NZCER review of NZQA's internal Automated Text Scoring (ATS) Model Performance

Model marking performance statistics and recommendations

Jessie Dong
with Charles Darr and Elliot Lawes

NZCER
Rangahau Mātauranga o Aotearoa

NZQA
Mana Tohu Mātauranga o Aotearoa
New Zealand Qualifications Authority

Rangahau Mātauranga o Aotearoa | New Zealand Council for Educational Research
Te Pakokori
Level 4, 10 Brandon St
Wellington
New Zealand

www.nzcer.org.nz

https://doi.org/10.18296/rep.0094

NZCER review of NZQA's internal Automated Text Scoring (ATS) Model Performance

Model marking performance statistics and recommendations

Jessie Dong
with Charles Darr and Elliot Lawes

February 2026

NZCER
Rangahau Mātauranga o Aotearoa

NZQA
Mana Tohu Mātauranga o Aotearoa
New Zealand Qualifications Authority

He ihirangi | Contents

Tables

Figures

NZQA management comment

We thank NZCER for this comprehensive and well-researched report and are pleased at the finding that NZQA's Automated Text Scoring (ATS) model demonstrates a high level of consistency and comparability with human markers.

We are particularly satisfied to note that NZQA's in-house ATS model performs comparably across gender, ethnicity, and socioeconomic indicators, consistent with patterns observed in human marking. We also recognise that, on average, it is currently slightly more conservative than human markers, with a narrower score distribution.

As we move closer to deploying the model, we will work through all recommendations made in this report and maintain quality assurance of results, including human oversight, ongoing validation, and fairness monitoring. We are committed to building internal capability to ensure the model continues to operate equitably as student populations and assessment contexts evolve.

We have already actioned NZCER's recommendation to investigate six outliers identified during their analysis. We are confident that any future issues with essay text will be identified through the guardrails that have recently been implemented as part of the NZQA ATS model. This will enable these responses to be considered for human marking.

We thank NZCER for their thoughtful analysis, which will help ensure students receive accurate and timely results from the Co-Requisite Writing assessment in 2026 and beyond.

1. Purpose

The purpose of this report is to present an independent analysis of the New Zealand Qualifications Authority's (NZQA's) Automated Text Scoring (ATS) model.[1] It evaluates the model's performance, bias/fairness, and readiness for marking literacy writing assessments associated with US 32405, as well as its potential transferability to other assessment standards.

The review was carried out by statistical and psychometric staff at the New Zealand Council for Educational Research (NZCER).

1 In this report, NZQA's internal Automated Text Scoring (ATS) model is referred to as the **NZQA ATS model**, or the **ATS model** for brevity.

2. Introduction

Background

NZQA is using machine scoring to mark the extended written responses that are part of the literacy writing assessment (US 32405). To achieve this, NZQA has used a commercial marking model. Over the past 2 years, NZQA has also invested in developing its own in-house ATS model.

The commercial marking model was used for live scoring in both the 2025 assessment events (AE1 and AE2). To date, the NZQA ATS model has been used to support the quality assurance process of marked responses.

Scope

This review assesses the marking performance of the NZQA ATS model for literacy writing. It includes statistical analyses and visualisations of the ATS model's scores against human marking from the first 2025 live assessment events (AE1).

Data from the 2024 Pilot were reviewed as a supplement to the 2025 AE1 analysis, where it helped identify any significant differences or emerging patterns.

3. Methods

Datasets

NZQA provided NZCER with human-marked data and NZQA ATS model scores from the 2024 Pilot and 2025 AE1 assessments for this review. Corresponding essay texts and high-level demographic flags were included to support a bias/fairness review. The assessment prompts, marking rubrics, and a brief of the NZQA ATS model were also supplied to provide context.

For consistency and comparability, two subsets of essays that were marked by both human markers and the ATS model for each of the assessment events were selected for analysis.[2]

Essays with no text were excluded from the datasets, as they are not scorable. Scores of 0 and 1 were collapsed into a single score of 1 to ensure that all scores are on a consistent 1–4 scale across datasets.

Details of the clean datasets are summarised in Table 1. Note the 2025 AE1 dataset comprises scores for two assessments from Week 1 and Week 2 of the assessment period, each containing two extended written prompts (Question 1 and Question 2). Different prompts were used each week. In the remainder of this report, each question or prompt is referred to as a task and the responses as essays.

TABLE 1 **Summary of the reviewed dataset**

Dataset	Task	Description	Count
2025.AE1	W1Q1	Question 1 in week 1, AE1 2025	10783
2025.AE1	W1Q2	Question 2 in week 1, AE1 2025	10436
2025.AE1	W2Q1	Question 1 in week 2, AE1 2025	8258
2025.AE1	W2Q2	Question 2 in week 2, AE1 2025	7992
2024.Pilot	W1Q1	Question 1 in week 1, Pilot 2024	17503
2024.Pilot	W1Q2	Question 2 in week 1, Pilot 2024	17497
2024.Pilot	W2Q1	Question 1 in week 2, Pilot 2024	17492
2024.Pilot	W2Q2	Question 2 in week 2, Pilot 2024	17486

Evaluation metrics

Exact Agreement, Adjacent Agreement, Cohen's Kappa (CK), and Quadratic Weighted Kappa (QWK) were used as the primary indicators of agreement between AI model scores and human marks. Pearson's Correlation was used to assess linear consistency between the two sets of scores. Matthews Correlation Coefficient (MCC) was additionally computed as a confusion-matrix-based measure of the overall consistency between two sets of scores. These metrics are summarised in Table 2.

2 Human marking occurred for the subset of responses that were close to the achieved / not achieved boundary.

TABLE 2 **Evaluation metrices used to review the model performance**

Metric	Description
Exact Agreement	Percentage of essays where the model score equals the human score.
Adjacent Agreement	Percentage of essays where the model score is within ±1 score point of the human score.
Exact + Adjacent	Percentage of essays where the model score is in exact or adjacent agreement with the human score.
Pearson's Correlation	Measure of the linear correlation between model and human scores.
Cohen's Kappa (CK)	Unweighted measure of agreement between the human and model score after accounting for chance (sensitive to skew).
Quadratic Weighted Kappa (QWK)	Weighted measure of agreement between the model and human scores after accounting for chance, giving partial credit for ordinal differences between scores.
Matthews Correlation Coefficient (MCC)	Confusion-matrix-based measure of overall classification consistency between model-predicted and human-assigned score categories, balancing correct and incorrect outcomes across score levels.

Among these metrics, QWK is the most informative for evaluating automated essay scoring models. Because essay scores are ordinal (i.e., ordered categories with meaningful distances between levels), QWK provides a more accurate reflection of marking reliability than unweighted metrics. It is widely adopted in large-scale assessment research, AES benchmarking (e.g., Kaggle competitions[3]), and operational scoring systems as the standard indicator of model–human agreement.

The metrics listed in Table 2 were calculated for each assessment by each rubric element, as well as for a holistic overall score that was generated by averaging the four element scores.

In addition to these agreement metrics, we also reviewed several descriptive statistical measures, including the mean, median, standard deviation, and effect size when comparing the model with human or different groups.

Rounding

The NZQA ATS model outputs scores that are continuous numbers with 2 decimal places, while human marks are discrete integers. To enable direct comparison and agreement analysis, the ATS model scores were rounded to the nearest whole number.

It is worth noting that, by default, Python and R use "bankers' rounding" (also called round half to even) which rounds .5 values to the nearest even number (e.g., 2.5 → 2, 3.5 → 4). This rounding method reduces aggregate rounding bias in large datasets, but it may not be the best practice for scoring in education settings as it might shift the score distributions slightly and introduce unfairness compared to the "round half up" method.

3 https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2
https://www.kaggle.com/competitions/asap-aes/overview

To maintain alignment with common assessment practice, this review used the round half up method, where .5 values are always rounded up to the next whole number.

However, we retained the original float scores for Pearson Correlation to avoid artificial discretisation, because rounding continuous outputs can reduce variance and distort relationships between variables, leading to artificially lower correlations.

