Data Mining Pipeline Report

This report presents a comprehensive analysis of eye-tracking data, processed through a multi-stage pipeline designed to extract meaningful behavioral insights. The pipeline systematically cleans raw gaze data, identifies outliers, labels participant responses, and engineers features related to Areas of Interest (AOIs) and cognitive phases. The objective is to provide a robust framework for understanding user interaction patterns in response to presented stimuli, suitable for academic research and publication.

The analysis covers data from a total of **{n\_participants} unique participants** and **{n\_questions} unique questions**. Each stage of the pipeline is detailed below, including the methodologies applied, key variables computed, and the rationale behind the processing steps.

Overview

**Unique participants:** 23

**Unique questions:** 15

Stage 1 — Data Cleaning and Interaction Time (t\_ij) Computation

**Objective:** This initial stage focuses on refining raw eye-tracking data by removing erroneous gaze samples and calculating the total interaction time for each participant-question pair.

Methodology:

* **Invalid Gaze Sample Removal:** Gaze samples are considered invalid and subsequently removed if their `BPOGV` (Binocular Point of Gaze Validity) value is not equal to 1, or if their gaze coordinates (`BPOGX`, `BPOGY`) are precisely (0,0). These conditions typically indicate data loss or tracking errors.
* **Interaction Time (t\_ij) Computation:** For each unique combination of participant, question, and exam part, the total interaction duration, denoted as **t\_ij**, is calculated. This metric represents the cumulative time a participant spent viewing a specific question. Following this, interactions shorter than 1 second are removed, as they are considered too brief to represent meaningful engagement.

**Summary Statistics for t\_ij:** Mean = 57.41s, Median = 54.90s, Standard Deviation = 26.26s

Stage 2 — Fast Outlier Detection (Lower Bound - LB)

**Objective:** This stage identifies and flags unusually short interaction times (t\_ij) that may represent superficial engagement or premature responses, using a statistical lower bound (LB) threshold.

Methodology:

* **Quartile and Interquartile Range (IQR) Computation:** For each unique question and exam part, the first quartile (Q1), median, third quartile (Q3), and Interquartile Range (IQR = Q3 - Q1) of the `t\_ij` values are calculated. These statistics provide a robust measure of the central tendency and spread of interaction times, minimizing the influence of extreme values.
* **Lower Bound (LB) Calculation:** The Lower Bound (LB) is computed as $Q1 - 1.5 imes IQR$. This formula is a standard method for identifying potential outliers in a dataset, where values falling below the LB are considered statistically anomalous.
* **Time Validity Flagging:** An interaction is flagged as **`invalid\_time`** if its `t\_ij` value is less than the calculated `LB` for that specific question and part. This identifies interactions that are significantly shorter than the typical engagement duration.

Sample of Computed Thresholds (LB)

The table below shows a sample of the calculated Q1, Median, Q3, IQR, and LB values for different question-part combinations. These thresholds are crucial for identifying outliers in interaction times.

| **question\_id** | **part** | **Q1** | **median** | **Q3** | **n\_all** | **IQR** | **LB** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Q1 | Part 1 | 45.810300 | 58.309810 | 58.979770 | 5 | 13.169470 | 26.056095 |
| Q1 | Part 2 | 44.069695 | 55.883185 | 70.666438 | 14 | 26.596743 | 4.174581 |
| Q10 | Part 1 | 28.819580 | 30.792480 | 49.300080 | 3 | 20.480500 | -1.901170 |
| Q10 | Part 2 | 54.819820 | 61.214720 | 71.706550 | 13 | 16.886730 | 29.489725 |
| Q11 | Part 1 | 38.397815 | 58.108855 | 79.377622 | 6 | 40.979807 | -23.071896 |
| Q11 | Part 2 | 42.688960 | 52.529050 | 57.412590 | 9 | 14.723630 | 20.603515 |
| Q12 | Part 1 | 37.199770 | 37.832760 | 63.454225 | 7 | 26.254455 | -2.181912 |
| Q12 | Part 2 | 35.208675 | 50.727480 | 65.870242 | 10 | 30.661567 | -10.783676 |
| Q13 | Part 1 | 53.002750 | 65.487215 | 72.381453 | 4 | 19.378702 | 23.934696 |
| Q13 | Part 2 | 45.910760 | 57.003910 | 84.200620 | 9 | 38.289860 | -11.524030 |
| Q14 | Part 1 | 37.789322 | 49.186890 | 63.218445 | 4 | 25.429123 | -0.354361 |
| Q14 | Part 2 | 38.116670 | 54.002930 | 68.842058 | 10 | 30.725387 | -7.971411 |
| Q15 | Part 1 | 37.835790 | 43.331660 | 74.650630 | 7 | 36.814840 | -17.386470 |
| Q15 | Part 2 | 24.991210 | 36.559810 | 44.865720 | 9 | 19.874510 | -4.820555 |
| Q2 | Part 1 | 41.515492 | 57.888640 | 84.774867 | 6 | 43.259375 | -23.373570 |
| Q2 | Part 2 | 30.727365 | 38.982730 | 49.778018 | 10 | 19.050653 | 2.151386 |
| Q3 | Part 1 | 49.308745 | 57.957765 | 76.604272 | 8 | 27.295527 | 8.365454 |
| Q3 | Part 2 | 50.429495 | 61.719080 | 68.450925 | 7 | 18.021430 | 23.397350 |
| Q4 | Part 1 | 64.509710 | 69.356260 | 85.431990 | 3 | 20.922280 | 33.126290 |
| Q4 | Part 2 | 42.213860 | 46.995020 | 58.558110 | 13 | 16.344250 | 17.697485 |

