the last W seconds corresponds to a raw sequence length of approximately $150W$ samples before truncation or padding. For deep learning models requiring fixed-length tensors, sequences were truncated to a maximum length $T_{\max}$ (e.g., 300 or 500 time steps depending on the model and configuration).

8.2.1 Why These Factors Matter

These factors are expected to affect the signal-to-noise ratio of the model input. For example, restricting the input to answer-focused AOI settings reduces unrelated gaze behavior and emphasizes inspection and commitment dynamics, while time windowing controls whether the model focuses on late-stage decision behavior or also includes earlier reading phases.

This windowing strategy is also closely related to the experimental construct itself. As described in Chapter 4, participants were not informed that some trials in the **Timer-No-Correct** condition contained no correct answer. Under such trials, a participant may engage with the item and inspect the options systematically, yet still be forced to finalize a response under uncertainty because no correct solution exists. In such cases, the eventual forced or random-like commitment is expected to emerge most clearly in the final phase of the trial. For this reason, late-trial windows were treated as especially plausible candidates for classification.

At the same time, the final selected window was not imposed a priori as a fixed assumption. Rather, multiple end-of-trial windows were tested empirically in order

to determine which interval yielded the strongest predictive performance. Therefore, the reported best configuration reflects both a behavioral motivation and an empirical model-selection result.

Table 8.1: Main input-configuration factors explored across deep learning experiments.

Factor	**Values explored**
Model families	LSTM, CNN, Hybrid CNN–LSTM, Transformer, MLP
Trial-part selection	Timer trials only; No-Timer + Timer-No-Correct; All parts
AOI filtering	Answer-focused AOI setting; broader AOI setting / all AOIs
Time windows	Last 4–12 seconds; full trial in selected runs
Feature selection	Selected 13 features; broader feature set when enabled

8.3 Experimental Configuration

All reported results were obtained using models implemented in PyTorch and trained with the Adam optimizer. Unless otherwise stated, the reported experiments used:

- Batch size: 16
- Learning rate: 0.0004 (best LSTM configuration)
- Weight decay: 5×10^{-5}
- Maximum epochs: 400
- Early stopping patience: 30

Class imbalance was addressed using weighted binary cross-entropy loss (`BCEWithLogitsLoss`) with loss reweighting derived from the training split. Deterministic seeding was enabled to improve reproducibility across Python, NumPy, and PyTorch, and deterministic CUDA settings were used where supported.

8.3.1 Model Selection Protocol

To avoid attributing performance differences to a single arbitrary choice, the reported results are based on a consistent model selection protocol:

1. **Input configuration exploration:** for each model type, multiple combinations of trial-part selection, AOI filtering, time-window length ($W \in \{4, \ldots, 12\}$ seconds), and feature sets were evaluated.

2. **Hyperparameter tuning:** after identifying strong-performing input configurations, model hyperparameters (e.g., hidden size, dropout, learning rate, weight decay, and early stopping patience) were adjusted to improve stability and peak performance.

3. **Model comparison on a fixed split:** all reported "best" numbers are based on the same fixed, stratified train/test split (seed = 42), enabling consistent comparison across runs.

This protocol helps reduce the risk that the reported improvements are caused solely by a favorable time window or filter setting, but rather reflect a combination of (i) a strong input representation and (ii) an appropriate balance between model capacity and regularization for the small-data regime. At the same time, because the reported best results emerged from exploratory comparisons conducted on a fixed hold-out split, they should be interpreted as optimistic estimates rather than fully unbiased generalization measures.

8.4 Architectures Evaluated

Five model families were evaluated in the experiments reported in this chapter:

- LSTM
- CNN (temporal 1D convolution)
- Hybrid CNN–LSTM
- Transformer encoder

- MLP baseline (flattened/aggregated inputs)

The full architectural specifications, layer configurations, and model diagrams are presented in Chapter 7. In the present chapter, these models are considered only from the perspective of their experimental performance under the explored input and training configurations.

8.5 Evaluation Metrics

The predictive performance of the models is evaluated using standard binary-classification metrics derived from the confusion matrix, where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

- **Accuracy**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-score**

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **ROC-AUC**

  ROC-AUC summarizes the ranking quality of the classifier across decision thresholds and is computed as the area under the receiver operating characteristic curve.

- **Youden Index**

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

These metrics are reported throughout the chapter to characterize both overall classification quality and the trade-off between correctly identifying **RANDOM** trials and avoiding false positives on **NOT_RANDOM** trials.