Rounded scores were also used in the visualisations, including the confusion matrices and density distributions, to ensure consistency and clearer interpretation of score patterns.

Bias/fairness investigation

Bias and fairness were examined across key demographic groups: Gender, Māori/Non-Māori, Pacific/Non-Pacific, and Equity Index (EQI) groups.

For each subgroup, the same set of evaluation metrics used above (e.g., Exact Agreement, Adjacent Agreement, and QWK) was calculated to assess model and human agreement within groups and to identify any systematic differences in performance.

In addition, subgroup mean scores and standard deviations were compared between the model outputs and human marks. Effect sizes were computed to quantify the magnitude of any score differences between groups (e.g., Māori vs. Non-Māori, Male vs. Female).

This "two-dimensional" approach allows for both:

- vertical analysis—examining consistency of model performance *within* each demographic group (fairness of scoring reliability), and
- horizontal analysis—comparing *between* groups to detect any systematic score gaps or bias patterns.

Together, these analyses provide a comprehensive view of fairness and subgroup performance consistency across the NZQA ATS model versus human marking.

In addition, we also reviewed outliers (large disagreements between model and human scores) to see how big the effect is for these essays that are marked very differently by human and AI models.

4. Results and findings

4.1 Performance statistics

Overall, the NZQA ATS model showed strong and reliable performance across all rubric elements for all tasks in the first assessment event 2025 (AE1), with agreement levels exceeding typical human-to-human agreement.[4] The detailed performance statistics by task and rubric element for the ATS model are presented in Table 3.

The main insights include:

- Scores from the NZQA ATS model are well aligned with human marking. The model achieved a QWK = 0.717, showing substantial agreement with human marking on the ordered 1–4 rubric scale. This level of agreement is higher than the average agreement typically seen between two human markers (QWK ≈ 0.6). Additional measures including CK (= 0.517) and MCC (= 0.523) also show moderate to strong reliability.
- Agreement with human scores was consistently high. The NZQA ATS model achieved an Exact Agreement of around 0.797 (0.734–0.844) and a Combined Exact and Adjacent Agreement close to 1.00. This means that nearly all ATS model marked scores were within one point of human marks.
- Correlations between the NZQA ATS model and human marks were also strong (Pearson's r = 0.826). This indicated that the model closely reflected the relative ranking of essays across the score scale, rubric elements, and tasks.
- The model performed consistently across all rubric dimensions, with minimal variation and small standard deviations, demonstrating robustness and generalisability across tasks.

Similar performance patterns were also observed in the 2024 Pilot data (Appendix A) supporting the evidence of stability of the NZQA ATS model across assessment events. We also note that the model performance improved in 2025 AE1 compared to the 2024 Pilot. This improvement could be linked to the additional model training that took place following the Pilot.

4 Research suggests that the average quadratic QWK score between two human raters is approximately 0.6 per rubric element (see Kumar & Boulanger, 2021). Comparable results were observed in NZCER's internal PAT Writing assessment trial involving 3,821 double-marked essays.

TABLE 3 **NZQA ATS model marking performance statistics in comparison with human marks, by task and rubric element in AE1 2025**

Task	Rubric element	Model	N	Exact	Adj	Exact+Adj	Pearson	CK	QWK	MCC
W1Q1	Accuracy	NZQA	10783	0.74	0.26	1	0.784	0.541	0.672	0.543
W1Q1	Content	NZQA	10783	0.759	0.238	0.997	0.782	0.354	0.622	0.359
W1Q1	Language	NZQA	10783	0.796	0.202	0.999	0.791	0.486	0.678	0.496
W1Q1	Structure	NZQA	10783	0.804	0.195	0.999	0.809	0.429	0.678	0.442
W1Q2	Accuracy	NZQA	10436	0.734	0.265	0.999	0.804	0.543	0.681	0.543
W1Q2	Content	NZQA	10436	0.819	0.18	0.998	0.852	0.512	0.759	0.518
W1Q2	Language	NZQA	10436	0.799	0.2	0.999	0.838	0.527	0.736	0.534
W1Q2	Structure	NZQA	10436	0.819	0.18	0.999	0.856	0.533	0.761	0.539
W2Q1	Accuracy	NZQA	8258	0.745	0.254	1	0.791	0.551	0.672	0.555
W2Q1	Content	NZQA	8258	0.843	0.154	0.997	0.829	0.501	0.73	0.505
W2Q1	Language	NZQA	8258	0.822	0.177	0.999	0.818	0.531	0.713	0.537
W2Q1	Structure	NZQA	8258	0.829	0.17	0.999	0.827	0.483	0.722	0.494
W2Q2	Accuracy	NZQA	7992	0.76	0.239	0.999	0.825	0.573	0.693	0.573
W2Q2	Content	NZQA	7992	0.844	0.151	0.995	0.873	0.568	0.794	0.581
W2Q2	Language	NZQA	7992	0.811	0.187	0.998	0.863	0.574	0.769	0.579
W2Q2	Structure	NZQA	7992	0.83	0.167	0.998	0.877	0.568	0.786	0.571
mean				**0.797**	**0.201**	**0.998**	**0.826**	**0.517**	**0.717**	**0.523**
sd				0.037	0.038	0.001	0.032	0.058	0.049	0.057

4.2 Summary statistics

Overall, the NZQA ATS model produced slightly lower mean scores (2.58) than human markers (2.66) on average across rubric elements and tasks, shown in Table 4. This small downward difference in the ATS model represents a small negative effect size of (-0.12).

TABLE 4 **Summary statistics for NZQA's ATS model in comparison with human marks in AE1 2025**

Task	Rubric	N	median_ ATS	median_ human	mean_ ATS	mean_ human	sd_ ATS	sd_ human	effect_ size
W1Q1	Accuracy	10783	2.52	2.00	2.43	2.44	0.57	0.72	-0.02
W1Q1	Content	10783	2.73	3.00	2.60	2.88	0.55	0.68	-0.45
W1Q1	Language	10783	2.80	3.00	2.66	2.76	0.56	0.68	-0.15
W1Q1	Structure	10783	2.88	3.00	2.73	2.92	0.57	0.68	-0.31
W1Q2	Accuracy	10436	2.43	2.00	2.33	2.32	0.65	0.76	0.01
W1Q2	Content	10436	2.95	3.00	2.78	2.82	0.70	0.76	-0.05
W1Q2	Language	10436	2.79	3.00	2.62	2.70	0.67	0.75	-0.11
W1Q2	Structure	10436	2.85	3.00	2.68	2.79	0.68	0.76	-0.15
W2Q1	Accuracy	8258	2.50	2.00	2.40	2.36	0.57	0.71	0.07
W2Q1	Content	8258	2.82	3.00	2.68	2.81	0.58	0.68	-0.21
W2Q1	Language	8258	2.79	3.00	2.65	2.71	0.57	0.67	-0.10
W2Q1	Structure	8258	2.91	3.00	2.76	2.85	0.59	0.69	-0.14
W2Q2	Accuracy	7992	2.33	2.00	2.22	2.18	0.68	0.77	0.04
W2Q2	Content	7992	2.89	3.00	2.67	2.70	0.76	0.83	-0.04
W2Q2	Language	7992	2.73	3.00	2.52	2.61	0.73	0.82	-0.12
W2Q2	Structure	7992	2.77	3.00	2.55	2.68	0.74	0.82	-0.17
mean			**2.73**	**2.75**	**2.58**	**2.66**	**0.64**	**0.74**	**-0.12**
sd			0.18	0.45	0.16	0.22	0.07	0.06	0.13

The standard deviation of the NZQA ATS model score is slightly smaller than it is for human marks (0.64 vs. 0.74), suggesting a slightly narrower scoring range or tighter marking behaviour across rubric elements and tasks.

Interestingly, the median scores for the ATS model and human scoring are closely aligned (2.73 vs. 2.75). This suggests that there is no major shift in the central tendency.

To help us understand what caused the small downwards differences in average score while the central tendency remained similar, we tried some visualisations (Figure 1) and observed that the small downward difference likely results from instances where essays marked as "4" by human markers were scored as "3" by the ATS model.