Stage 3 — Behavioral Labeling (Unusual/Normal Performance - UP/NP)

**Objective:** This stage assigns behavioral labels (Unusual Performance - UP, Normal Performance - NP, Invalid, or Not Applicable) to each participant's response based on their correctness and interaction time relative to a statistically derived upper fence.

Methodology:

* **Filtering for Valid Records:** Labeling is performed exclusively on records deemed valid from Stage 2 (i.e., not flagged as `invalid\_time`).
* **Correct Answer Statistics (Q1\_C, median\_C, Q3\_C, IQR\_C):** Similar to Stage 2, quartile and IQR values are computed, but specifically for `t\_ij` values associated with **only valid correct answers** for each question and part. This creates a baseline for efficient, correct responses.
* **Upper Fence for Correct Answers (UF\_C):** The Upper Fence for Correct answers (UF\_C) is calculated as $Q3\_C + 1.5 imes IQR\_C$. This threshold helps identify correct responses that took an unusually long time, potentially indicating a less efficient problem-solving process despite arriving at the correct answer.

Labeling Logic:

The following rules are applied sequentially to assign a behavioral label:

1. If `UF\_C` cannot be computed (e.g., no valid correct answers for a given question/part), the label is set to NA\_no\_correct (Not Applicable - No Correct Answers).
2. If the participant's answer is **incorrect**, the label is set to UP (Unusual Performance).
3. If the participant's answer is **correct** but their `t\_ij` is greater than `UF\_C`, the label is also set to UP (Unusual Performance), indicating an unusually long time for a correct response.
4. In all other cases (correct answer and `t\_ij` ≤ `UF\_C`), the label is set to NP (Normal Performance).

Label Distribution

The distribution of assigned behavioral labels across all valid interactions is as follows:

| **label** | **count** |
| --- | --- |
| UP | 119 |
| NP | 95 |
| NA\_no\_correct | 14 |
| INVALID | 5 |

Sample of Thresholds for Correct Answers (UF\_C)

This table provides a sample of the calculated Q1\_C, Median\_C, Q3\_C, IQR\_C, and UF\_C values, derived exclusively from correct responses. These thresholds are used to differentiate between normal and unusual performance among correct answers.