8.6 Comparative Results Across Architectures

Table 8.2 summarizes the best observed performance per architecture on the fixed hold-out test set. These values correspond to the *best observed run* for each architecture under the explored combinations of trial parts, AOI filters, time windows, feature settings, and hyperparameters. They should therefore be interpreted as best-case architecture-specific results rather than averages across all runs.

Table 8.2: Best observed performance per architecture on the hold-out test set.

Model	Accuracy	F1	Precision	Recall	ROC-AUC	(TN,FP,FN,TP)
LSTM	**0.7647**	**0.8286**	0.7436	**0.9355**	0.6774	(10,10,2,29)
CNN	0.7255	0.8056	0.7250	0.9062	**0.7270**	(8,11,3,29)
Hybrid CNN–LSTM	0.7255	0.7879	0.7429	0.8387	0.7194	(11,9,5,26)
Transformer	0.6731	0.7463	0.7143	0.7812	0.6141	(10,10,7,25)
MLP	0.6731	0.7792	0.6667	0.9375	0.5250	(5,15,2,30)

Overall, the LSTM achieved the best observed test accuracy and the strongest recall for the **RANDOM** class on the reported hold-out split. CNN-based models were competitive and achieved the highest ROC-AUC, whereas the Transformer underperformed relative to recurrent and convolutional models, which is broadly consistent with expectations in small-data regimes. The MLP baseline achieved very high recall but suffered from a large number of false positives, indicating that ignoring temporal structure substantially reduces specificity for **NOT_RANDOM**.

These comparisons should not be interpreted as strictly architecture-only effects, since the best reported run for each model family may correspond to a different input configuration. In addition, the ROC-AUC values remain only moderate overall, indicating that class separation is partial even when threshold-dependent metrics such as recall and F1 appear stronger. This suggests that temporal modeling contributes to discrimination performance to a certain extent, but does not fully resolve overlap between the two behavioral classes.

8.7 Model-Family-Specific Best Configurations

To complement the architecture definitions in Chapter 7, Table 8.3 summarizes the best-performing reported configuration for each evaluated model family. This table is provided to ensure balanced reporting of the final hyperparameter choices across models rather than for the LSTM alone.

Table 8.3: Best-performing reported configuration per model family.

Model	**Best-performing reported configuration**
LSTM	Hidden size = 384; 3 layers; dropout = 0.15; FC hidden dimension = 192; sequence length = 300; learning rate = 0.0004; weight decay = 5×10^{-5}; batch size = 16; early stopping patience = 30; max epochs = 400.
CNN	Channels = [96, 192, 384, 384]; kernel sizes = $7 \rightarrow 5 \rightarrow 3$; dropout = 0.15; FC hidden dimension = 192; learning rate = 0.0008; weight decay = 5×10^{-5}; batch size = 16.
Hybrid CNN–LSTM	CNN channels = 128; LSTM hidden size = 256; 2 LSTM layers; dropout = 0.15; learning rate = 0.0004; weight decay = 5×10^{-5}; batch size = 16.
Transformer	d_{model} = 192; heads = 12; encoder layers = 5; feedforward dimension = 768; dropout = 0.15; GELU activation; learning rate = 0.0004; weight decay = 5×10^{-5}; batch size = 16.
MLP	Hidden layers = [256, 128, 64, 32]; dropout = 0.15; BatchNorm enabled; learning rate = 0.0004; weight decay = 5×10^{-5}; batch size = 16.

The table shows that the final reported settings were broadly comparable in terms of regularization and optimization, while the principal differences across models lay in representational capacity and inductive bias.

8.8 Best Performing Model

The best-performing reported configuration used a three-layer LSTM network with hidden size 384 and dropout 0.15. Table 8.4 reports the key hyperparameters of the

final selected model.

Table 8.4: Best LSTM configuration (final selected model).

Parameter	Value
Hidden size	384
Number of layers	3
Dropout	0.15
FC hidden dimension	192
Sequence length	300
Learning rate	0.0004
Weight decay	5×10^{-5}
Batch size	16
Early stopping patience	30
Max epochs	400

8.8.1 Final Test Metrics

On the held-out test set ($t = 0.50$), the selected LSTM achieved:

- Accuracy: 76.47%
- Precision: 74.36%
- Recall: 93.55%
- F1-score: 82.86%
- ROC-AUC: 67.74%

Taken together, these results indicate that the selected LSTM is particularly effective at identifying **RANDOM** trials, although this comes at the cost of limited specificity for **NOT_RANDOM** trials. In the context of the present thesis, this model is therefore treated as the best-performing architecture under the reported evaluation setup, while the CNN remains noteworthy for achieving the highest ROC-AUC among the tested models.