Figure 1 compares the score distributions for human markers and the NZQA ATS model. The human-scored distributions show a distinct peak at score 4, which is reduced in the distributions for the ATS model. The taller peak at score 3 in NZQA's results indicates that some top scores were slightly shifted down.

A few additional observations were made:

- Wider range in human marks: Human markers used a broader range of scores, with visible clusters around 2 and 3, and smaller peaks at 1 and 4. It also shows greater variability between rubric elements, reflecting differences in focus across the rubric.
- Slightly tighter score distribution in the NZQA ATS model: The ATS model produced score distributions that closely follow the human pattens, with a mild centralisation around score 3 and slightly fewer 4s. This small compression accounts for the marginally lower standard deviation while maintaining a clear human-like pattern.
- Patterns are consistent across all assessments (W1Q1–W2Q2), suggesting systematic rather than prompt-specific behaviour.

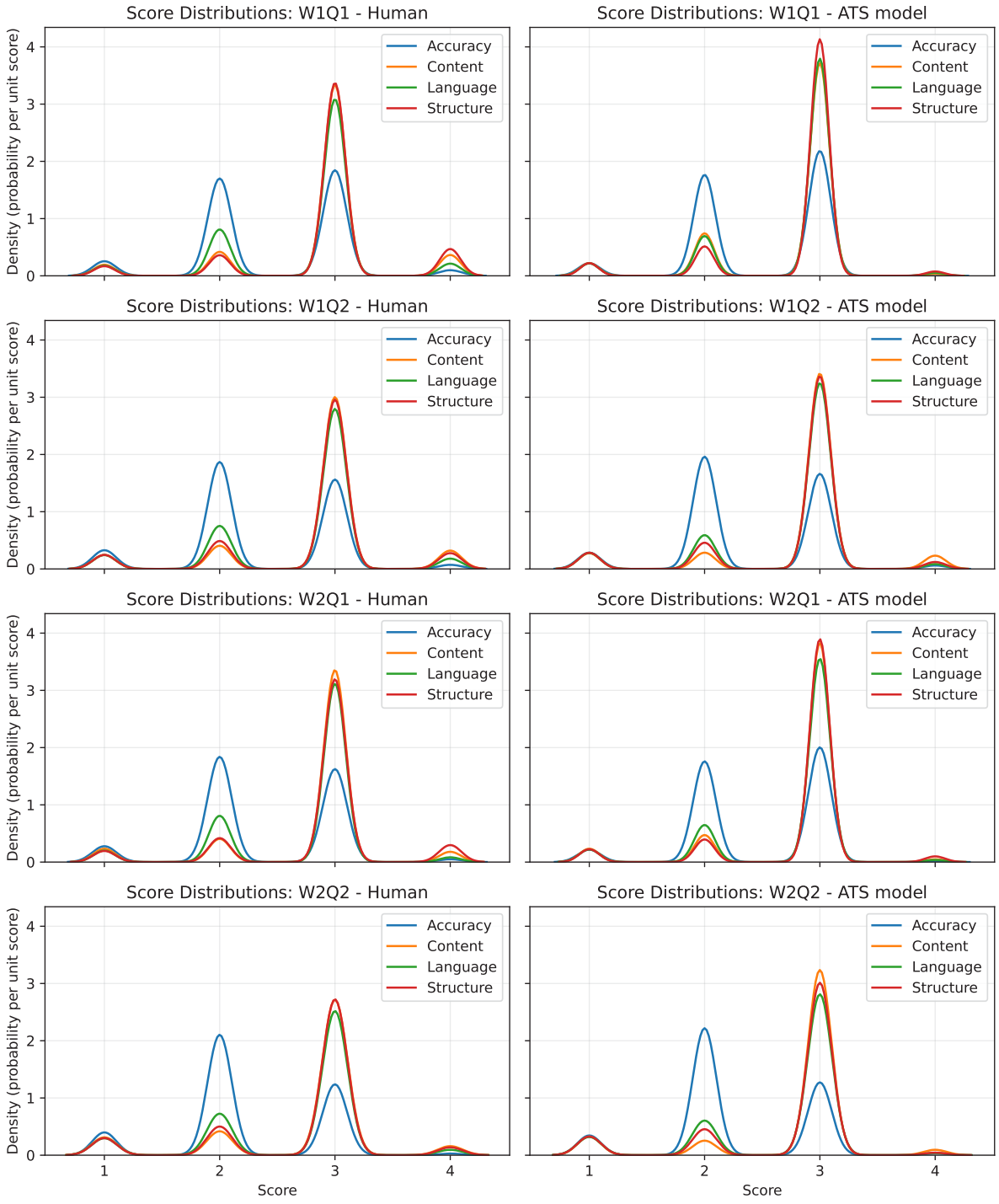


FIGURE 1 **Score distributions of human marks and the NZQA ATS model across four tasks and four rubric dimensions in AE1 2025**

4.3 Model behaviour analysis

This section examines how the NZQA ATS model behaves in relation to human marking patterns. The analysis uses three complementary perspectives: confusion matrices, score distribution comparison, and intra-rubric relationships.

- **Confusion matrices** show how model-assigned scores match human scores across the marking scale.
- **Score distribution plots** illustrate how the model and human markers use the scoring scale for each rubric element.
- **Intra-rubric relationships** assess how scores from different rubric elements correlate within human and model marking.

Together, these perspectives provide a comprehensive view of how closely the ATS model replicates human marking behaviour.

Confusion matrices

A confusion matrix is a table used to compare two sets of scores. In this case, the NZQA ATS model's predictions against human marks. The rows represent the actual human scores, and the columns represent the model-predicted scores. Each cell shows how many essays received a particular combination of scores. A strong alignment between model and human marking appears as a dark diagonal line, where most predictions fall exactly on or very near the correct score.[5]

Figure 2 presents the confusion matrices for the NZQA ATS model versus human scores, across all tasks for Week 1 and Week 2 in AE1 2025.

The confusion matrices show strong alignment between the ATS model and human marking, with most of the ATS model predictions concentrated along the diagonal. This indicates a high level of score agreement across all rubric elements and tasks. Minor deviations appear mainly between adjacent scores (e.g., 2<–>3 or 3<–>4), suggesting that most discrepancies are within one score point rather than large misclassifications.

The confusion matrices for the ATS model versus human scores in the 2024 Pilot show similar patterns and are available in Appendix B.

5 In a confusion matrix, darker diagonal cells indicate exact matches between model and human scores. Adjacent cells represent 1-point differences. A dense diagonal with minimal off-diagonal spread shows reliable marking and minimal bias.