| **question\_id** | **part** | **Q1\_C** | **median\_C** | **Q3\_C** | **n\_correct\_valid** | **IQR\_C** | **UF\_C** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Q1 | Part 1 | 58.644790 | 58.979770 | 68.767120 | 3 | 10.122330 | 83.950615 |
| Q1 | Part 2 | 228.007810 | 228.007810 | 228.007810 | 1 | 0.000000 | 228.007810 |
| Q10 | Part 1 | 28.819580 | 30.792480 | 49.300080 | 3 | 20.480500 | 80.020830 |
| Q10 | Part 2 | 55.168000 | 55.858220 | 70.350740 | 3 | 15.182740 | 93.124850 |
| Q11 | Part 1 | 65.950713 | 78.052915 | 83.346907 | 4 | 17.396195 | 109.441200 |
| Q11 | Part 2 | 45.145502 | 47.602045 | 50.058587 | 2 | 4.913085 | 57.428215 |
| Q12 | Part 1 | 36.715330 | 37.684210 | 84.541260 | 5 | 47.825930 | 156.280155 |
| Q12 | Part 2 | 24.597040 | 24.597040 | 24.597040 | 1 | 0.000000 | 24.597040 |
| Q13 | Part 1 | 65.487215 | 68.183410 | 76.579495 | 3 | 11.092280 | 93.217915 |
| Q13 | Part 2 | 58.063780 | 70.216800 | 78.055425 | 3 | 19.991645 | 108.042892 |
| Q14 | Part 1 | 37.789322 | 49.186890 | 63.218445 | 4 | 25.429123 | 101.362129 |
| Q14 | Part 2 | 52.670410 | 58.236090 | 72.377380 | 5 | 19.706970 | 101.937835 |
| Q15 | Part 1 | 39.334495 | 58.999605 | 78.115475 | 4 | 38.780980 | 136.286945 |
| Q15 | Part 2 | 24.897950 | 24.991210 | 44.865720 | 5 | 19.967770 | 74.817375 |
| Q2 | Part 1 | 40.643872 | 68.136825 | 90.468530 | 4 | 49.824658 | 165.205516 |
| Q2 | Part 2 | 34.527135 | 45.783930 | 63.233275 | 3 | 28.706140 | 106.292485 |
| Q3 | Part 1 | 55.852662 | 65.389650 | 92.590157 | 6 | 36.737495 | 147.696400 |
| Q3 | Part 2 | 61.719080 | 62.723630 | 74.178220 | 5 | 12.459140 | 92.866930 |
| Q4 | Part 1 | 64.509710 | 69.356260 | 85.431990 | 3 | 20.922280 | 116.815410 |
| Q4 | Part 2 | 39.971193 | 46.132520 | 58.528963 | 8 | 18.557770 | 86.365617 |

Stage 4 — Area of Interest (AOI) Features & Cognitive Phases

**Objective:** This final processing stage extracts granular features related to specific Areas of Interest (AOIs) on the screen and delineates distinct cognitive phases (Reading and Answering) within each interaction.

Methodology:

* **Cognitive Phase Duration Computation:** The total interaction time (`t\_ij`) is segmented into two primary cognitive phases: **Reading Duration** and **Answering Duration**. This segmentation is critical for understanding how participants allocate their attention during problem-solving.
  + **Question→Choice Transition:** The transition point from the Reading phase to the Answering phase is determined by identifying the first gaze sample that falls within any of the defined `Choice` AOIs after initially fixating on the `Question` AOI. If `BKID` (Button/Key ID) data is available, it is used to precisely mark the moment a participant interacts with an option.
  + **Midpoint Fallback:** In cases where AOI transition data or `BKID` is not available or ambiguous, a fallback mechanism is employed where the midpoint of the total `t\_ij` is used to approximate the transition between reading and answering phases.
* **AOI Time Aggregation:** For each interaction, the cumulative gaze duration within predefined Areas of Interest (AOIs) is calculated. These AOIs typically include: `Question` (the question text area), `Choice\_A`, `Choice\_B`, `Choice\_C`, `Choice\_D` (individual answer options), `Timer` (the countdown timer area), and `Submit` (the submission button area). These aggregated times provide insights into attentional distribution.

Sample of Final Processed Features (Stage 4)

The table below displays a sample of the enriched dataset after Stage 4, including behavioral labels, phase durations, and aggregated AOI gaze times. These features form the basis for further in-depth analysis.