8.9 Confusion Matrix and Error Profile

The confusion matrix of the best LSTM model on the test set is:

$$\begin{bmatrix} 10 & 10 \\ 2 & 29 \end{bmatrix}$$

This error profile indicates strong sensitivity to the **RANDOM** class, but clearly weaker specificity for **NOT_RANDOM**. In practical terms, half of the **NOT_RANDOM** trials in the test set were misclassified as **RANDOM**, corresponding to a specificity of only 0.50. This is the dominant error mode of the model.

This pattern is consistent with moderate class imbalance in the best-performing subset and with partial overlap in gaze dynamics between the two behavioral classes. At the same time, the small number of false negatives suggests that the model captures a substantial portion of the gaze patterns associated with random-like responding.

8.10 Threshold Analysis

A threshold sweep was conducted on the best LSTM model to evaluate operating-point sensitivity. Table 8.5 reports the sweep summary on the test set.

Table 8.5: Threshold sweep for the best LSTM model (test set).

t	Acc	F1	TN	FP	FN	TP	Youden
0.35	0.6667	0.7848	3	17	0	31	0.1500
0.40	0.6863	0.7895	5	15	1	30	0.2177
0.45	0.7451	0.8219	8	12	1	30	0.3677
0.50	**0.7647**	**0.8286**	10	10	2	29	**0.4355**
0.55	0.6863	0.7576	10	10	6	25	0.3065
0.60	0.6471	0.7097	11	9	9	22	0.2597
0.65	0.6275	0.6780	12	8	11	20	0.2452
0.70	0.5686	0.5769	14	6	16	15	0.1839

The best performance on the tested grid remained at $t = 0.50$, and the Youden index also peaked at this threshold. This indicates that threshold tuning alone does not resolve the observed false-positive behavior. The dominant source of error therefore

appears to be overlap in the learned representations of the two classes, rather than merely an improperly chosen operating point.

8.11 Model-Wise Exploration of Filters and Time Windows

A substantial portion of experimentation focused on identifying which input construction choices maximize discriminability. For each of the five model families (LSTM, CNN, Hybrid CNN–LSTM, Transformer, and MLP), the same general exploration strategy was applied:

- vary **trial-part selection** (timer-only, combined subsets, or all parts),
- vary **AOI filtering** (answer-focused AOI setting versus broader AOI settings when enabled),
- sweep **time-window length** over multiple values (last 4–12 seconds, and optionally full-trial duration),
- compare **compact 13-feature inputs** against larger feature sets (when enabled).

This design aims to reduce the likelihood that the "best model" is merely the result of a favorable architecture choice, but rather the outcome of systematically aligning the model with the most informative view of the gaze signal.

Table 8.6 summarizes the best observed input configuration for each of the five main model families. The purpose of this table is not to claim exhaustive optimality, but to show the dominant configuration pattern that emerged from the experimental search.

Table 8.6: Best observed input configuration per model family.

Model	**Trial parts**	**AOI**	**Time window**	**Best accuracy**
LSTM	Timer trials	Answer-focused AOI setting	9s	76.47%
CNN	Timer trials	All AOIs	11s	72.55%
Hybrid CNN–LSTM	No-Timer + Timer-No-Correct	Answer-focused AOI setting	8s	72.55%
Transformer	No-Timer + Timer-No-Correct	All AOIs	11s	67.31%
MLP	No-Timer + Timer-No-Correct	All AOIs	11s	67.31%

This summary highlights that the best configuration was not identical across model families. In particular, the LSTM performed best under the Timer-trials + answer-focused AOI + 9-second setting, whereas CNN and Transformer models benefited from broader AOI coverage in their best observed runs. Accordingly, the reported model-family differences reflect joint effects of architecture and input configuration rather than architecture alone.

8.11.1 Interpretation of Windowing Effects

From a behavioral standpoint, the last-seconds windowing strategy is motivated by the definition of the labels: the distinction between **RANDOM** and **NOT_RANDOM** depends on behavior close to answer engagement and response commitment. Therefore, shorter windows may emphasize commitment dynamics, while longer windows may include earlier phases (reading/comparison) that may or may not be discriminative for the label.

