NZQA Automated Text Scoring (ATS) review

Review of model marking performance, quality assurance, and retraining approach for AE2 2025 and AE1 2026

Jessie Dong
with Charles Darr

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.0105

NZQA Automated Text Scoring (ATS) review

Review of model marking performance, quality assurance, and retraining approach for AE2 2025 and AE1 2026

Jessie Dong
with Charles Darr

August 2026

NZCER
Rangahau Mātauranga o Aotearoa

NZQA
Mana Tohu Mātauranga o Aotearoa
New Zealand Qualifications Authority

He ihirangi | Contents

NZQA management comment 1

Executive summary 2

1. Purpose and scope 3
- Purpose of this review 3
- Scope of this review 3
- Organisation of the report 3

2. Methods 4
- Data 4
- Documents 5
- Model performance metrics 5
- Rounding 6
- Model behaviour investigation 7
- Bias/fairness investigation 7

3. ATS performance results and findings 8
- 3.1 Performance statistics 8
- 3.2 Summary statistics 11
- 3.3 ATS performance across assessment events 14
- 3.4 Model behaviour analysis 15
- 3.5 Bias and fairness investigation by subgroups 21

4. Quality assurance review findings 25
- 4.1 Quality assurance framework and selection methodology for human marking 25
- 4.2 Comments 25

5. Retraining approach review findings 27
- An illustrative sampling approach informed by a published study 27

6. Discussion 29

7. Recommendations 30
- Fitness for the purpose of marking Literacy Writing 32405 assessments 30
- Transferability to other assessment standards (beyond Literacy Writing) 30
- Retraining and human marking assurance approach 30

References 32
- Appendix 1 33
- Appendix 2 34
- Appendix 3 35
- Appendix 4 37
- Appendix 5 38
- Appendix 6 41

Tables

Table 1	Summary of the reviewed data in AE2 2025	4
Table 2	Summary of the reviewed data in AE1 2026	5
Table 3	Evaluation metrics for reviewing model performance	6
Table 4	ATS–human agreement performance statistics by task and rubric element in AE1 2026 (rubric score scale 0–4)	9
Table 5	ATS–human agreement performance statistics of the total scores by task in AE1 2026 (rubric score scale 0–16)	9
Table 6	ATS–human agreement performance statistics by task and rubric element in AE2 2025 (rubric score scale 0–4)	10
Table 7	ATS–human agreement performance statistics of the total scores by task in AE2 2025 (rubric score scale 0–16)	11
Table 8	Summary statistics for ATS and human rubric element scores by task in AE1 2026	12
Table 9	Summary statistics for ATS and human total scores by task in AE1 2026	12
Table 10	Summary statistics for ATS and human rubric element scores by task in AE2 2025	13
Table 11	Summary statistics for ATS and human total scores by task in AE2 2025	14
Table 12	Spearman correlation among human-marked rubric element scores in AE1 2026	20
Table 13	Spearman correlation among rounded ATS rubric element scores in AE1 2026	20
Table 14	Spearman correlation among unrounded ATS rubric element scores in AE1 2026	20
Table 15	Mean ATS–human agreement performance statistics by demographic subgroup in AE1 2026 across all tasks and rubric elements	22
Table 16	Pairwise subgroup mean score differences for ATS and human marking by task in AE1 2026	24

Figures

Figure 1	QWK of ATS marking vs. human scores across assessment events	15
Figure 2	Confusion matrices for the ATS model score vs. human marks by rubric element across all tasks in AE1 2026	16
Figure 3	Total score distribution comparison between human marks and ATS scores in AE1 2026.	17
Figure 4	Rubric element score distribution comparison between human marks and ATS scores across assessment tasks in AE1 2026	19
Figure 5	Comparison of score density distributions from human marks and the ATS model by rubric element in AE1 2026	21
Figure 6	Score distributions by gender for human and ATS models in AE1 2026	23

NZQA management comment

This second NZCER review of NZQA's Automated Text Scoring (ATS) model uses results from the Literacy Writing Co-requisite Event 2 of 2025, and Event 1 of 2026. We thank NZCER for the thorough analysis and rigorous scrutiny they have applied in assessing the model's performance.

NZCER's first review of the ATS model, completed in early 2026, gave NZQA the necessary assurance to proceed with using our ATS model for Assessment Event 1 of 2026 for marking Literacy Writing. This second review was designed to confirm the performance results post Assessment Event 1 of 2026, as well as review our quality assurance approach (which scripts are identified for further human marking), and our approach for retraining the model after each event.

We are pleased that the review confirms our model approach, quality assurance arrangements, and retraining processes are fit for purpose. The findings also reinforce our confidence that the model performs consistently across gender, ethnicity, and socioeconomic groups, with no substantial differences identified in ATS–human agreement.

We welcome the recommendations NZCER has identified for refinements to our quality assurance and model retraining approaches, and we are considering how to action these for Literacy Writing future marking rounds.

Executive summary

NZQA commissioned this review to provide an independent evaluation of its Automated Text Scoring (ATS) model and to support decisions about its continued use in marking assessments for the co-requisite Literacy Writing standard 32405. The review examined model marking performance across Assessment Event 2, 2025 (AE2 2025) and Assessment Event 1, 2026 (AE1 2026) alongside the quality assurance, human marking, and retraining processes, and considered the potential transferability of the ATS approach to other assessment standards.

Overall, the ATS model is considered fit for continued operational use in marking the Literacy Writing assessment.

In AE1 2026, when ATS was used as the production marking model, it demonstrated strong operational performance. At the rubric element level, exact match between ATS and human scores was 79.6%, exact-or-adjacent agreement reached 99.8%, and the mean Quadratic Weighted Kappa (QWK) was 0.763. At the total score level, 84.5% of ATS scores were within one point of the corresponding human score, and the QWK was 0.872. Performance in AE2 2025 was similarly strong.

On average, ATS scored slightly lower than human markers. The mean ATS–human score difference was −0.07 with a small effect size in AE1 2026. The ATS model also appears to differentiate somewhat less strongly than human markers among the "Content", "Language", and "Structure" rubric elements, which is an area worth continuing to monitor. The subgroup analysis did not identify substantial differences in ATS–human agreement across demographic groups, although ongoing monitoring is recommended.

The quality assurance and retraining approaches are generally sound. Recommended refinements include strengthening validity checks, automating the identification of empty responses, broadening the selection of high-scoring scripts, and considering a random or stratified-random sample to provide a more representative estimate of overall performance. Retraining data should be sampled separately by week and question, with continued attention to less common scores and weaker performing rubric elements. Given the high ATS–human agreement for scripts scored near the cut scores, the size and human marking approach for this cohort could also be reviewed.

While the broader ATS approach may be transferable to other assessment standards, this would need to be established separately for each standard through model redesign, appropriate training data, rubric alignment, validation, fairness and robustness testing, and ongoing human oversight.

The improvements observed from AE2 2025 to AE1 2026 are consistent with a beneficial retraining effect, although differences in questions, student populations, human marking, and dataset selection may also have contributed. The datasets used in this review were purposively selected quality assurance samples focused on more challenging marking cases, so the reported statistics should not be treated as estimates of performance across all responses.