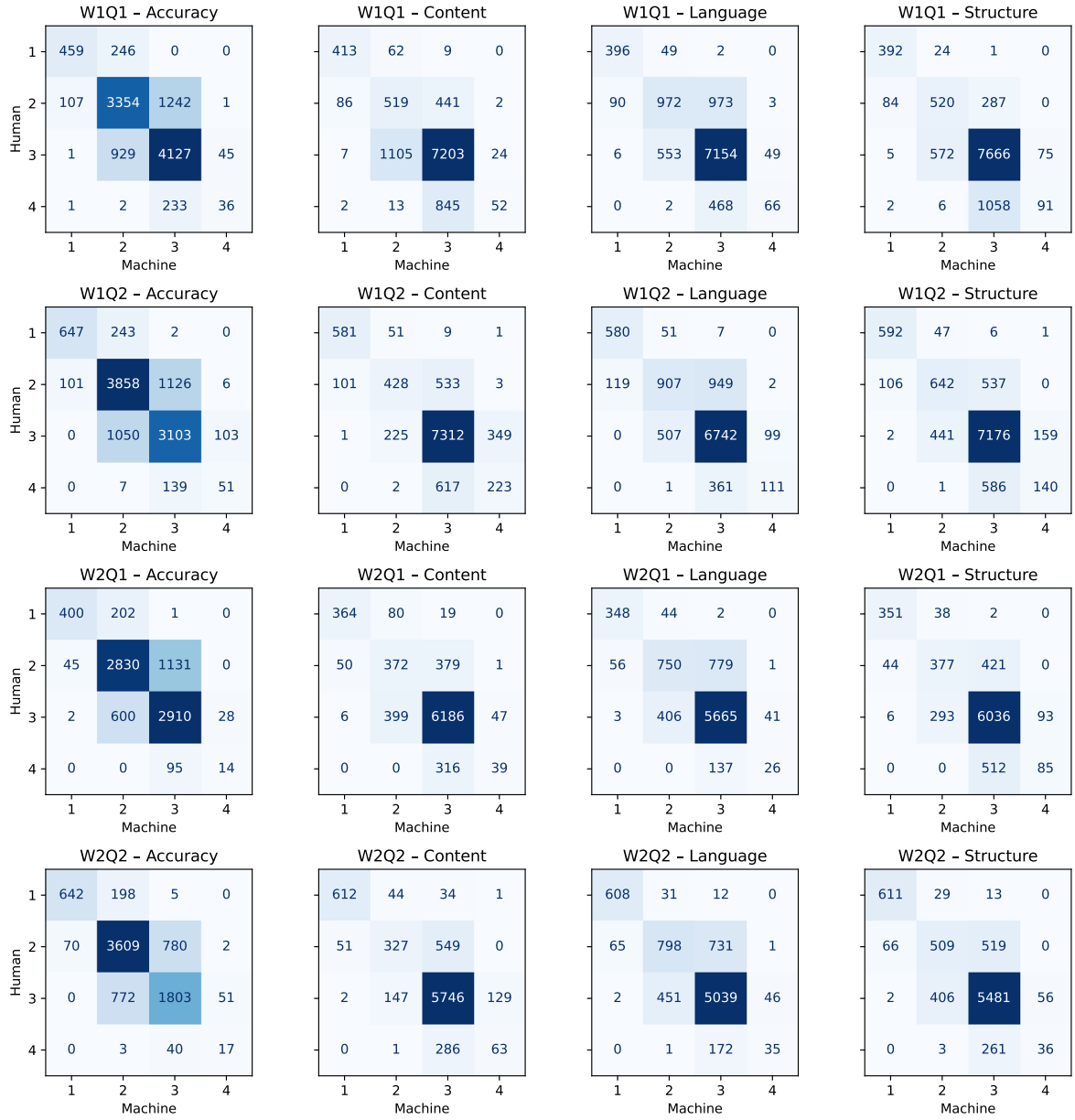


FIGURE 2 **Confusion matrices for the NZQA ATS model vs. human, by rubric element across all tasks in AE1 2025**

Outliers

A small number of essays (six in total from AE1 2025) showed differences greater than two score points between human and NZQA's model marks. These outliers are located at the corners of the confusion matrix shown in Figure 2.

To help us better understand how big the effect is for these outliers, the complete set of scores and essay texts for each task were pulled out and investigated. It appears that the NZQA model did not "misbehave" on these essays. Instead, it was traced to text processing issues where parts of the essay text were truncated or missed before scoring by the ATS model. The affected cases have been passed on to the NZQA team for review.

It is important that NZQA investigates these cases further to identify the cause of the text exporting or processing errors and confirm where in the data pipeline the truncation occurred. Ensuring that the full essay text is consistently and accurately processed before scoring will help prevent similar issues in future assessments and maintain the reliability and fairness of the marking system.

Score distribution analysis

Figure 3 shows the density plots of score distributions for human markers and the NZQA ATS model, by rubric element. These plots show how frequently each score (1–4) was assigned and provide insights into how the model and human markers use the 1–4 marking scale.

This analysis provides a different perspective from the distribution plots in Figure 1. While Figure 1 compares score distributions across the four rubric elements, Figure 3 focuses on comparing how each rubric element individually has been scored by humans and the NZQA ATS model. It highlights the distribution patterns within each rubric element and how closely the models replicate human marking behaviour.

The overlapping curves show that the model closely replicates human use of the marking scale, with highly similar shapes across the score range. Both human and model scores are concentrated around score 3, reflecting consistent alignment in recognising mid-range essay quality.

A small difference is observed at the upper end of the scale, where the NZQA model assigns fewer 4s compared to human markers, resulting in marginally taller peaks at score 3. This pattern, also seen in Figure 1, suggests a mild conservative tendency in the model's scoring for this dataset.

However, the overall alignment remains strong, with the model closely reflecting human judgement across all rubric elements and tasks. It is important to note that this dataset primarily includes essays near the cut scores, so the lower frequency of score 4 does not indicate any limitation in the model's ability to assign top scores.

FIGURE 3 **Comparison of score density distribution[6] from human marks and the NZQA ATS model, by rubric element in AE1 2025**

Intra-rubric relationships

To evaluate the relationship between rubric elements, Spearman correlations[7] were used because of the discrete and ordinal nature of the scores.

Tables 5 and 6 show Spearman correlations between the four rubric elements (Accuracy, Content, Language, and Structure) for human and NZQA ATS model scores (rounded), respectively.

6 In the figure, each curve represents how frequently a score was assigned.
7 The Spearman correlation captures how the rank order of scores across rubric elements aligns, without assuming equal intervals or linearity, which makes it more suitable for discrete, ordinal data.

Human-marked scores show moderate correlations between rubric elements (r = 0.45–0.60). This is expected, as each rubric assesses a related but distinct writing construct. The moderate relationships indicate that human markers differentiate between rubric elements while still recognising natural overlaps in writing quality (e.g., strong structure often accompanies strong language).

The NZQA ATS model shows slightly higher correlations between the rubric elements (r = 0.54–0.83) than human markers. The strongest relationships appear between Content, Language, and Structure (r ≈ 0.80+), while Accuracy remains more independent (r = 0.54–0.60). This pattern is consistent with the nuanced relationships found in human scoring and indicates that the ATS model reflects the same underlying structure of writing quality dimensions.

TABLE 5 **Correlation between rubric element scores marked by human in AE1 2025**

	Accuracy_human	Content_human	Language_human	Structure_human
Accuracy_human	1.00	0.45	0.47	0.47
Content_human	0.45	1.00	0.55	0.60
Language_human	0.47	0.55	1.00	0.53
Structure_human	0.47	0.60	0.53	1.00

TABLE 6 **Correlation between rubric element scores marked by the NZQA ATS model in AE1 2025**

	Accuracy_nzqa	Content_nzqa	Language_nzqa	Structure_nzqa
Accuracy_nzqa	1.00	0.58	0.60	0.54
Content_nzqa	0.58	1.00	0.81	0.83
Language_nzqa	0.60	0.81	1.00	0.82
Structure_nzqa	0.54	0.83	0.82	1.00

4.4 Holistic score performance

Holistic scores, calculated as the average of the four rubric element scores, provide an overall measure of model performance at the essay level. An analysis of the holistic score complements the rubric-level analysis by summarising how well the model replicates human marking when all elements are considered together.

Table 7 presents the holistic score performance statistics for the NZQA ATS model, compared with human marking.

The ATS model achieved Exact Agreement with human-marked responses between 0.798 and 0.855, and Pearson Correlations between 0.882 and 0.921 across all assessments. Reliability measures (QWK = 0.673–0.796; Kappa = 0.437–0.601) indicate substantial agreement and strong consistency. Task W1Q1 shows slightly lower alignment (QWK = 0.673) compared with other tasks (QWK = 0.75–0.796), but overall performance remains within a strong and reliable range.

Overall, the NZQA ATS model demonstrates high reliability, strong agreement, and stable performance across all tasks at the holistic level.