This motivation is especially relevant for the **Timer-No-Correct** condition. Because participants were not told that some items lacked a correct answer, these trials

could still involve apparently genuine inspection and search behavior before a final forced choice was made. Under this interpretation, random-like responding in such trials may be expressed most strongly near the end of the trial, when the participant can no longer resolve the item and must nevertheless commit to one option. This was one of the main reasons for giving special attention to late-trial windows during the deep learning experiments.

The final selected configuration (Timer trials, answer-focused AOI setting, and a 9-second window, truncated to $T = 300$) reflects one of the more stable trade-offs observed during this exploration. At the same time, because this late-window representation is close to the temporal segment from which the target construct is operationalized, it may also create favorable predictive conditions relative to earlier or broader input views.

8.12 Iterative Optimization Narrative (What Improved and Why)

A key observation throughout experimentation was the high variance across runs caused by random initialization in a small-data regime. Consequently, improvements were assessed through qualitative comparison of multiple logged runs rather than through a fixed repeated-run protocol reported quantitatively in this thesis.

A representative tuning trajectory for the LSTM is summarized in Table 8.7. The baseline configuration achieved approximately mid-60% accuracy, while the tuned configuration reached 76.47% on the fixed hold-out split.

Table 8.7: Representative hyperparameter tuning summary for LSTM.

Stage	Hidden	Dropout	LR	Weight Decay	Patience	Best Acc
Baseline	256	0.25	0.0005	0.0001	20	≈0.67
Tuned	384	0.15	0.0004	5×10^{-5}	30–50	**0.7647**

In practice, broad architecture comparison was performed first, followed by more detailed optimization of the strongest candidate configurations. Because LSTM emerged as the most promising model family, the most detailed hyperparameter tuning focused on this architecture.

In qualitative terms, increasing hidden size improved temporal modeling capacity, while reducing dropout and weight decay avoided over-regularization. A slightly lower learning rate improved convergence stability, and moderate early stopping patience (around 30–60) produced the most robust behavior during exploratory tuning.

8.13 Limitations

Several limitations must be considered when interpreting these results:

- The dataset size is relatively small for deep learning models, increasing sensitivity to random initialization and limiting generalization.
- The evaluation uses a single train/test split rather than participant-level cross-validation; therefore, performance may vary under stricter subject-independent evaluation.
- Eye-tracking signals exhibit substantial individual variability; with limited training samples, models may not fully capture participant-specific strategies.
- Reported scores correspond to the best observed run per architecture rather than average performance across repeated runs; they should therefore be interpreted as indicative of achievable performance under the explored setup, not as definitive estimates of expected performance under repeated resampling.
- Because one experimental condition (**Timer-No-Correct**) is operationally assigned the **RANDOM** label by design, some predictive signal may reflect condition-specific structure rather than purely latent behavioral randomness.
- The strongest-performing inputs focused on late-trial windows close to response commitment; although this is behaviorally motivated, it may also make classification easier by emphasizing signal segments most closely related to the target definition.
- Because the reported best configurations emerged from exploratory comparisons on a fixed hold-out split, the resulting performance estimates may be somewhat optimistic.

8.14 Summary

The deep learning experiments show that gaze dynamics contain predictive information about random-like responding behavior. Among the evaluated models, the LSTM achieved the best observed overall balance of performance on the fixed hold-out split, reaching 76.47% accuracy and 82.86% F1-score under the reported setup.

At the same time, false positives on **NOT_RANDOM** trials remain the dominant error type, specificity remains limited, and threshold tuning does not substantially alter this trade-off. The results also show that input configuration choices—especially trial-part selection, AOI filtering, and time-window selection—play a critical role in performance and must be interpreted jointly with model architecture.

These findings support the feasibility of predicting behaviorally derived randomness labels from gaze trajectories, while also motivating future work with larger datasets, stricter subject-independent evaluation protocols such as LOPO cross-validation, and more conservative validation strategies for configuration selection.

Chapter 9

Conclusion

This thesis introduced a gaze-based framework for distinguishing random from non-random responding behavior in digital assessment environments, with a focus on combining interpretable behavioral modeling and data-driven prediction.

At the core of this work is a three-stage behavioral pipeline that transforms raw eye-tracking streams into structured, trial-level behavioral labels. The pipeline integrates (i) conservative data quality filtering, (ii) AOI-based detection of stable engagement with answer options, and (iii) condition-aware timing thresholds to operationalize random-like responding. This design emphasizes transparency and reproducibility, ensuring that the resulting labels are grounded in observable gaze dynamics rather than outcome-based correctness alone.