1. Purpose and scope

Purpose of this review

This review is intended to support NZQA's use of its Automated Text Scoring (ATS) model for marking writing assessments associated with the co-requisite Literacy Writing (32405) assessment standard.

The review provides an independent evaluation of model marking performance for Assessment Event 2 in 2025 (AE2)[1] and Assessment Event 1 in 2026 (AE1).[2] It also considers the marking assurance plan and retraining approach and provides recommendations to support future implementation.

Scope of this review

The scope of this review includes:

- Review of the ATS model marking performance for AE2 2025 and AE1 2026, including performance statistics for total scores and for each rubric element. Key statistical indicators include agreement with human marking, score consistency, and distribution patterns.
- Bias and fairness analysis for AE2 2025 and AE1 2026, including subgroup performance where data are available.
- Reference to the AE1 2025 review findings, with commentary on comparative marking performance where valid comparisons can be made. Some comparisons may be limited by differences in score scale range between review periods.
- Review of the model retraining approach for AE1 2026.
- Review of NZQA's marking assurance plan, including the sampling process, human marking approach, and use of blind or non-blind human marking in this context.
- Advice on the potential transferability of the ATS model approach to other assessment standards, including standards associated with the new qualification.
- Recommendations on the ATS model's marking performance, fitness for ongoing use in Literacy Writing assessment, retraining approach, and human marking assurance approach.

Organisation of the report

This review report begins by outlining its purpose and scope (Section 1) and explaining the review methods, including the datasets and the analytical approach (Section 2). It then presents findings on model performance (Section 3), quality assurance (Section 4), and retraining approaches (Section 5), followed by a discussion of the key findings (Section 6) and recommendations (Section 7).

1 The ATS model was not used for operational marking in AE2 2025; instead, it was used as part of the quality assurance process.
2 The ATS model was used for operational marking in AE1 2026.

2. Methods

Data

Two sets of data were reviewed. One set is from AE2 2025, and the other is from AE1 2026.

Each assessment event involves two parallel assessment forms (Week 1 and Week 2). Each form includes two extended response items (Question 1 and Question 2). Each question was scored using a rubric comprising four elements: **Accuracy, Content, Language, and Structure** (New Zealand Qualifications Authority [NZQA], 2026, p. 3). Below is a brief description of what the rubric elements assess:

Accuracy: Correct use of text conventions (e.g., grammar, punctuation, spelling), including correct use of sentence struct0ures, tenses, singular/plural forms, pronouns, and verb forms.

Content: Quality of ideas/information and their development, focusing on what is said rather than how it is said.

Language: Appropriateness for purpose and audience (e.g., sentence types and variety, word choice, and language features).

Structure: Overall flow of ideas across the text as a whole (e.g., connection and coherence).

Each rubric element is scored on a scale of 0–4, where 0 indicates no evidence and 4 indicates strong evidence. Separate ATS models are used to score Question 1 and Question 2 because of their different text lengths.

Both datasets contain students' National Student Number (NSN), question number, week number, human-marked scores, ATS-marked scores, low-confidence flags and safety flags produced by ATS, and high-level demographic flags.

A summary of the cleaned data is provided in Tables 1 and 2. Note: empty responses were excluded from the datasets.

TABLE 1 **Summary of the reviewed data in AE2 2025**

Event	Week	Question	Count
AE2 2025	Week 1	Question 1	9,309
AE2 2025	Week 1	Question 2	9,071
AE2 2025	Week 2	Question 1	5,568
AE2 2025	Week 2	Question 2	5,298
Total			29,246

TABLE 2 **Summary of the reviewed data in AE1 2026**

Event	Week	Question	Count
AE1 2026	Week 1	Question 1	8,686
AE1 2026	Week 1	Question 2	8,425
AE1 2026	Week 2	Question 1	7,774
AE1 2026	Week 2	Question 2	7,521
Total			32,406

It is worth noting that the two datasets were selected differently. The AE1 2026 dataset included three cohorts: scripts flagged as low confidence by the ATS model; a sample of scripts scored 1 or 4; and scripts with scores near the expected cut scores. In contrast, the AE2 2025 dataset was selected using scores from a different production model and did not include a cohort targeting scores of 1 and 4. In both cases, human markers conducted blind marking, meaning the markers were not shown the automated scores.

Documents

In addition to the dataset, the documents below were also reviewed:

- Quality assurance framework for ATS
- Methodology for the selection of scripts for human marking for literacy writing unit standard 32405—Assessment Event 1, 2026
- Approach for retraining the ATS model—Event 1, 2026
- Retraining ATS model: augmented data selection and results.

Model performance metrics

The key evaluation metrics used to review model performance are exact agreement, adjacent agreement, Pearson's correlation coefficient, root mean squared error, Cohen's kappa, Quadratic Weighted Kappa, and Matthews' correlation coefficient. A description of each metric is provided in Table 3.

Summary statistics including mean, median, standard deviation, differences between model means and human means, and effect size were calculated to support the model performance review.

TABLE 3 **Evaluation metrics for reviewing model performance**

Metric	Description	Range	Score type used
Exact agreement	The percentage of essays where the model score is the same as the human score.	[0,1]	Rounded score
Adjacent agreement	The percentage of essays where the human and the model scores differ by exactly one point.		
Pearson's correlation coefficient (PCC)	A measure of the linear relationship between model scores and human scores.	[–1,1] +1: perfect positive linear relationship 0: no linear relationship –1: perfect negative linear relationship	Scores with two decimal places
Root mean squared error (RMSE)	A measure of the typical difference between model scores and human scores, expressed in the original score units.	[0, ∞) Lower values indicate smaller differences between model and human scores	Scores with two decimal places
Cohen's kappa (CK)	A measure of agreement between model scores and human scores after accounting for agreement expected by chance. It can be sensitive to skewed score distributions.	[–1,1] Higher values indicate better agreement +1: perfect agreement between model scores and human scores 0: agreement is no better than chance <0: indicates systematic disagreement	Rounded score
Quadratic weighted kappa (QWK)[3]	A weighted measure of agreement between model scores and human scores after accounting for chance. It penalises big differences more harshly than smaller ones.		
Matthews' correlation coefficient (MCC)	A correlation-based measure of how closely model and human scores align, balancing correct and incorrect scoring across the entire score range.		

Rounding

The NZQA ATS model outputs scores as continuous numbers with two decimal places, whereas human marks are discrete integers. To enable model–human agreement analysis, the ATS model scores were rounded to the nearest whole number for some of the metrics, as listed in Table 3.

This review used the round-half-up method, where .5 values are always rounded up to the next whole number. Rounded scores were also used in the visualisations, including the confusion matrices and density distributions, to ensure consistency and clearer interpretation of score patterns.

However, we retained the original scores with two decimal places for calculating PCCs to avoid artificial discretisation.[4] Original scores were also retained for RMSE calculations, as specified in Table 3.

3 QWK is one of the most widely used measures for human–AI agreement in the Automated Essay Scoring field.
4 Rounding continuous outputs can reduce variance and distort relationships between variables, leading to artificially lower correlations.

Model behaviour investigation

To strengthen the performance review, model behaviour was analysed by investigating outliers with large model–human score differences, reviewing score distributions, and examining intra-rubric correlations to better understand how ATS models differ from human markers.

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 listed in Table 3 was calculated to assess model–human agreement within each group 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.