| **participant\_id** | **question\_id** | **part** | **t\_ij** | **label** | **Reading\_duration\_s** | **Answering\_duration\_s** | **Question** | **Choice\_A** | **Choice\_B** | **Choice\_C** | **Choice\_D** | **Timer** | **Submit** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| participant\_1 | Q1 | Part 2 | 175.19141 | UP | 75.39795 | 99.79346 | 89.19044 | 2.46047 | 2.98731 | 1.51268 | 0.23730 | 0.00000 | 0.0 |
| participant\_1 | Q10 | Part 2 | 85.02930 | UP | 63.14648 | 21.88282 | 34.31785 | 4.96877 | 7.86035 | 3.90968 | 5.08984 | 0.11816 | 0.0 |
| participant\_1 | Q11 | Part 2 | 37.12207 | UP | 18.01807 | 19.10400 | 15.61621 | 0.00000 | 0.47509 | 0.00000 | 0.00000 | 0.00000 | 0.0 |
| participant\_1 | Q12 | Part 2 | 84.43994 | UP | 35.72949 | 48.71045 | 26.89939 | 1.91650 | 1.19042 | 0.94189 | 0.67822 | 0.31494 | 0.0 |
| participant\_1 | Q14 | Part 1 | 82.24365 | NP | 6.45117 | 75.79248 | 53.05224 | 0.79785 | 0.52588 | 1.49415 | 0.61621 | 0.20801 | 0.0 |
| participant\_1 | Q15 | Part 2 | 44.86572 | NP | 22.41992 | 22.44580 | 12.68162 | 1.33885 | 2.22315 | 0.88380 | 0.30810 | 0.00000 | 0.0 |
| participant\_1 | Q4 | Part 2 | 32.05957 | NP | 23.64599 | 8.41358 | 9.84717 | 1.11230 | 0.00000 | 2.22265 | 0.16308 | 0.00000 | 0.0 |
| participant\_1 | Q5 | Part 1 | 79.48242 | NA\_no\_correct | 6.36914 | 73.11328 | 42.02001 | 0.73829 | 1.25926 | 0.36816 | 0.81007 | 0.00000 | 0.0 |
| participant\_1 | Q6 | Part 2 | 91.59570 | UP | 45.56250 | 46.03320 | 39.96582 | 5.85449 | 1.30665 | 1.46193 | 1.50488 | 0.52930 | 0.0 |
| participant\_1 | Q9 | Part 1 | 84.47266 | UP | 57.03711 | 27.69092 | 13.21044 | 1.64794 | 2.08398 | 1.71338 | 0.84424 | 0.48535 | 0.0 |
| participant\_10 | Q10 | Part 2 | 54.47778 | NP | 27.02844 | 27.44934 | 33.26595 | 0.77894 | 2.30401 | 0.16131 | 1.12476 | 0.00000 | 0.0 |
| participant\_10 | Q11 | Part 1 | 37.59235 | NP | 18.66156 | 18.93079 | 14.57801 | 0.35602 | 0.00000 | 0.86670 | 0.42206 | 0.10871 | 0.0 |
| participant\_10 | Q12 | Part 1 | 42.36719 | UP | 21.01923 | 21.34796 | 12.40683 | 3.11597 | 2.34991 | 0.53576 | 0.73632 | 0.00000 | 0.0 |
| participant\_10 | Q14 | Part 2 | 29.82794 | UP | 26.02277 | 3.80517 | 8.39649 | 0.00000 | 0.65582 | 0.70453 | 0.42152 | 0.00000 | 0.0 |
| participant\_10 | Q15 | Part 1 | 40.83320 | NP | 20.39679 | 20.43641 | 15.02144 | 0.18066 | 1.46771 | 0.28095 | 0.63702 | 0.26123 | 0.0 |
| participant\_10 | Q2 | Part 2 | 23.27034 | NP | 11.60155 | 11.66879 | 2.19786 | 6.30323 | 0.15537 | 1.29373 | 0.00000 | 0.00000 | 0.0 |
| participant\_10 | Q3 | Part 2 | 132.43420 | UP | 48.74388 | 83.69032 | 57.77601 | 2.41020 | 5.17435 | 0.79577 | 1.59511 | 0.00000 | 0.0 |
| participant\_10 | Q4 | Part 1 | 101.50772 | NP | 50.49939 | 51.00833 | 60.26283 | 2.86984 | 0.98617 | 0.48207 | 1.52591 | 0.00000 | 0.0 |
| participant\_10 | Q6 | Part 2 | 58.32938 | UP | 28.85715 | 29.47223 | 28.59945 | 2.21514 | 1.61481 | 2.75372 | 1.21298 | 0.27448 | 0.0 |
| participant\_10 | Q7 | Part 2 | 81.98227 | UP | 40.86740 | 41.11487 | 20.99789 | 13.15733 | 3.28691 | 3.99327 | 3.95969 | 0.00000 | 0.0 |

Key Variables and Definitions

This section provides a glossary of key variables and terms used throughout the data mining pipeline and in this report, crucial for a thorough understanding of the analysis.