The experimental framework was also designed to support methodological control at the level of item construction, interface design, and data acquisition. In particular, the final set of mathematical questions was based on researcher-designed items that were screened in advance for comparable perceived difficulty. This preliminary validation step helped reduce excessive variability associated with overly easy or overly difficult items and strengthened the interpretability of the observed engagement patterns. In addition, the data-collection environment relied on a custom-built interface whose layout and response mechanism were intentionally structured to support stable AOI extraction and interpretable trial-level timing signals. A separate technical pilot phase further helped verify the stability of recording, logging, and export procedures before the main experiment.

The empirical results suggest that the proposed pipeline can produce relatively sta-

ble and interpretable behavioral markers across experimental conditions. In particular, the use of condition-specific thresholds allows the framework to adapt to different task constraints (e.g., time pressure), while maintaining a consistent operational definition of engagement.

Building on these labels, a range of deep learning architectures was evaluated to assess whether temporal gaze patterns contain predictive information about responding behavior. The results indicate that such information is likely present: sequence models, particularly LSTM-based architectures, were able to capture discriminative temporal patterns and achieved the strongest performance under the reported evaluation setup (accuracy 76.47%, F1-score 82.86%).

However, these results should be interpreted with appropriate caution. Performance estimates are based on a fixed train/test split and correspond to the best observed configurations under an exploratory setup. Moreover, classification performance is characterized by high recall but relatively low specificity, indicating that distinguishing **NOT_RANDOM** behavior remains challenging. This suggests that the two behavioral classes exhibit partial overlap in gaze dynamics, and that further improvements may require richer representations or additional contextual signals.

Taken together, the findings of this thesis point toward two main observations. First, gaze-based behavioral features provide a potentially meaningful signal for characterizing engagement and identifying random-like responding beyond traditional accuracy-based metrics. Second, while deep learning models can leverage this signal for prediction, their performance is strongly influenced by input construction choices and evaluation design, particularly in small-data regimes.

9.1 Future Work

Several directions for future research emerge from this work.

- **Stronger evaluation protocols.** Applying participant-level or leave-one-participant-out cross-validation would provide a more reliable estimate of generalization and reduce the risk of optimistic performance estimates associated with fixed splits.
- **Larger and more diverse datasets.** Increasing the number of participants and task types would improve model robustness and enable more reliable training of

higher-capacity architectures.

- **Multimodal behavioral modeling.** Integrating additional signals such as mouse movements, response times, or interaction logs could help disambiguate cases where gaze patterns alone are insufficient.

- **Improved temporal representations.** While Transformer-based models were explored, their performance was limited in the present small-data setting. Future work could investigate their use under larger datasets, improved regularization strategies, or hybrid architectures that better balance global attention and inductive bias.

- **Refinement of behavioral definitions.** Further work could explore alternative or complementary definitions of random responding, including probabilistic labeling schemes or continuous engagement measures rather than binary thresholds.

Overall, this work provides evidence supporting the feasibility of combining gaze analytics and machine learning to study responding behavior in digital assessments. While the proposed framework offers a structured and interpretable starting point, it also highlights the importance of careful evaluation, data design, and modeling choices in developing reliable behavioral inference systems.

Bibliography

[1] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. van de Weijer, *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, 2011.

[2] M.-J. Tsai, H.-T. Hou, M.-L. Lai, W.-Y. Liu, and F.-Y. Yang, "Visual attention for solving multiple-choice science problem: An eye-tracking analysis," *Computers & Education*, vol. 58, no. 1, pp. 375–385, 2012.

[3] F. Goldhammer, J. Naumann, A. Stelter, K. Tóth, H. Rölke, and E. Klieme, "The time-on-task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment," *Journal of Educational Psychology*, vol. 106, no. 3, pp. 608–626, 2014.

[4] S. L. Wise, "Rapid-guessing behavior: Its identification, interpretation, and implications," *Educational Measurement: Issues and Practice*, vol. 36, no. 4, pp. 52–61, 2017.

[5] S. L. Wise and X. Kong, "Response time effort: A new measure of examinee motivation in computer-based tests," *Applied Measurement in Education*, vol. 18, no. 2, pp. 163–183, 2005.

[6] C. Chen and J. Epps, "Using task-induced pupillary response to improve visual-attention-based user authentication," *Pattern Recognition Letters*, vol. 82, pp. 9–15, 2017.