3. ATS performance results and findings

3.1 Performance statistics

AE1 2026

At the rubric element level, the ATS model showed strong overall agreement with human marking in AE1 2026. Across the four tasks, 79.6% of ATS scores exactly matched the human score, and 99.8% were either exact or adjacent. Results were reasonably consistent across tasks, with a mean QWK of 0.763, PCC of 0.8, RMSE of 0.455, and MCC of 0.535. “Accuracy” was generally the weakest rubric element: its QWK ranged from 0.645 to 0.705, whereas “Content”, “Language”, and “Structure” more often achieved QWK values above 0.75. The performance statistics by task and rubric for AE1 2026 are presented in Table 4.

Exact agreement was lower when the four rubric scores were added into a total score, as expected, because the total has a wider 0–16 range and therefore more opportunities for small discrepancies. Exact total score agreement was 40.1%, but 84.5% of total scores were within one point of the human total.

The high total score PCC (0.887) and QWK (0.872), together with an RMSE of about 1.20 points, indicate that the model generally ranked scripts similarly to human markers even when it did not reproduce the exact total score. The performance statistics of the total score by task are presented in Table 5.

TABLE 4 **ATS–human agreement performance statistics by task and rubric element in AE1 2026 (rubric score scale 0–4)**

Task	Rubric	n	Exact	Adjacent	Exact+Adj	PCC	RMSE	CK	QWK	MCC
Wk1Q1	Accuracy	8,686	0.676	0.323	0.998	0.712	0.575	0.444	0.645	0.445
	Content	8,686	0.835	0.162	0.997	0.790	0.418	0.523	0.766	0.530
	Language	8,686	0.873	0.126	0.999	0.828	0.361	0.587	0.808	0.588
	Structure	8,686	0.834	0.165	0.999	0.811	0.412	0.526	0.775	0.534
	mean		***0.805***	***0.194***	***0.998***	***0.785***	***0.442***	***0.520***	***0.749***	***0.524***
Wk1Q2	Accuracy	8,425	0.691	0.308	0.998	0.720	0.561	0.403	0.653	0.417
	Content	8,425	0.855	0.142	0.997	0.825	0.394	0.559	0.812	0.565
	Language	8,425	0.840	0.158	0.998	0.826	0.409	0.541	0.799	0.542
	Structure	8,425	0.811	0.187	0.998	0.807	0.442	0.489	0.775	0.498
	mean		***0.799***	***0.199***	***0.998***	***0.795***	***0.452***	***0.498***	***0.760***	***0.506***
Wk2Q1	Accuracy	7,774	0.698	0.300	0.998	0.751	0.554	0.485	0.681	0.491
	Content	7,774	0.766	0.227	0.993	0.763	0.507	0.482	0.725	0.502
	Language	7,774	0.838	0.162	1.000	0.831	0.404	0.581	0.801	0.582
	Structure	7,774	0.819	0.181	0.999	0.828	0.428	0.549	0.791	0.552
	mean		***0.780***	***0.218***	***0.998***	***0.793***	***0.473***	***0.524***	***0.750***	***0.532***
Wk2Q2	Accuracy	7,521	0.719	0.279	0.998	0.770	0.537	0.492	0.705	0.497
	Content	7,521	0.845	0.152	0.997	0.859	0.407	0.632	0.838	0.635
	Language	7,521	0.822	0.177	0.999	0.845	0.426	0.598	0.818	0.599
	Structure	7,521	0.813	0.186	0.999	0.839	0.438	0.584	0.812	0.589
	mean		***0.800***	***0.199***	***0.998***	***0.828***	***0.452***	***0.577***	***0.793***	***0.580***
Overall	***mean***		***0.796***	***0.202***	***0.998***	***0.800***	***0.455***	***0.530***	***0.763***	***0.535***

TABLE 5 **ATS–human agreement performance statistics of the total scores by task in AE1 2026 (rubric score scale 0–16)**

Task	Score	n	Exact	Adj	Exact+Adj	PCC	RMSE	CK	QWK	MCC
Wk1Q1	Total	8,686	0.426	0.436	0.862	0.877	1.16	0.202	0.863	0.211
Wk1Q2	Total	8,425	0.386	0.443	0.83	0.881	1.251	0.148	0.856	0.157
Wk2Q1	Total	7,774	0.392	0.451	0.843	0.886	1.191	0.213	0.877	0.219
Wk2Q2	Total	7,521	0.4	0.444	0.845	0.903	1.196	0.218	0.893	0.224
Overall	***mean***		***0.401***	***0.444***	***0.845***	***0.887***	***1.200***	***0.195***	***0.872***	***0.203***

AE2 2025

The ATS model also performed well in AE2 2025. At the rubric element level, exact agreement was 72.7%, exact-or-adjacent agreement was 99.7%, PCC was 0.808, QWK was 0.762, and RMSE was 0.536. Wk1Q1 was the weakest task, with a mean QWK of 0.701, driven particularly by lower agreement for "Content". Wk2Q2 was the strongest task, with a mean QWK of 0.799. These results suggest that the model was usually within one score point of the human mark, but its ability to make exact agreements depended on the task and rubric. The full set of performance statistics by task and rubric is presented in Table 6.

For total scores, exact agreement was 36.2% and exact-or-adjacent agreement was 76.4%. The mean task level QWK was 0.857, and the PCC was 0.884, indicating strong ordinal agreement. The RMSE of 1.513 shows that the size of total score differences was nevertheless meaningful. Overall, the results were good, but exact scoring and calibration were weaker than in AE1 2026. The performance statistics of the total score by task are presented in Table 7.

TABLE 6 **ATS–human agreement performance statistics by task and rubric element in AE2 2025 (rubric score scale 0–4)**

Task	Rubric	n	Exact	Adjacent	Exact+Adj	PCC	RMSE	KP	QWK	MCC
Wk1Q1	Accuracy	9,309	0.731	0.267	0.998	0.786	0.527	0.525	0.734	0.526
	Content	9,309	0.659	0.333	0.992	0.672	0.606	0.271	0.625	0.286
	Language	9,309	0.788	0.209	0.997	0.789	0.472	0.450	0.741	0.452
	Structure	9,309	0.720	0.278	0.998	0.770	0.539	0.352	0.703	0.364
	mean		***0.725***	***0.272***	***0.996***	***0.754***	***0.536***	***0.400***	***0.701***	***0.407***
Wk1Q2	Accuracy	9,071	0.689	0.307	0.997	0.802	0.571	0.493	0.746	0.498
	Content	9,071	0.747	0.250	0.997	0.827	0.522	0.438	0.787	0.448
	Language	9,071	0.755	0.242	0.997	0.836	0.508	0.460	0.792	0.464
	Structure	9,071	0.744	0.254	0.998	0.829	0.520	0.451	0.791	0.457
	mean		***0.734***	***0.263***	***0.997***	***0.824***	***0.530***	***0.461***	***0.779***	***0.467***
Wk2Q1	Accuracy	5,568	0.700	0.298	0.997	0.799	0.559	0.510	0.742	0.514
	Content	5,568	0.750	0.246	0.996	0.811	0.517	0.418	0.776	0.435
	Language	5,568	0.766	0.231	0.997	0.827	0.497	0.475	0.790	0.481
	Structure	5,568	0.721	0.277	0.998	0.823	0.539	0.395	0.773	0.413
	mean		0.734	0.263	0.997	0.815	0.528	***0.450***	***0.770***	***0.461***
Wk2Q2	Accuracy	5,298	0.663	0.332	0.995	0.802	0.594	0.474	0.750	0.480
	Content	5,298	0.717	0.277	0.995	0.838	0.554	0.423	0.805	0.435
	Language	5,298	0.752	0.244	0.996	0.858	0.516	0.489	0.822	0.495
	Structure	5,298	0.725	0.271	0.996	0.851	0.542	0.456	0.817	0.464
	mean		***0.714***	***0.281***	***0.996***	***0.837***	***0.552***	***0.461***	***0.799***	***0.469***
Overall	***mean***		***0.727***	***0.270***	***0.997***	***0.808***	***0.536***	***0.443***	***0.762***	***0.451***