* **`participant\_id`**: A unique identifier assigned to each study participant.
* **`question\_id`**: A unique identifier for each question presented to participants.
* **`part`**: Denotes the section of the exam (e.g., 'Part 1', 'Part 2') to which a question belongs.
* **`BPOGV` (Binocular Point of Gaze Validity)**: A metric indicating the validity of the recorded gaze sample. A value of 1 typically signifies valid gaze data.
* **`t\_ij` (Interaction Time)**: The total duration, in seconds, that participant `i` spent interacting with question `j`.
* **`Q1`, `median`, `Q3`**: The first quartile, median, and third quartile of `t\_ij` values, respectively, calculated per question and part.
* **`IQR` (Interquartile Range)**: The difference between the third and first quartiles (`Q3 - Q1`), representing the spread of the middle 50% of `t\_ij` values.
* **`LB` (Lower Bound)**: A statistical threshold calculated as $Q1 - 1.5 imes IQR$, used to identify unusually short interaction times (outliers).
* **`invalid\_time`**: A flag indicating that an interaction's `t\_ij` fell below the `LB`, suggesting an outlier.
* **`is\_correct`**: A binary variable (1 or 0) indicating whether the participant's answer to a question was correct.
* **`Q1\_C`, `median\_C`, `Q3\_C`**: The first quartile, median, and third quartile of `t\_ij` values, calculated exclusively for **correct answers** per question and part.
* **`IQR\_C` (Interquartile Range for Correct Answers)**: The `IQR` calculated specifically for `t\_ij` values of correct answers.
* **`UF\_C` (Upper Fence for Correct Answers)**: A statistical threshold calculated as $Q3\_C + 1.5 imes IQR\_C$, used to identify unusually long interaction times for correct answers.
* **`label`**: The behavioral label assigned to each interaction:
  + NP (Normal Performance): Correct answer with `t\_ij` within expected range.
  + UP (Unusual Performance): Incorrect answer, or correct answer with `t\_ij` exceeding `UF\_C`.
  + INVALID: Interaction flagged due to `invalid\_time` in Stage 2.
  + NA\_no\_correct: Not Applicable, due to insufficient correct answers to compute `UF\_C`.
* **`Reading\_duration\_s`**: The estimated time, in seconds, a participant spent reading the question and options.
* **`Answering\_duration\_s`**: The estimated time, in seconds, a participant spent actively considering and selecting an answer.
* **AOI (Area of Interest)**: Predefined regions on the screen (e.g., Question, Choice A, Timer, Submit) used to aggregate gaze data.

Notes

* If the background image is not available, the heatmap will be generated without a background.
* The coordinates for writing texts and rectangles on the background are read from config.ini; if not present, default values are used.

Overlay Coordinates (from config.ini)

No [Overlay] section found in config.ini — defaults used.

Per-stage Analysis Summary

This section provides a concise summary of key findings and statistics derived from each stage of the data processing pipeline.

Stage 1 — Data Cleaning & t\_ij Computation Summary

**Total Valid Interactions (post-cleaning):** 233 records.

**Interaction Time (t\_ij) Statistics:**

* Mean t\_ij: 57.41 seconds
* Median t\_ij: 54.90 seconds
* Standard Deviation of t\_ij: 26.26 seconds

These statistics indicate the central tendency and variability of participant engagement times after initial data cleaning and filtering of very short interactions.

Stage 2 — Outlier Detection Summary

**Number of Question-Part Groups with Computed Lower Bounds (LB):** 30.

The outlier detection process identified interactions with unusually short durations, which are critical for understanding potentially disengaged or rushed responses. A sample of the computed LB thresholds is provided above.

Stage 3 — Behavioral Labeling Summary

**Distribution of Behavioral Labels:**

* **UP**: 119 instances.
* **NP**: 95 instances.
* **NA\_no\_correct**: 14 instances.
* **INVALID**: 5 instances.

This distribution provides a high-level overview of participant performance and engagement patterns, categorizing responses into Normal Performance (NP), Unusual Performance (UP), Invalid interactions, and cases where correct answer thresholds could not be established (NA\_no\_correct).

Stage 4 — AOI Features & Cognitive Phases Summary

**Cognitive Phase Durations:**

* **Reading Duration (s):** Mean = 26.70, Median = 24.87, N = 233
* **Answering Duration (s):** Mean = 30.73, Median = 27.45, N = 233

These metrics offer insights into how participants divide their attention between understanding the question and formulating a response. The aggregated AOI times (presented in the sample table above) further detail specific attentional foci.