TABLE 7 **Model marking performance statistics for holistic scores[8] across all assessment in AE1 2025—NZQA ATS model vs. human**

Task	N	Exact	Adjacent	Exact + Adj	Pearson	Kappa	QWK	MCC
W1Q1	10783	0.798	0.201	0.999	0.882	0.437	0.673	0.446
W1Q2	10436	0.835	0.164	0.999	0.91	0.561	0.773	0.565
W2Q1	8258	0.855	0.144	0.999	0.904	0.554	0.75	0.56
W2Q2	7992	0.84	0.158	0.998	0.921	0.601	0.796	0.608

4.5 Bias and fairness analysis by subgroups

Bias and fairness were evaluated across key demographic dimensions: Gender, Māori/Non-Māori, Pacific/Non-Pacific, and EQI groups. The same set of performance metrics including Exact Agreement, Adjacent Agreement, CK, QWK, Pearson's Correlation, and MCC were calculated for each subgroup. In addition, subgroup mean scores, standard deviations, and effect sizes were analysed to detect any systematic differences in model behaviour or outcomes.

In summary, no evidence of demographic bias was found. The NZQA ATS model mirrors human scoring across gender, ethnicity, and socioeconomic indicators. Across all demographic groups, the ATS model showed consistent performance relative to human marking. It appears that the model replicated what is existing in the human marked data.

The score distribution by gender for human marks and the NZQA ATS model is presented in Figure 4. It is clear that the shapes, medians (white dots), and interquartile spreads are almost identical. Visualisations for Māori/Non-Māori, Pacific/Non-Pacific, and EQI groups are available in Appendix C.

Similarly, Māori/Non-Māori, Pacific/Non-Pacific, and EQI group comparisons showed nearly identical means and standard deviations for human and AI-scoring, with small effect sizes. This indicates that any observed differences are statistically negligible. The model's scoring behaviour remained balanced and predictable across all subpopulations, reflecting good generalisability and fairness in application.

8 Holistic score = average of four rubric element scores.

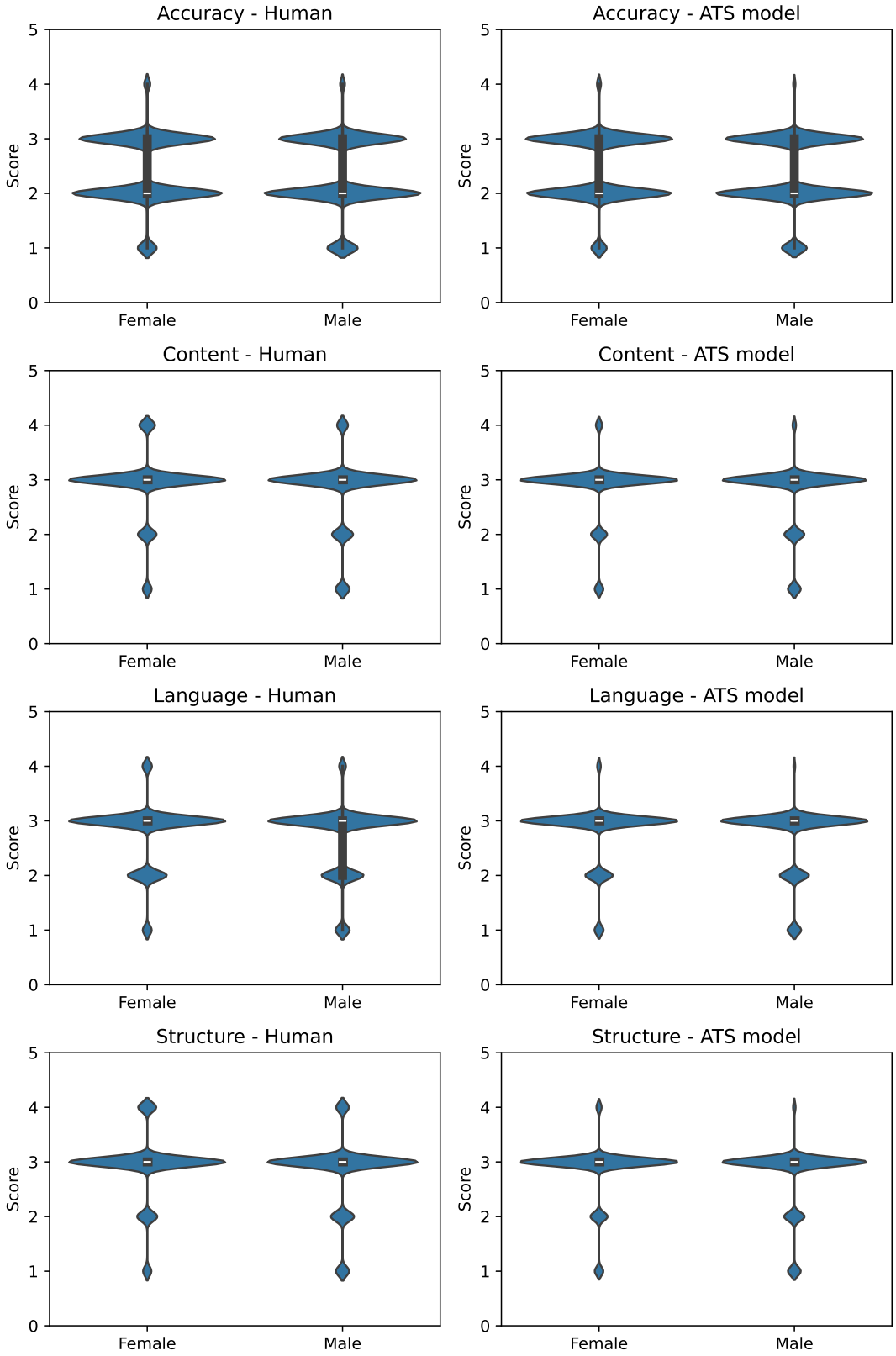


FIGURE 4 **Score distribution by gender for human, NZQA ATS models in AE1 2025**

TABLE 8 **NZQA ATS model marking performance statistics in comparison with human marks, by subgroups in AE1 2025 (mean values across tasks)**

Gender	N	Exact	Adj	Exact+Adj	Pearson	CK	QWK	MCC
Male	22131	0.800	0.198	0.998	0.844	0.540	0.739	0.546
Female	15270	0.793	0.205	0.998	0.789	0.479	0.674	0.486
Māori	10233	0.739	0.257	0.996	0.779	0.441	0.657	0.453
Non-Māori	27236	0.744	0.252	0.996	0.763	0.414	0.640	0.426
Pacific	4211	0.731	0.266	0.997	0.781	0.444	0.662	0.454
Non-Pacific	33258	0.744	0.251	0.996	0.766	0.419	0.643	0.431
More	7019	0.738	0.258	0.996	0.813	0.488	0.697	0.498
Moderate	21870	0.765	0.232	0.997	0.790	0.442	0.664	0.455
Fewer	7229	0.763	0.234	0.997	0.788	0.422	0.652	0.436

A small difference was observed in the subgroup analysis (Table 8). The NZQA ATS model showed a slightly higher correlation with human marks for male students (r = 0.844) compared to female students (r = 0.789). Similarly, the mean QWK values were higher for males than for females (0.739 vs. 0.674). Comparable but smaller differences were also found across other subgroups. These results suggest that each subgroup performs somewhat differently in writing assessments.

However, the differences between subgroup means for human marks and ATS model scores across all tasks are consistently less than 0.1 (Appendix D). This indicates that the model closely mirrors human marking performance, with no evidence of systemic bias. The complete set of pairwise statistics for each subgroup is provided in Appendix D.

4.6 Potential transferability to other assessments

Transferability to other assessment contexts cannot be fully assessed without more information about those assessments. However, based on the description of the NZQA ATS model, it represents a strong and modern approach for text-based assessments.

The NZQA ATS model uses a transformer-based large language model (DeBERTa-v3-Large) within a supervised learning framework. This is the current state of the art in automated essay scoring. The model is trained on human-marked data and retrained/refined after each exam cycle as new data become available. This allows the model to evolve and maintain alignment with current marking standards.

Because Question 1 and Question 2 in each assessment have different word limits, NZQA trained two separate models, one for shorter response tasks and one for longer response tasks, to ensure fairness and prevent essay length from influencing scoring outcomes. The inclusion of guardrails for off-topic and safety detection, a word-count moderation rule, and an LLM explainer for score transparency further strengthens this approach.