[7] U. Kroehne and F. Goldhammer, "How to conceptualize, represent, and analyze log data from technology-based assessments?" *Measurement: Interdisciplinary Research and Perspectives*, vol. 16, no. 2, pp. 63–72, 2018.

[8] I. Molenaar and M. van Boekel, "Sequencing and combination of eye movements and situational interest in adaptive learning environments," *Learning and Instruction*, vol. 49, pp. 78–91, 2017.

[9] M. Nyström and K. Holmqvist, "An adaptive algorithm for fixation, saccade, and glissade detection in eye-tracking data," *Behavior Research Methods*, vol. 42, no. 1, pp. 188–204, 2010.

[10] F. Goldhammer and T. Martens, "Reading engagement and response time in pisa 2009," *International Journal of Testing*, vol. 13, no. 3, pp. 191–216, 2013.

[11] E. Ulitzsch, S. Pohl, L. Khorramdel, U. Kroehne, and M. von Davier, "A response-time-based latent response mixture model," *Psychometrika*, vol. 87, no. 2, pp. 593–619, 2022.

[12] S. Greiff, S. Wüstenberg, and F. Avvisati, "Computer-generated log-file analyses as a window into students' minds?" *Educational Psychologist*, vol. 50, no. 1, pp. 12–28, 2015.

[13] Q. He and M. von Davier, "Analyzing process data from problem-solving items," *Educational Measurement: Issues and Practice*, vol. 35, no. 3, pp. 20–30, 2016.

[14] A. T. Duchowski, *Eye Tracking Methodology: Theory and Practice*, 3rd ed. Springer, 2017.

[15] A. Rouinfar, A. L. White, M. C. Tannenbaum, D. Sabers, and A. F. Heckler, "Linking attentional processes and conceptual problem solving," *Frontiers in Psychology*, vol. 5, p. 1094, 2014.

[16] M. A. Just and P. A. Carpenter, "A theory of reading: From eye fixations to comprehension," *Psychological Review*, vol. 87, no. 4, pp. 329–354, 1980.

[17] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychological Bulletin*, vol. 124, no. 3, pp. 372–422, 1998.

[18] X. Liu, X. Zhang, W.-W. Chen, and S.-M. Yuan, "Eye movement analysis of digital learning content for educational innovation," *Sustainability*, vol. 12, no. 6, p. 2437, 2020.

[19] A. F. Klaib, N. O. Alsrehin, W. Y. Melhem, H. O. Bashtawi, and A. A. Magableh, "Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and internet of things technologies," *Expert Systems with Applications*, vol. 166, p. 114037, 2021.

[20] K. Majrashi and M. Hamilton, "Eye tracking, usability, and user experience: A systematic review," *Behaviour & Information Technology*, vol. 41, no. 3, pp. 569–590, 2022.

[21] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[22] H. Jarodzka, K. Scheiter, P. Gerjets, and T. van Gog, "In the eyes of the beholder: How experts and novices interpret dynamic stimuli," *Learning and Instruction*, vol. 20, no. 2, pp. 146–154, 2010.

[23] T. van Gog, H. Jarodzka, K. Scheiter, P. Gerjets, and F. Paas, "Attention guidance during example study via the model's eye movements," *Computers in Human Behavior*, vol. 25, no. 3, pp. 785–791, 2009.

[24] X. Kong and S. L. Wise, "The effect of test-taking effort on item response time and accuracy," *Applied Measurement in Education*, vol. 20, no. 2, pp. 163–180, 2007.

[25] S. L. Wise, "Effort analysis: Individual score validation of achievement test data," *Applied Measurement in Education*, vol. 32, no. 3, pp. 237–252, 2019.

[26] B. Steichen, G. Carenini, and C. Conati, "User-adaptive information visualization using eye gaze data," in *Proceedings of the International Conference on Intelligent User Interfaces*, 2013, pp. 317–328.

[27] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Deepgaze iii," *Journal of Vision*, vol. 22, no. 5, p. 7, 2022.

[28] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: A fully convolutional neural network for predicting human eye fixations," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4446–4456, 2017.

[29] Y. Cao, Y. Ding, R. W. Proctor, V. G. Duffy, Y. Liu, and X. Zhang, "Detecting users' usage intentions for websites employing deep learning on eye-tracking data," *Information Technology and Management*, vol. 22, no. 4, pp. 281–292, 2021.

[30] M. Öder, Ş. Eraslan, and Y. Yeşilada, "Automatically classifying familiar web users from eye-tracking data," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 30, no. 1, pp. 233–248, 2022.

[31] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 505–513.