TABLE 7 **ATS–human agreement performance statistics of the total scores by task in AE2 2025 (rubric score scale 0–16)**

Task	Rubric	n	Exact	Adjacent	Exact+Adj	PCC	RMSE	KP	QWK	MCC
Wk1Q1	Total	9,309	0.347	0.413	0.76	0.854	1.457	0.185	0.825	0.185
Wk1Q2	Total	9,071	0.369	0.402	0.771	0.894	1.486	0.222	0.869	0.224
Wk2Q1	Total	5,568	0.366	0.403	0.769	0.887	1.505	0.215	0.859	0.217
Wk2Q2	Total	5,298	0.366	0.389	0.755	0.9	1.604	0.223	0.875	0.225
Overall	***mean***		***0.362***	***0.402***	***0.764***	***0.884***	***1.513***	***0.211***	***0.857***	***0.213***

3.2 Summary statistics

AE1 2026

The AE1 2026 summary statistics show a small but consistent tendency for ATS to assign lower scores than human markers. This pattern was also seen in the last review for AE1 2025. Across rubric elements, the overall ATS mean was 2.64, compared with 2.71 for human marking, giving a mean difference of –0.07 and a small effect size of -0.11. The largest rubric level difference occurred for "Language" in Wk1Q2, where the mean difference was –0.17, and the effect size was –0.28. "Content" in Wk2Q1 showed almost no systematic difference.

The same pattern is visible in total scores. The overall ATS mean total was 10.57, compared with 10.86 for human marking, a difference of –0.29 and an effect size of –0.12. All four tasks had negative differences, ranging from –0.18 to –0.43. Underscoring therefore remained present, but it was smaller than in AE2 2025, which ranged from –0.45 to –0.57.

The machine-to-human standard deviation ratio[5] was 0.954 for AE1 2026. This is close to 1, indicating that the overall ATS score spread was broadly similar to the human spread, though slightly narrower. The model's remaining issue is therefore better characterised as mild conservative calibration and local boundary errors.

5 This is calculated as machine_sd / human_sd

TABLE 8 **Summary statistics for ATS and human rubric element scores by task in AE1 2026**

		ATS			Human				
Task	**Rubric**	**median**	**mean**	**sd**	**median**	**mean**	**sd**	**mean_ difference**	**effect_ size**
Wk1Q1	Accuracy	2.45	2.42	0.56	2.00	2.44	0.66	−0.02	−0.03
	Content	2.83	2.74	0.57	3.00	2.86	0.62	−0.12	−0.20
	Language	2.87	2.78	0.57	3.00	2.87	0.57	−0.10	−0.17
	Structure	3.03	2.92	0.60	3.00	2.96	0.62	−0.04	−0.07
	mean	***2.80***	***2.71***	***0.58***	***2.75***	***2.78***	***0.62***	***−0.07***	***−0.12***
Wk1Q2	Accuracy	2.30	2.28	0.59	2.00	2.31	0.64	−0.03	−0.05
	Content	2.89	2.79	0.63	3.00	2.88	0.62	−0.09	−0.14
	Language	2.77	2.67	0.62	3.00	2.85	0.61	−0.17	−0.28
	Structure	2.80	2.70	0.62	3.00	2.85	0.64	−0.14	−0.23
	mean	***2.69***	***2.61***	***0.61***	***2.75***	***2.72***	***0.63***	***−0.11***	***−0.17***
Wk2Q1	Accuracy	2.38	2.35	0.60	2.00	2.40	0.67	−0.05	−0.08
	Content	2.85	2.73	0.63	3.00	2.73	0.70	0.00	0.01
	Language	2.84	2.72	0.62	3.00	2.82	0.63	−0.10	−0.16
	Structure	2.98	2.85	0.65	3.00	2.88	0.66	−0.04	−0.06
	mean	***2.76***	***2.66***	***0.62***	***2.75***	***2.71***	***0.67***	***−0.05***	***−0.07***
Wk2Q2	Accuracy	2.28	2.26	0.63	2.00	2.28	0.68	−0.02	−0.02
	Content	2.90	2.74	0.69	3.00	2.80	0.71	−0.06	−0.09
	Language	2.79	2.65	0.68	3.00	2.76	0.69	−0.11	−0.16
	Structure	2.82	2.68	0.69	3.00	2.76	0.71	−0.08	−0.11
	mean	2.70	2.58	0.67	2.75	2.65	0.70	−0.07	−0.09
Overall	***mean***	***2.74***	***2.64***	***0.62***	***2.75***	***2.71***	***0.65***	***−0.07***	***−0.11***

TABLE 9 **Summary statistics for ATS and human total scores by task in AE1 2026**

		ATS			Human				
Task	**Score**	**median**	**mean**	**sd**	**median**	**mean**	**sd**	**mean_ difference**	**effect_ size**
Wk1Q1	Total	11.19	10.85	2.25	11.00	11.13	2.14	−0.28	−0.13
Wk1Q2	Total	10.77	10.44	2.40	11.00	10.88	2.20	−0.43	−0.19
Wk2Q1	Total	11.06	10.65	2.44	11.00	10.83	2.32	−0.18	−0.08
Wk2Q2	Total	10.83	10.33	2.64	11.00	10.59	2.50	−0.26	−0.10
Overall	***mean***	***10.96***	***10.57***	***2.44***	***11.00***	***10.86***	***2.29***	***−0.29***	***−0.12***

AE2 2025

The reported mean differences between model and human scores for AE2 2025 indicate systematic underscoring by ATS. Across the 16 rubric dimensions for all tasks, the mean machine–human difference was –0.13. All model–human differences were negative, and the overall effect size was –0.17. The largest reported rubric level difference between model and human was for "Structure" in Wk1Q1, with a mean difference of –0.20 and an effect size of –0.30.

Underscoring was more pronounced for total scores. The overall mean total score difference was –0.51, corresponding to an effect size of –0.18, and task level differences ranged from –0.45 to –0.57. Small negative differences across rubric elements therefore accumulated into a more noticeable difference in total scores.

The mean machine-to-human standard deviation ratio was 1.027. This is relatively close to 1, indicating that ATS reproduced the overall human score spread well in AE2 2025, despite its conservative scoring.