However, this approach also has limitations. It relies on the availability of high-quality human-marked data and is task-specific, designed primarily for literacy writing with structured, rubric-based marking. Applying the model to other types of assessments would require careful model adjustment, retraining, and validation to ensure accuracy and fairness in new contexts.

5. Discussion

Taken together, the combined results from performance metrics, summary statistics, confusion matrices, and score distributions confirm that the NZQA ATS model replicates human scoring with strong reliability and minimal bias.

The bias and fairness analysis found no evidence of systematic model bias across the demographic dimensions reviewed. The NZQA ATS model mirrors human scoring across gender, ethnicity, and socioeconomic indicators (as represented by EQI groups).

These results support its suitability for operational use, provided fairness monitoring continues as part of routine quality assurance.

Assumptions

This review assumes that no data leakage occurred prior to scoring using the marking model. Specifically, the NZQA ATS model did not have prior exposure to the dataset from the first live assessment event 2025 (AE1).

Limitations

The reviewed dataset focused on essays with total scores near the cut scores, representing only a subset of the national data. This limited range reduces variability and can lower QWK values. When most essays fall between 2 and 3, any disagreement at the extremes (1 or 4) disproportionately affects QWK. Thus, slightly lower QWK values reflect the dataset characteristics rather than reduced model accuracy.

To align scores across datasets, the review recoded the original 0–4 scale into a 1–4 scale by collapsing 0s and 1s. This compression may have slightly influenced agreement and correlation results, though the impact is considered small.

6. Recommendations

The NZQA ATS model is technically strong, reliable, and fair, achieving higher agreement with human markers than is typically observed between human markers. With continued monitoring, validation, and human moderation, it is well-positioned for operational deployment in literacy writing assessments. To support this aim, we have developed the following recommendations.

Readiness for operational use

The NZQA ATS model is ready for scaled use in literacy writing from a marking quality perspective, provided it operates within a moderated system with ongoing validation and human oversight. It should be noted that the design and implementation of the ATS marking system itself are outside the scope of this review.

Before full deployment, NZQA should ensure that:

- human moderation or oversight remains in place, with humans reviewing essays flagged for low confidence or boundary cases
- ongoing validation continues across prompts/tasks to confirm model stability.

Bias and fairness monitoring

- Increase training samples at score 4 (and potentially at score 1) to strengthen representation at scale extremes.
- Use balanced sampling across all score ranges during validation to improve calibration and fairness.
- Maintain subgroup performance reviews (Gender, Māori/Non-Māori, Pacific/Non-Pacific, EQI) as part of regular quality assurance.

Data processing and system checks

- Apply consistent rounding strategies to enhance marking consistency and fairness where rounding is needed.
- Investigate identified outliers and review text exporting and cleaning processes to ensure the wider assessment system does not unintentionally modify input data.

Reference

Kumar, V. S., & Boulanger, D. (2021). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence Education*, *31*, 538–584. https://doi.org/10.1007/s40593-020-00211-5

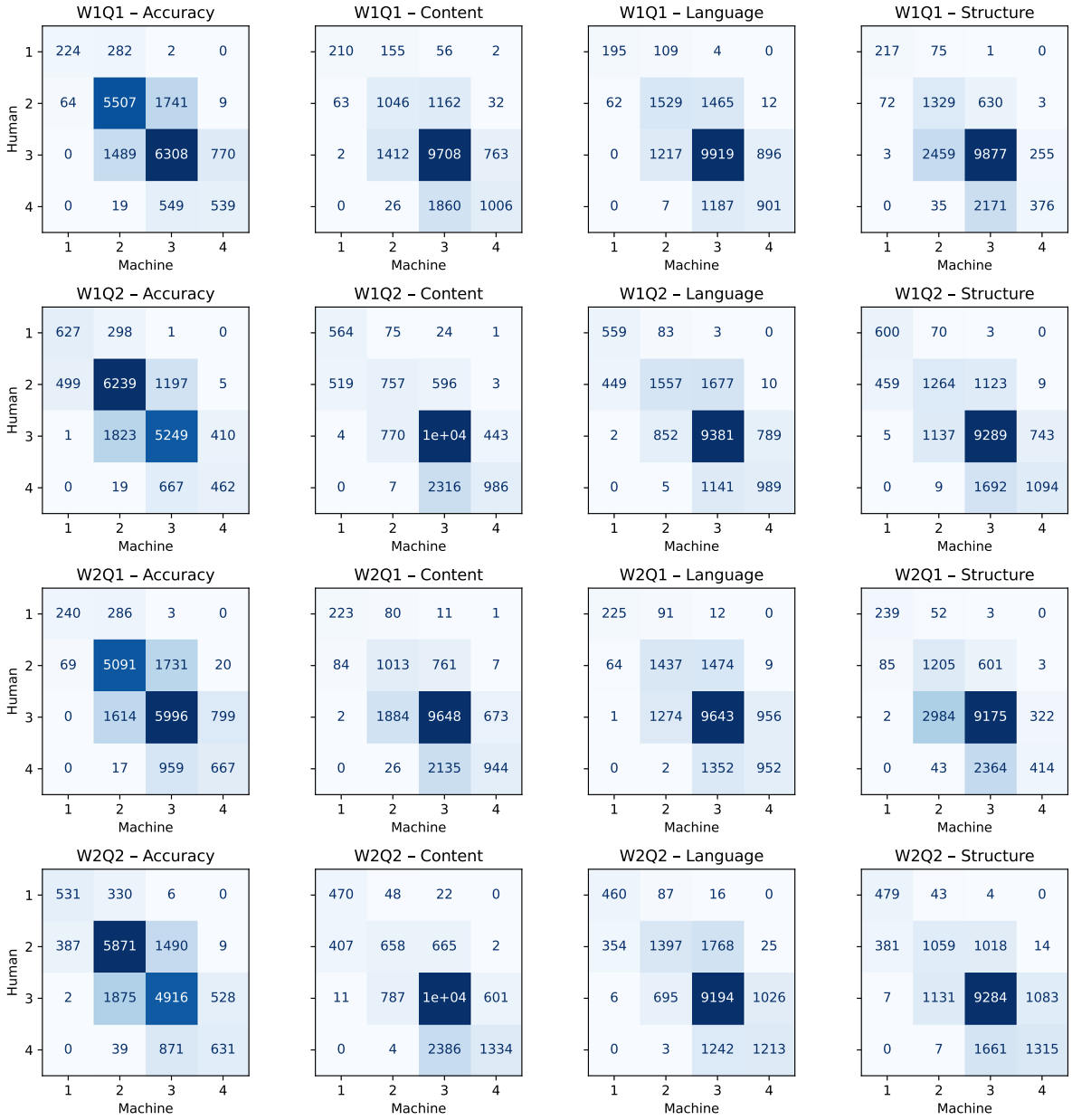
Appendices

APPENDIX A
NZQA ATS model marking performance statistics in comparison with human marks, by task and rubric element in 2024 Pilot assessment

Task	Rubric element	Model	N	Exact	Adj	Exact+Adj	Pearson	CK	QWK	MCC
W1Q1	Accuracy	NZQA	17503	0.719	0.28	0.998	0.738	0.515	0.661	0.515
W1Q1	Content	NZQA	17503	0.684	0.309	0.993	0.634	0.323	0.524	0.326
W1Q1	Language	NZQA	17503	0.717	0.282	0.999	0.675	0.39	0.574	0.391
W1Q1	Structure	NZQA	17503	0.674	0.323	0.998	0.659	0.271	0.492	0.279
W1Q2	Accuracy	NZQA	17497	0.719	0.28	0.999	0.776	0.533	0.705	0.534
W1Q2	Content	NZQA	17497	0.728	0.27	0.998	0.779	0.413	0.68	0.427
W1Q2	Language	NZQA	17497	0.714	0.285	0.999	0.76	0.445	0.68	0.45
W1Q2	Structure	NZQA	17497	0.7	0.299	0.999	0.767	0.418	0.677	0.422
W2Q1	Accuracy	NZQA	17492	0.686	0.312	0.998	0.729	0.471	0.648	0.471
W2Q1	Content	NZQA	17492	0.676	0.321	0.997	0.66	0.302	0.525	0.306
W2Q1	Language	NZQA	17492	0.701	0.298	0.999	0.677	0.368	0.567	0.369
W2Q1	Structure	NZQA	17492	0.631	0.366	0.997	0.657	0.216	0.466	0.224
W2Q2	Accuracy	NZQA	17486	0.683	0.313	0.997	0.761	0.482	0.678	0.482
W2Q2	Content	NZQA	17486	0.718	0.28	0.998	0.768	0.402	0.657	0.413
W2Q2	Language	NZQA	17486	0.701	0.296	0.997	0.748	0.424	0.655	0.43
W2Q2	Structure	NZQA	17486	0.694	0.304	0.998	0.747	0.396	0.651	0.397
mean				0.697	0.301	0.998	0.721	0.398	0.615	0.402
sd				0.024	0.024	0.001	0.051	0.087	0.077	0.085