Overall Statistical Summary

This section provides a high-level statistical overview of the entire dataset and the results of the pipeline.

* **Total Participants Analyzed:** 23
* **Total Questions Analyzed:** 15
* **Overall Mean Interaction Time (t\_ij):** 57.41 seconds
* **Overall Median Interaction Time (t\_ij):** 54.90 seconds
* **Overall Behavioral Label Distribution:**
  + UP: 119 instances
  + NP: 95 instances
  + NA\_no\_correct: 14 instances
  + INVALID: 5 instances

Visualizations

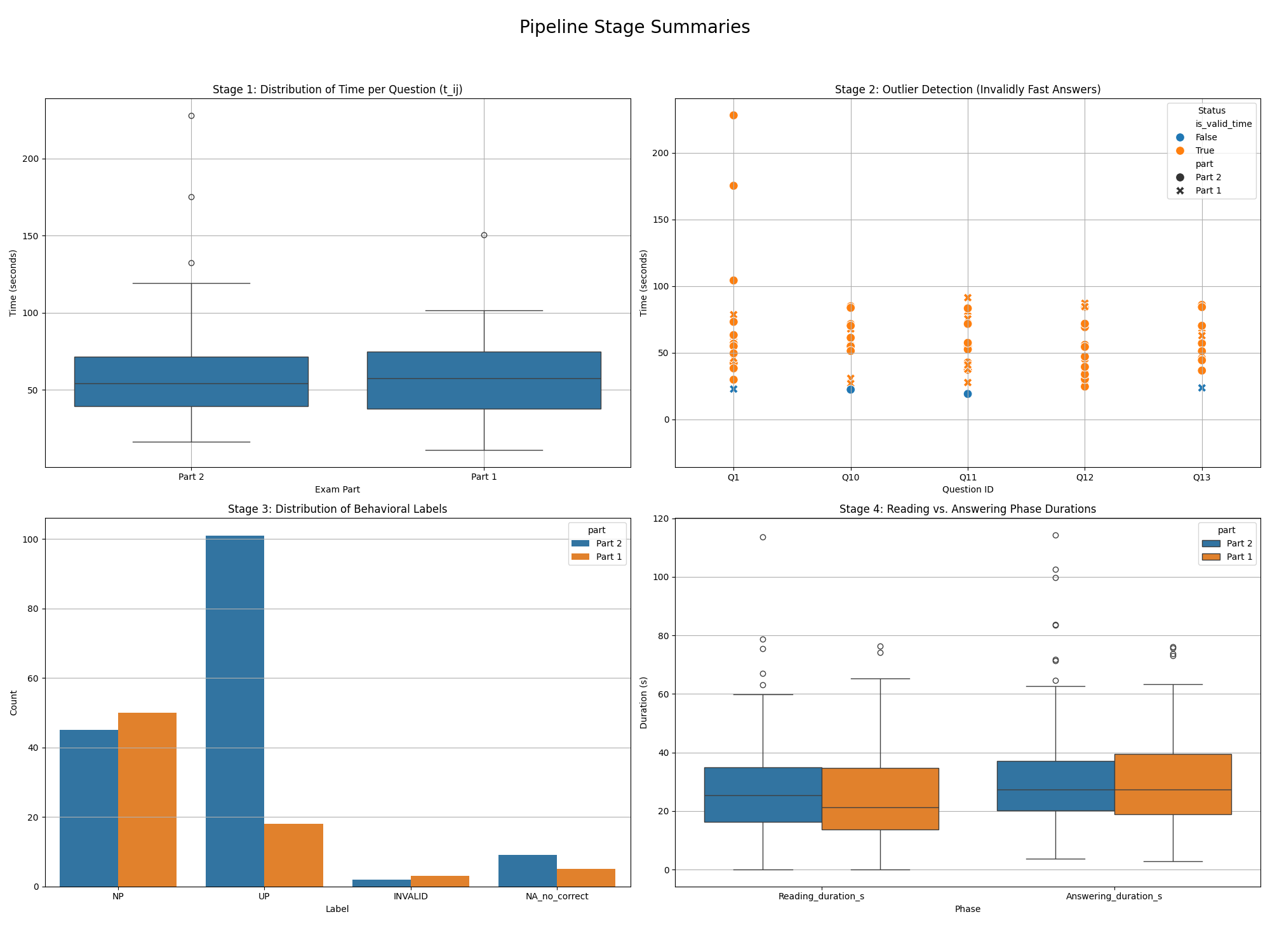
AOI Time Summary per Question



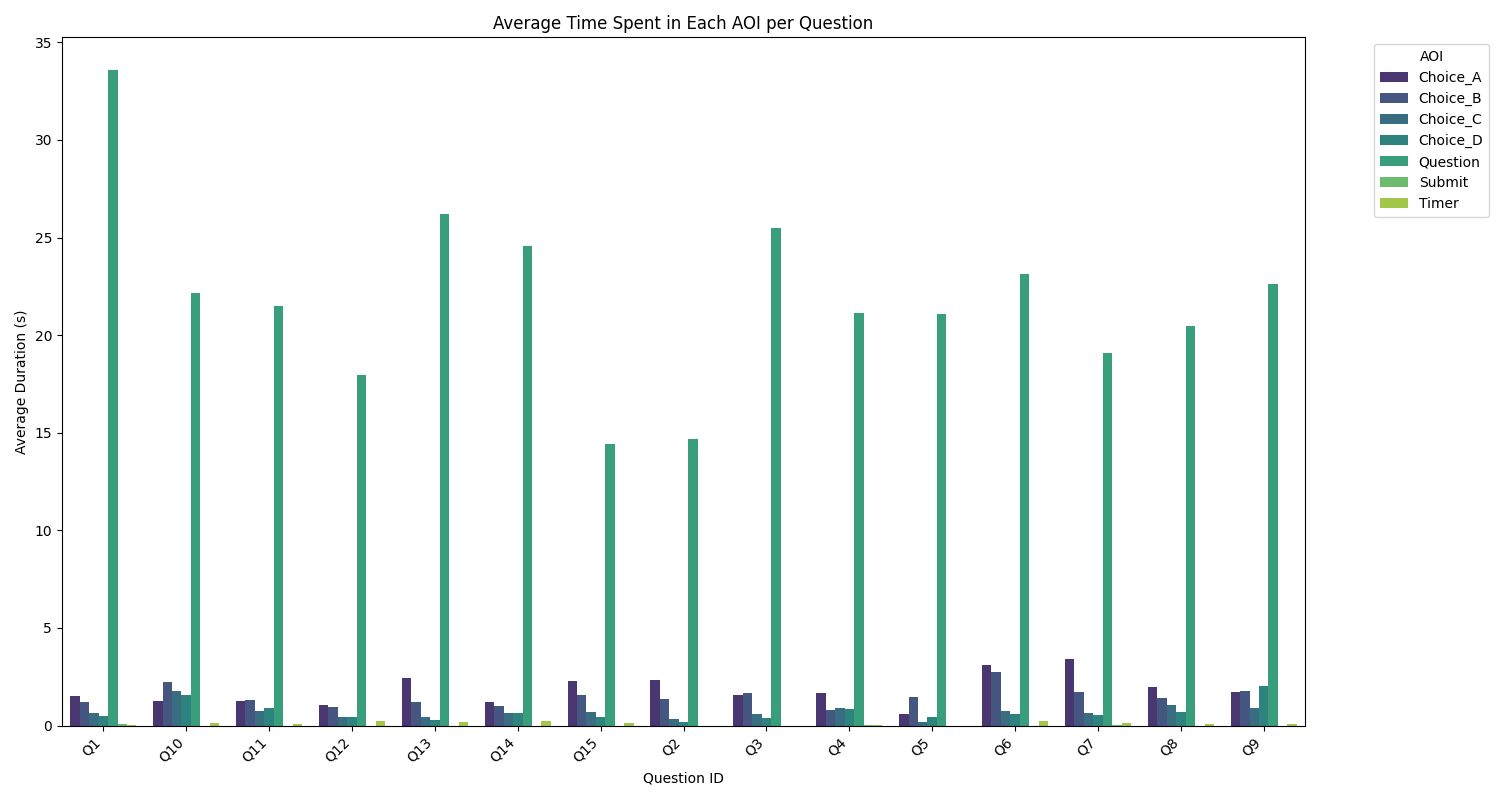
Numeric Summary (per question)