TABLE 10 **Summary statistics for ATS and human rubric element scores by task in AE2 2025**

		ATS			Human				
Task	**Rubric**	**median**	**mean**	**sd**	**median**	**mean**	**sd**	**mean_ difference**	**effect_ size**
Wk1Q1	Accuracy	2.64	2.51	0.66	3.00	2.60	0.70	–0.09	–0.13
	Content	2.84	2.68	0.62	3.00	2.81	0.71	–0.13	–0.20
	Language	2.92	2.74	0.64	3.00	2.86	0.61	–0.11	–0.18
	Structure	2.94	2.76	0.65	3.00	2.95	0.68	–0.20	–0.30
	mean	***2.84***	***2.67***	***0.64***	***3.00***	***2.81***	***0.68***	***–0.13***	***–0.20***
Wk1Q2	Accuracy	2.50	2.37	0.75	3.00	2.50	0.74	–0.12	–0.17
	Content	2.99	2.77	0.79	3.00	2.87	0.75	–0.10	–0.13
	Language	2.88	2.66	0.78	3.00	2.82	0.72	–0.16	–0.21
	Structure	2.95	2.72	0.79	3.00	2.82	0.75	–0.09	–0.12
	mean	***2.83***	***2.63***	***0.78***	***3.00***	***2.75***	***0.74***	***–0.12***	***–0.16***
Wk2Q1	Accuracy	2.44	2.30	0.71	3.00	2.43	0.74	–0.13	–0.18
	Content	2.91	2.69	0.75	3.00	2.82	0.74	–0.14	–0.18
	Language	2.86	2.62	0.75	3.00	2.77	0.71	–0.15	–0.20
	Structure	2.98	2.73	0.77	3.00	2.89	0.77	–0.16	–0.21
	mean	***2.80***	***2.59***	***0.74***	***3.00***	***2.73***	***0.74***	***–0.14***	***–0.19***
Wk2Q2	Accuracy	2.38	2.22	0.80	2.00	2.33	0.77	–0.11	–0.14
	Content	2.99	2.67	0.90	3.00	2.76	0.82	–0.09	–0.10
	Language	2.86	2.54	0.87	3.00	2.70	0.78	–0.15	–0.19
	Structure	2.98	2.65	0.91	3.00	2.75	0.82	–0.10	–0.12
	mean	***2.80***	***2.52***	***0.87***	***2.75***	***2.63***	***0.80***	***–0.11***	***–0.14***
Overall	***mean***	2.82	2.60	0.76	2.94	2.73	0.74	–0.13	–0.17

TABLE 11 **Summary statistics for ATS and human total scores by task in AE2 2025**

		ATS			Human				
Task	**Rubric**	**median**	**mean**	**sd**	**median**	**mean**	**sd**	**mean_ difference**	**effect_ size**
Wk1Q1	Total	11.36	10.70	2.52	12.00	11.23	2.33	−0.53	−0.22
Wk1Q2	Total	11.35	10.53	3.07	12.00	11.00	2.68	−0.48	−0.17
Wk2Q1	Total	11.21	10.34	2.93	11.00	10.91	2.66	−0.57	−0.20
Wk2Q2	Total	11.24	10.08	3.43	11.00	10.54	2.91	−0.45	−0.14
Overall	***mean***	***11.29***	***10.41***	***2.99***	***11.50***	***10.92***	***2.64***	***−0.51***	***−0.18***

3.3 ATS performance across assessment events

Across the three assessment events—AE1 2025, AE2 2025, and AE1 2026—at the rubric element level, the mean QWK increased from 0.717 in AE1 2025 to 0.762 in AE2 2025, then remained stable at 0.763 in AE1 2026, as shown in Figure 1. It is worth noting that comparisons with AE1 2025 should be treated with caution because its QWK was calculated on a 1–4 score range, whereas the later assessment events used a 0–4 range. Thus, the QWK in AE1 2025 should be interpreted only as an estimated reference point.

Several other measures improved between AE2 2025 and AE1 2026: rubric level exact agreement increased from 72.7% to 79.6%, while RMSE decreased from 0.536 to 0.455. At the total score level, exact agreement increased from 36.2% to 40.1%, exact-or-adjacent agreement increased from 76.4% to 84.5%, QWK increased from 0.857 to 0.872, and RMSE decreased from 1.513 to about 1.20. The average rubric level machine–human difference decreased from approximately −0.13 to −0.07, while the total score difference decreased from −0.51 to −0.29, indicating less underscoring. Taken together, the improvements are consistent with a beneficial retraining effect. However, the review cannot separate retraining effects from differences in questions, candidate responses, dataset selection, or human marking between assessment events.

It is worth noting that not every measure improved. Rubric level PCC decreased slightly from 0.808 in AE2 2025 to 0.800 in AE1 2026, while total score PCC was broadly stable. Total score unweighted kappa and MCC also decreased slightly despite improvement in QWK and RMSE.

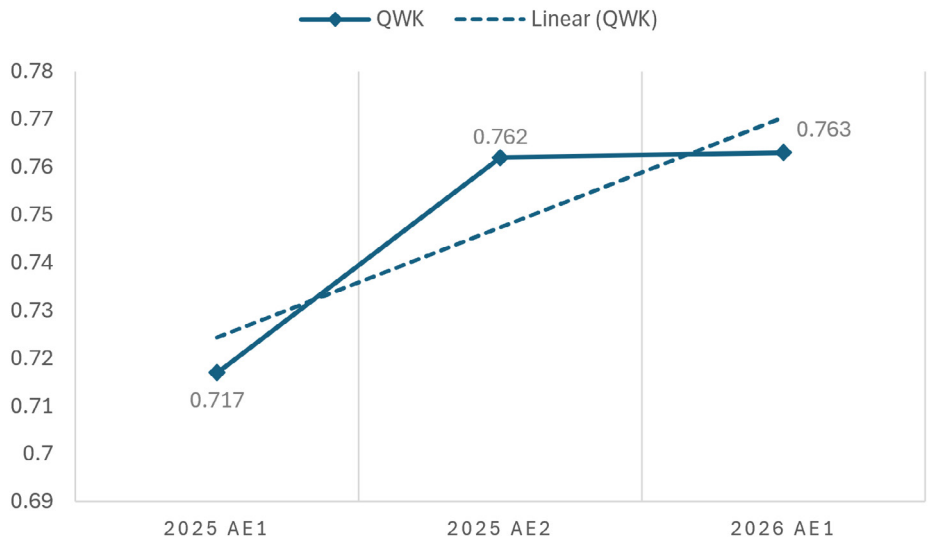


FIGURE 1 **QWK of ATS marking vs. human scores across assessment events**[6]

3.4 Model behaviour analysis

Outliers

Figure 2 shows the confusion matrices for the machine and human score by rubric element across all tasks. Most ATS scores that did not exactly match with human scores were adjacent to the human scores. Using a difference of two or more rubric points as a working definition of an outlier, there were approximately 254 outlying comparisons among 129,624 rubric level comparisons, or about 0.2%. Only 19 comparisons differed by three or four points. These counts confirm that large disagreements were rare, though they remain important because they represent cases where the model and human marker interpreted the response quite differently. It might be worth investigating these rare cases, especially those without a low_confidence flag. Three of these rare cases in 2026 AE1 were identified and will be forwarded to NZQA for further review.