APPENDIX B

Confusion matrices of the NZQA ATS model vs. human scores in the 2024 Pilot assessments



APPENDIX C
Figures for subgroup bias analysis in AE1 2025

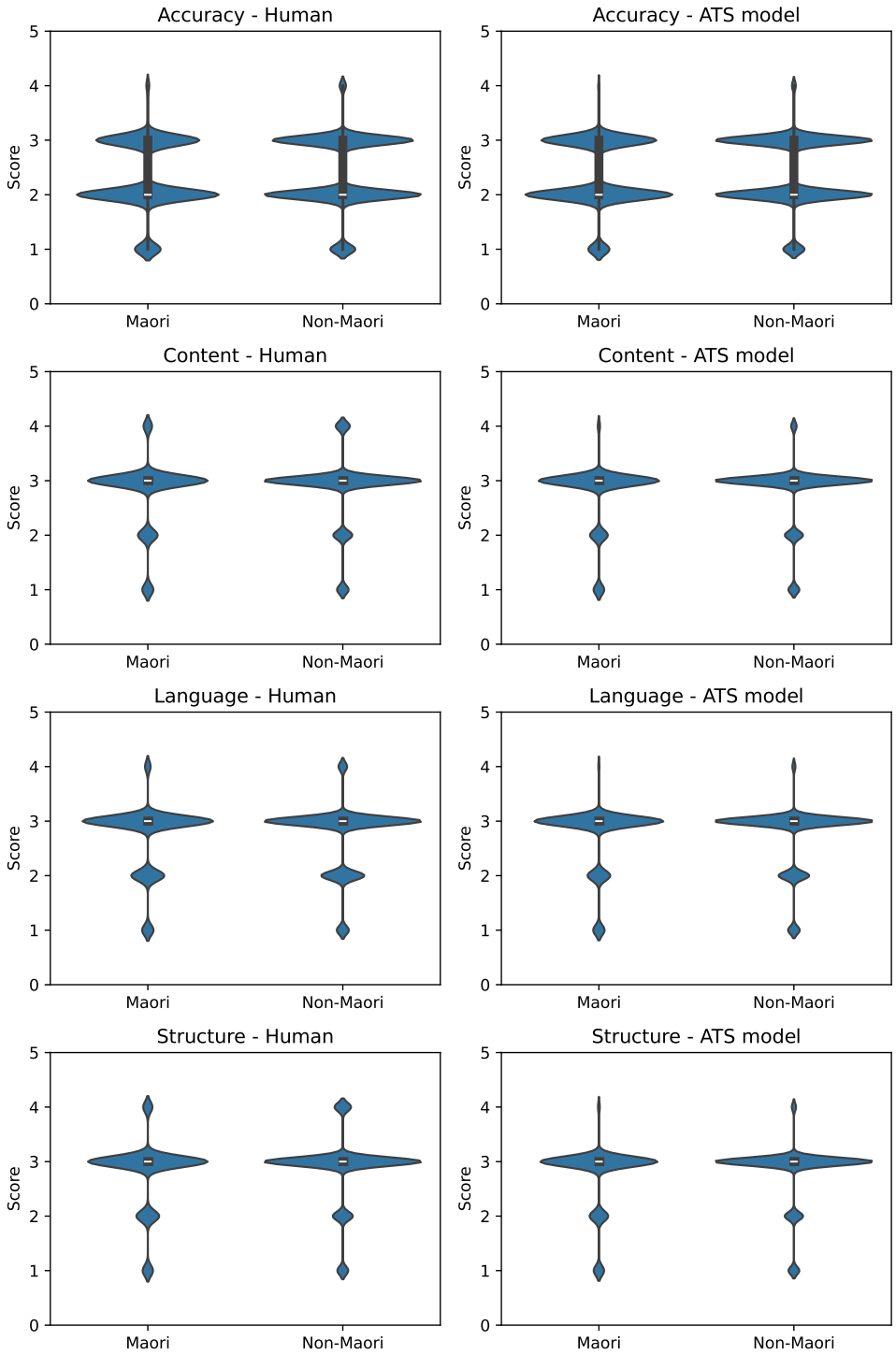


FIGURE C1 **Score distribution of human and the NZQA ATS model, by Māori/Non-Māori in 2025 AE1**

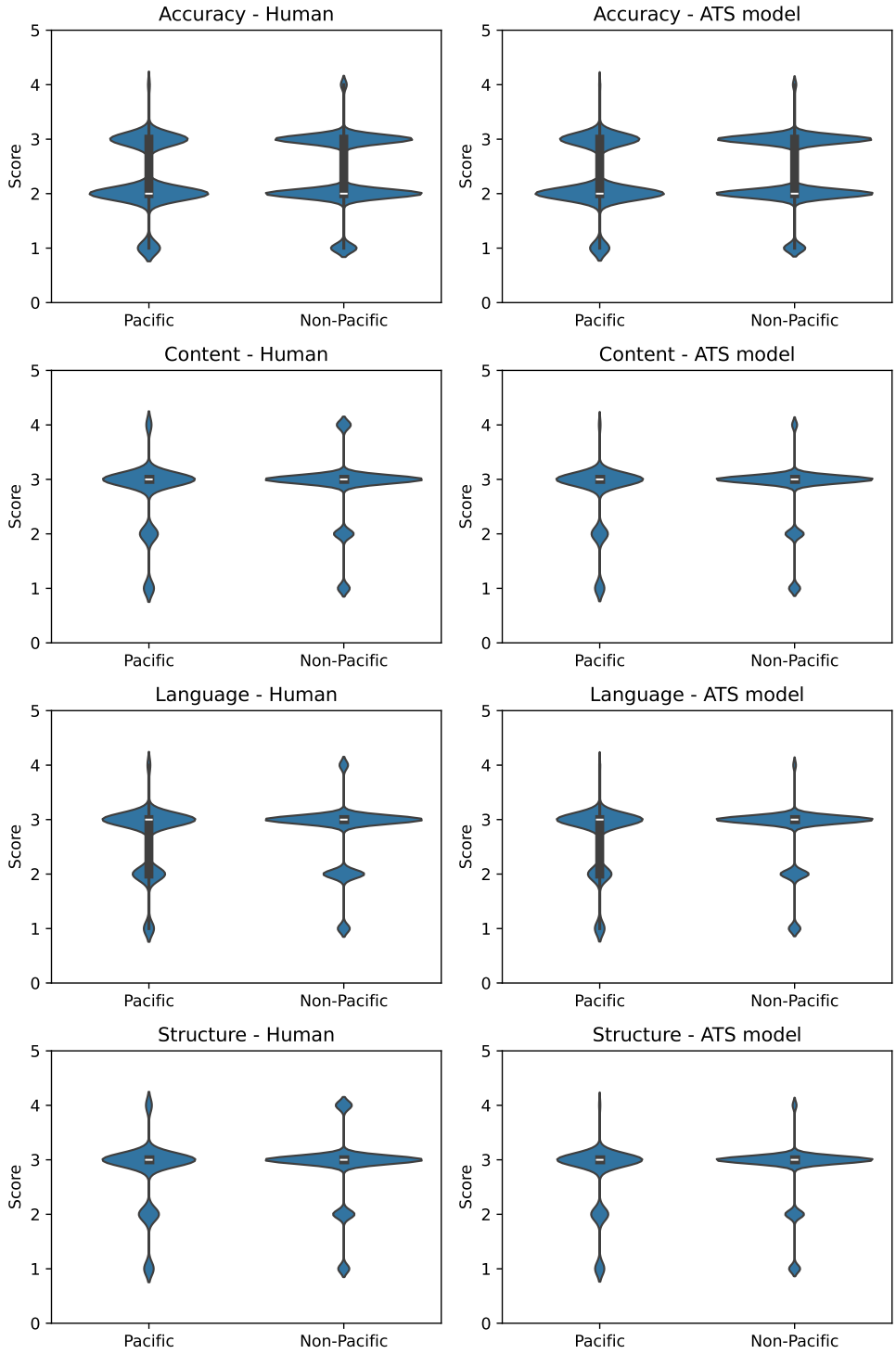


FIGURE C2 **Score distribution of human and the NZQA ATS model, by Pacific/Non-Pacific in 2025 AE1**