| **question\_id** | **Question\_s** | **Correct\_Answer\_s** | **Other\_Answers\_s** | **Question\_pct** | **Correct\_Answer\_pct** | **Other\_Answers\_pct** |
| --- | --- | --- | --- | --- | --- | --- |
| Q1 | 33.575 | 0.252 | 3.605 | 89.695 | 0.673 | 9.632 |
| Q10 | 22.147 | 1.232 | 5.606 | 76.408 | 4.250 | 19.342 |
| Q11 | 21.516 | 0.639 | 3.597 | 83.552 | 2.482 | 13.966 |
| Q12 | 17.940 | 0.345 | 2.534 | 86.173 | 1.655 | 12.172 |
| Q13 | 26.204 | 1.177 | 3.225 | 85.618 | 3.846 | 10.536 |
| Q14 | 24.555 | 0.290 | 3.161 | 87.677 | 1.036 | 11.288 |
| Q15 | 14.432 | 0.928 | 4.079 | 74.241 | 4.776 | 20.984 |
| Q2 | 14.683 | 0.259 | 3.935 | 77.783 | 1.371 | 20.846 |
| Q3 | 25.492 | 0.566 | 3.644 | 85.826 | 1.906 | 12.268 |
| Q4 | 21.120 | 0.479 | 3.727 | 83.393 | 1.891 | 14.715 |
| Q5 | 21.091 | 0.494 | 2.196 | 88.689 | 2.078 | 9.233 |
| Q6 | 23.143 | 0.440 | 6.737 | 76.331 | 1.450 | 22.219 |
| Q7 | 19.069 | 0.421 | 5.867 | 75.203 | 1.660 | 23.137 |
| Q8 | 20.445 | 0.769 | 4.319 | 80.071 | 3.013 | 16.916 |
| Q9 | 22.634 | 0.711 | 5.719 | 77.876 | 2.445 | 19.679 |

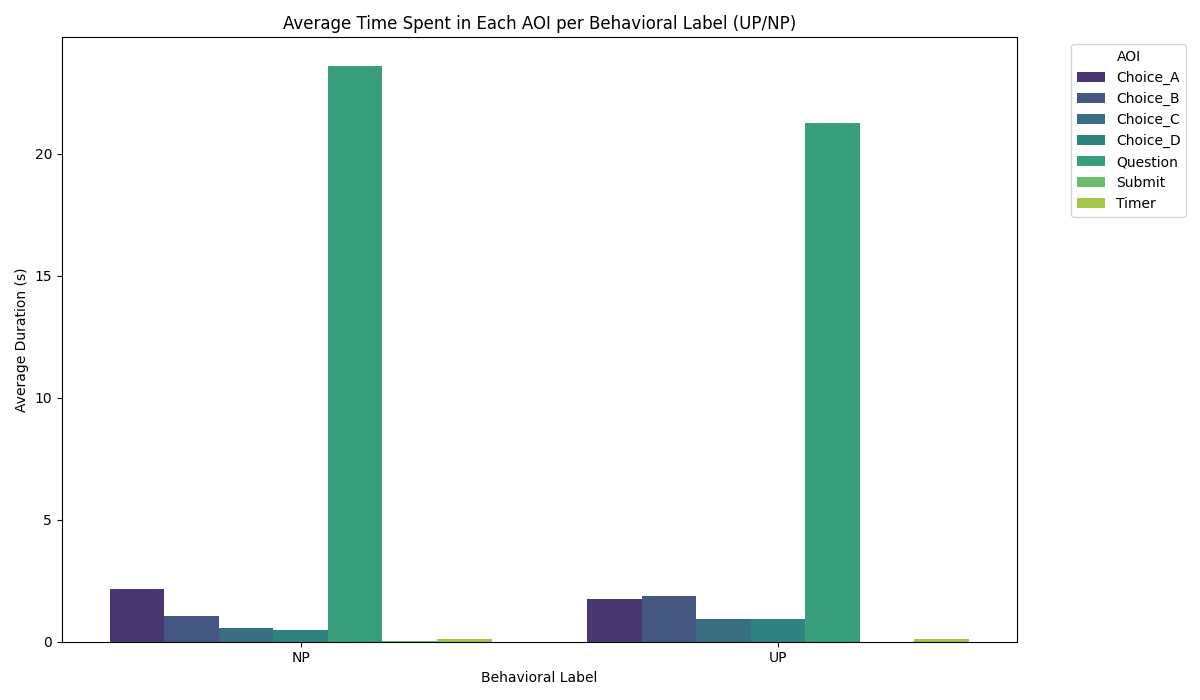
Pipeline Summary Plots



AOI Time per Question



AOI Time per Label



Participant Participant 1