6 Note: the QWK for 2025 AE1 was calculated using the score range 1–4, while the others use the score range 0–4.

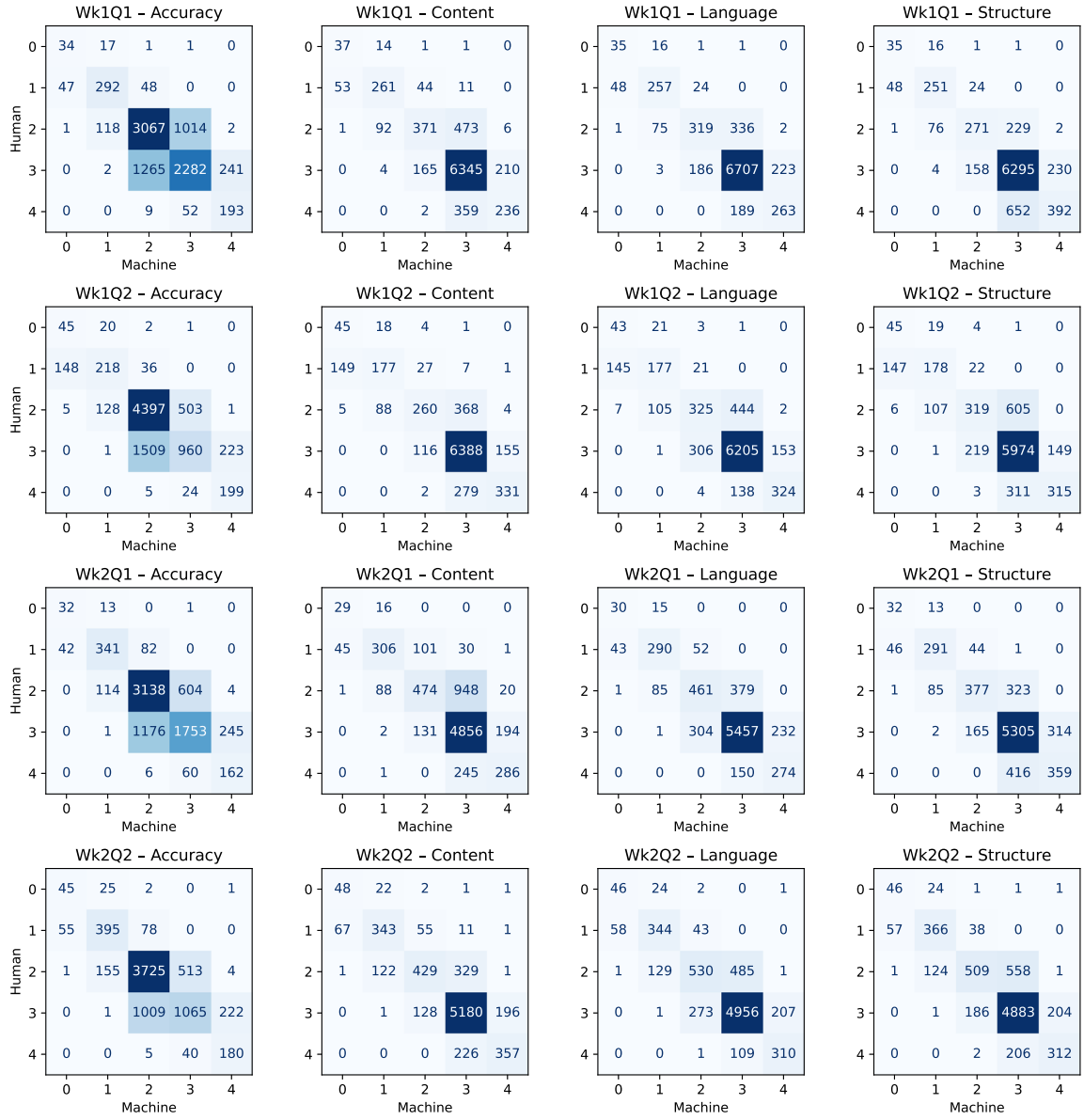


FIGURE 2 **Confusion matrices for the ATS model score vs. human marks by rubric element across all tasks in AE1 2026**

The confusion matrices figure for AE2 2025 is provided in Appendix 1.

Score distributions

As shown in Figure 3, the ATS and human total score distributions aligned broadly, but the ATS distributions are more concentrated around a total score of 11. Human scores are more broadly distributed across 11 and 12, with more scripts receiving 12. This is consistent with the summary statistics showing that ATS total scores were, on average, 0.29 points lower than human totals.

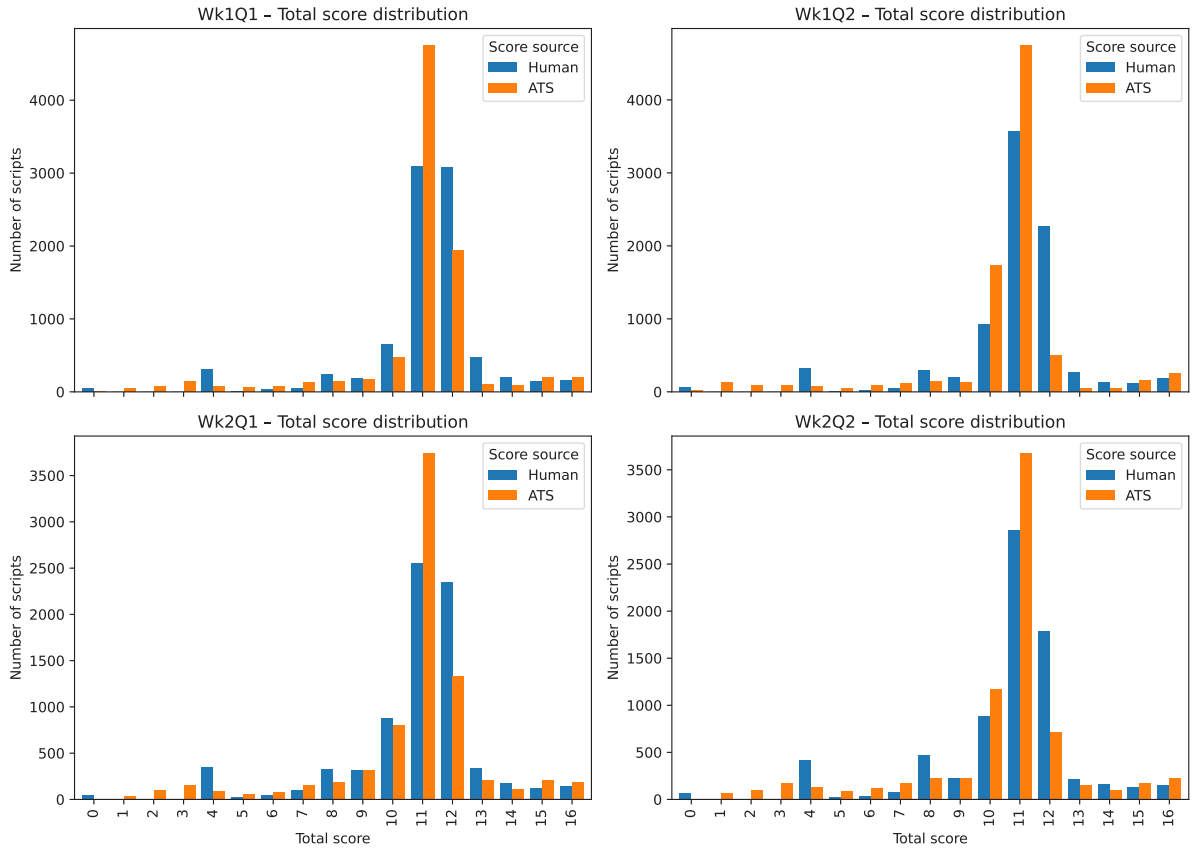


FIGURE 3 **Total score distribution comparison between human marks and ATS scores in AE1 2026.**

This concentration around 11 may help explain why QWK agreement is high while exact total score agreement remains comparatively modest. However, it is not easy to target the exact reasons as this is more likely an issue associated with accumulating the rubric element scores.