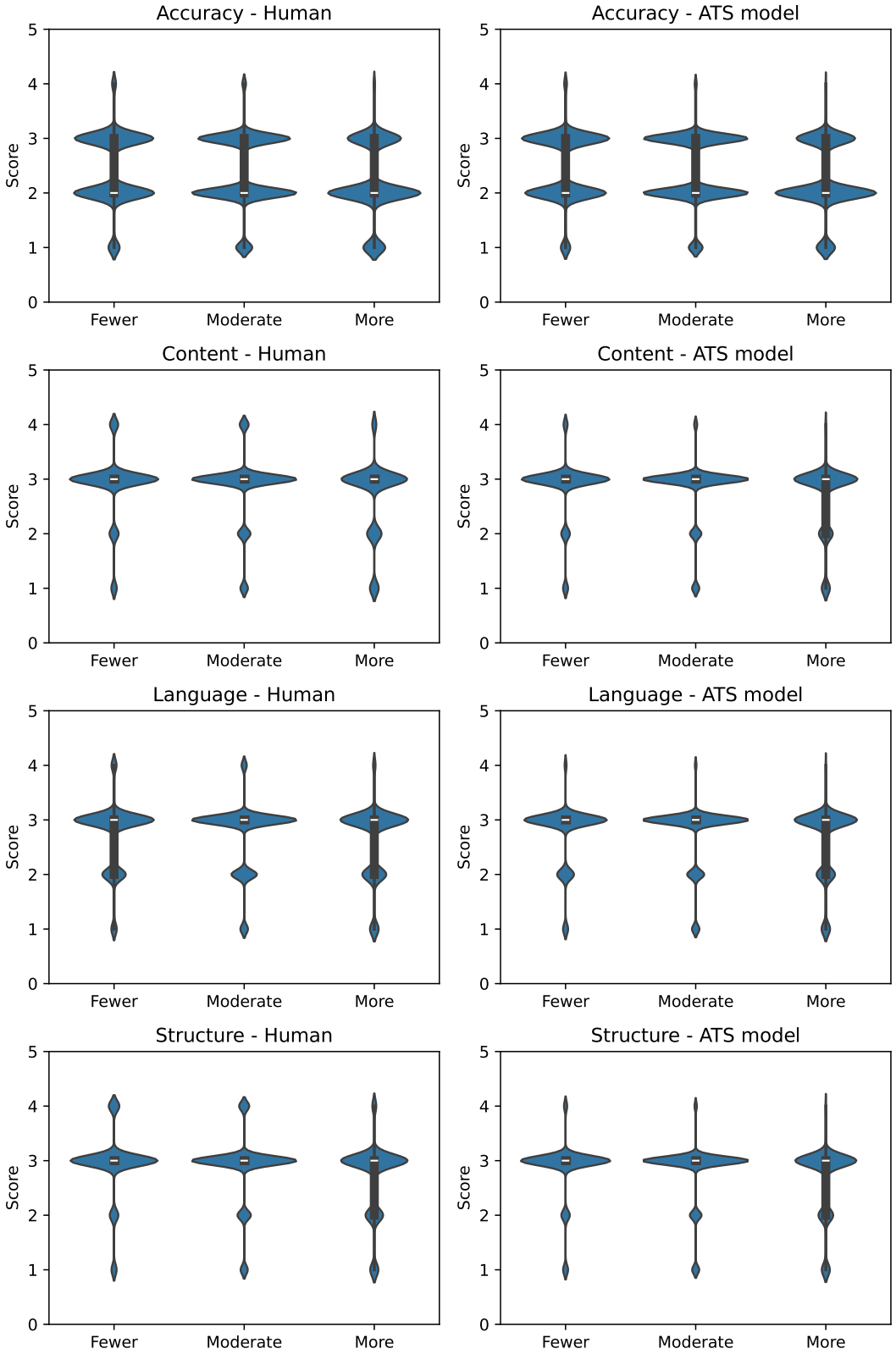


FIGURE C3 **Score distribution of human and the NZQA ATS model, by EQI groups in 2025 AE1**

APPENDIX D
Pairwise summary statistics for subgroup bias analysis in 2025 AE1

Task	Group variable	group_a	group_b	n_a	n_b	machine_ mean_diff	human_ mean_diff	gap_diff
W1Q1	Gender	Female	Male	4479	6289	0.475	0.482	-0.007
W1Q2	Gender	Female	Male	4377	6044	0.375	0.444	-0.069
W2Q1	Gender	Female	Male	3251	4988	0.436	0.422	0.014
W2Q2	Gender	Female	Male	3163	4810	0.271	0.279	-0.007
W1Q1	Māori	Maori	Non-Maori	2870	7913	-0.38	-0.351	-0.029
W1Q2	Māori	Maori	Non-Maori	2751	7685	-0.389	-0.388	-0.002
W2Q1	Māori	Maori	Non-Maori	2353	5905	-0.122	-0.15	0.028
W2Q2	Māori	Maori	Non-Maori	2259	5733	-0.322	-0.311	-0.011
W1Q1	Pacific	Pacific	Non-Pacific	1083	9700	-0.693	-0.687	-0.006
W1Q2	Pacific	Pacific	Non-Pacific	1041	9395	-1.039	-0.931	-0.108
W2Q1	Pacific	Pacific	Non-Pacific	1065	7193	-0.237	-0.245	0.008
W2Q2	Pacific	Pacific	Non-Pacific	1022	6970	-0.546	-0.469	-0.077
W1Q1	EQI Group	Moderate	More	6339	1948	0.912	0.85	0.062
W1Q1	EQI Group	Moderate	Fewer	6339	2102	-0.182	-0.16	-0.022
W1Q1	EQI Group	More	Fewer	1948	2102	-1.094	-1.01	-0.084
W1Q2	EQI Group	Moderate	More	6132	1848	1.18	1.095	0.085
W1Q2	EQI Group	Moderate	Fewer	6132	2065	-0.202	-0.189	-0.013
W1Q2	EQI Group	More	Fewer	1848	2065	-1.382	-1.284	-0.098
W2Q1	EQI Group	Moderate	More	4772	1645	0.614	0.61	0.005
W2Q1	EQI Group	Moderate	Fewer	4772	1554	0.02	0.019	0.001
W2Q1	EQI Group	More	Fewer	1645	1554	-0.594	-0.591	-0.003
W2Q2	EQI Group	Moderate	More	4627	1578	0.793	0.744	0.048
W2Q2	EQI Group	Moderate	Fewer	4627	1508	-0.141	-0.147	0.006
W2Q2	EQI Group	More	Fewer	1578	1508	-0.934	-0.892	-0.042

NZCER
Rangahau Mātauranga o Aotearoa

Rangahau Mātauranga o Aotearoa | New Zealand Council for Educational Research

www.nzcer.org.nz
facebook.com/nzcer
@NZCER
www.linkedin.com/company/new-zealand-council-for-educational-research