The machine-to-human standard deviation ratios were close to 1 in both assessment events, indicating that ATS broadly reproduced the overall spread of human scores. The ratio changed from 1.027 in AE2 2025 to 0.954 in AE1 2026. This suggests a shift from slightly greater to slightly lower dispersion than human scoring. The total score distribution figure for AE2 2025 is provided in Appendix 2.

At rubric level, both human and ATS scores are concentrated at scores 2 and 3, but the ATS distributions show less differentiation among “Content”, “Language”, and “Structure”, as shown in Figure 4. For the ATS model, these three rubric curves overlap closely around score 3 across all four tasks. The human distributions show more variation in the relative concentration and spread of these rubric elements, indicating that human markers distinguish among them more strongly.

The model behaviour on the “Accuracy” rubric element is different from the other rubric elements in both human and ATS marking, with substantial density at score 2. However, the patterns between ATS and human marking are very similar. Overall, the distribution at rubric level suggests that the model captures the broad scoring for each rubric but tends to produce a more uniform profile across rubric elements.

The Spearman correlation analysis supports the visual evidence that the ATS model differentiates less among rubric elements than human markers. For human scores, the correlations among "Content", "Language", and "Structure" range from 0.61 to 0.63, as shown in Table 12. For ATS scores, the corresponding correlations are much higher, ranging from 0.90 to 0.93, as shown in Table 13.[7] Scores of the "Accuracy" element are less strongly related to the other rubric elements for both marking sources, but its ATS correlations (0.57–0.60) are still higher than the human correlations (0.42–0.48). Interestingly, correlations for both human scores and ATS scores across rubric elements increased for AE1 2026 compared to AE1 2025. So, it is not easy to tell if this reflects a model retraining effect, a task-specific effect, or a rubric-based marking effect.

This may warrant future attention if the rubric level marking intends to evaluate different aspects of the writing and tends to allow more variance among rubric element scores. In this case, a useful future direction for improvement would be to strengthen rubric-specific training examples, particularly those in which performance is deliberately uneven across "Content", "Language", and "Structure". Model evaluation should then not only test agreement within each rubric element but also assess whether the pattern of correlations across rubric elements becomes more similar to that seen in human marking.

7 Spearman correlations were calculated using rounded ATS scores to enable direct comparison with the human marks and to maintain consistency with the analysis conducted in the previous AE1 2025 review. Table 14 reports the correlations based on the original unrounded ATS scores as supplementary information.

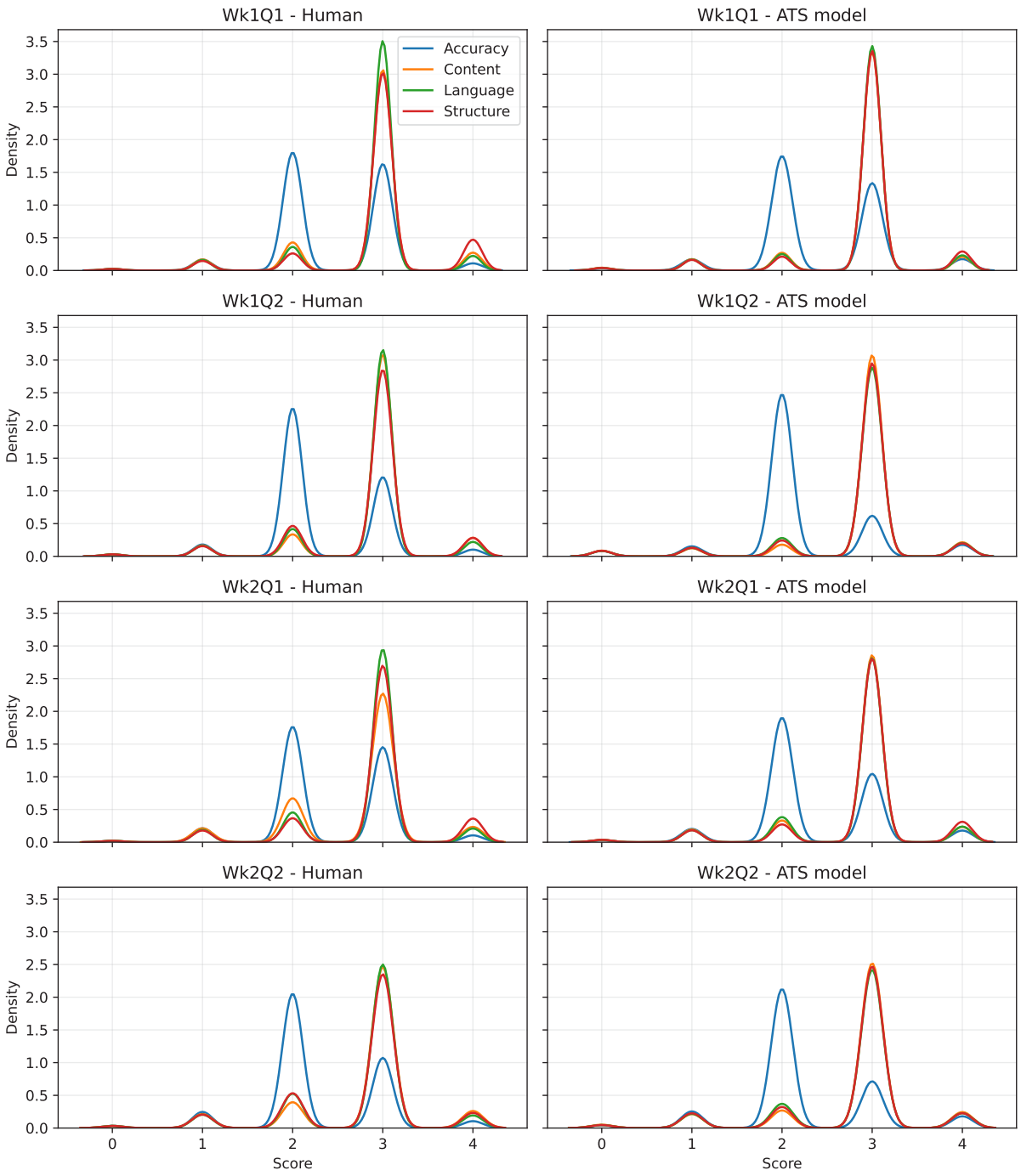


FIGURE 4 **Rubric element score distribution comparison between human marks and ATS scores across assessment tasks in AE1 2026**

The rubric score distribution figure for AE2 2025 is provided in Appendix 3.

TABLE 12 **Spearman correlation among human-marked rubric element scores in AE1 2026**

	Accuracy_human	Content_human	Language_human	Structure_human
Accuracy_human	1.00	0.42	0.48	0.43
Content_human	0.42	1.00	0.63	0.62
Language_human	0.48	0.63	1.00	0.61
Structure_human	0.43	0.62	0.61	1.00

TABLE 13 **Spearman correlation among rounded ATS rubric element scores in AE1 2026**

	Accuracy_ATS	Content_ATS	Language_ATS	Structure_ATS
Accuracy_ats	1.00	0.57	0.60	0.60
Content_ats	0.57	1.00	0.93	0.90
Language_ats	0.60	0.93	1.00	0.92
Structure_ats	0.60	0.90	0.92	1.00

TABLE 14 **Spearman correlation among unrounded ATS rubric element scores in AE1 2026**

	Accuracy_ats	Content_ats	Language_ats	Structure_ats
Accuracy_ats	1.00	0.45	0.69	0.53
Content_ats	0.45	1.00	0.80	0.90
Language_ats	0.69	0.80	1.00	0.79
Structure_ats	0.53	0.90	0.79	1.00

The density plots show close alignment between human and ATS score distributions for “Content”, “Language”, and “Structure” across the four tasks. The main peak patterns and the curves generally overlap well. “Accuracy” rubric element is the clearest exception: the relative density at scores 2 and 3 differs more noticeably between the ATS and human score distributions, with the model tending to have more scores of 2 and fewer scores of 3 in several tasks.

This aligns with the performance statistics, which show that “Accuracy” is the weakest rubric element. Its QWK ranges from 0.645 to 0.705 and its PCC from 0.712 to 0.770, with RMSE between 0.537 and 0.575. “Content”, “Language”, and “Structure” generally achieve higher QWK values and lower errors. Improvement work should therefore focus on the evidence used to distinguish “Accuracy” scores of 2 and 3 in future retraining.

The set of Spearman correlation tables for AE2 2025 is presented in Appendix 4.

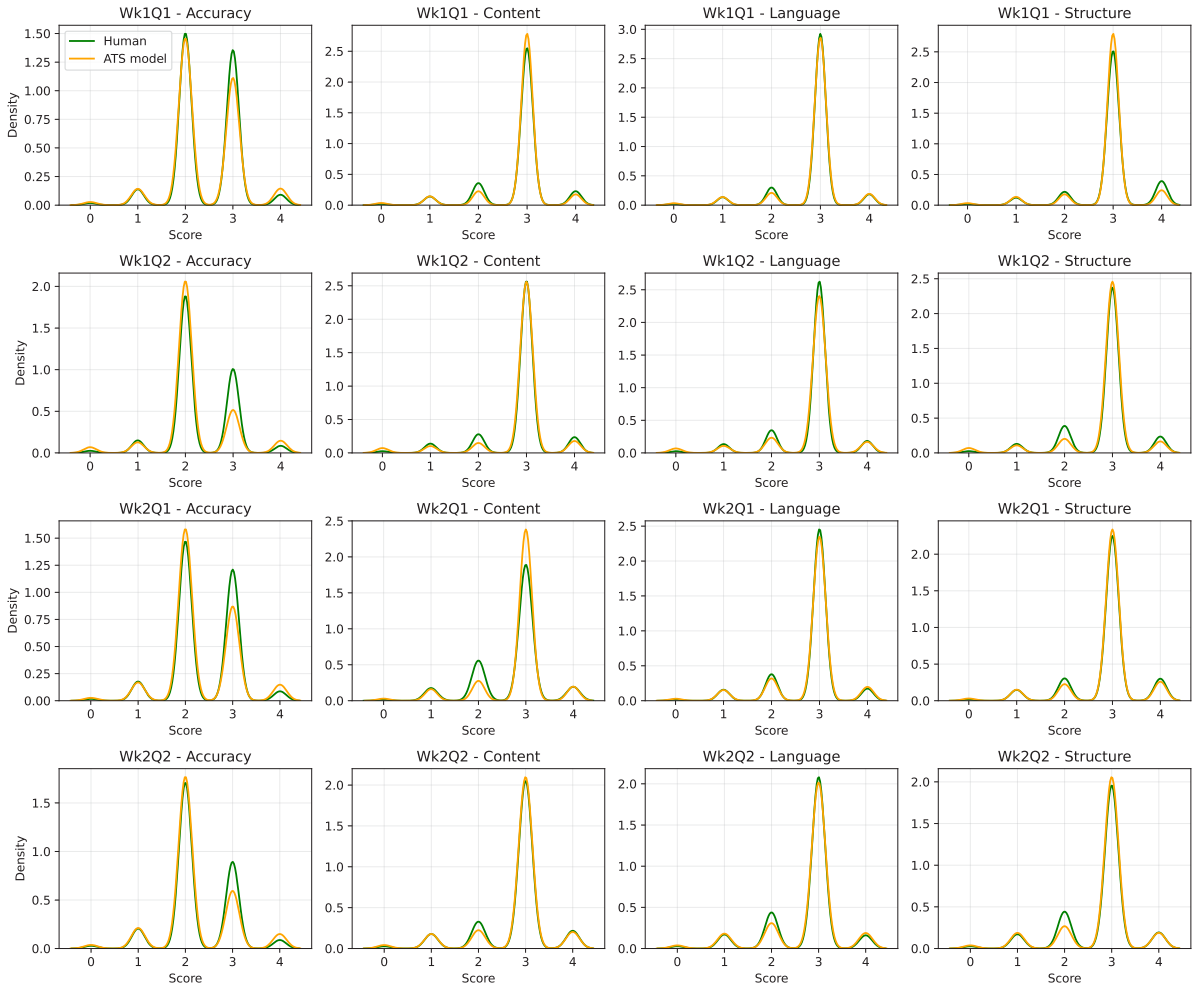


FIGURE 5 **Comparison of score density distributions from human marks and the ATS model by rubric element in AE1 2026**

3.5 Bias and fairness investigation by subgroups

The subgroup results do not show substantial additional bias introduced by ATS. The gender distributions are very similar for female, male, and gender-unknown candidates under both human and ATS marking, as shown in Table 15 and Figure 6.

The ethnicity comparisons were also similar. Exact agreement was 0.807 for Māori candidates and 0.791 for non-Māori candidates, with QWK values of 0.757 and 0.763 respectively. Pacific and non-Pacific candidates had the same exact agreement of 0.796 and QWK values of 0.764 and 0.761.

The mean absolute differences between subgroup ATS–human scoring gaps across gender, ethnicity, and EQI groups were less than 0.06, as shown in Table 16.

Based on the available descriptive evidence, ATS largely reproduced the subgroup patterns already present in human scores rather than introducing a large new disparity. Figure 6 shows the violin distribution of the scores by gender for human and ATS models in AE1 2026. Other subgroup figures for Māori/non-Māori, Pacific/Non-Pacific, and EQI Groups are available in Appendix 5. The pairwise summary statistics for AE2 2025 are presented in Appendix 6.

TABLE 15 **Mean ATS–human agreement performance statistics by demographic subgroup in AE1 2026 across all tasks and rubric elements**

Subgroup	n	Exact	Adj	Exact+Adj	PCC	RMSE	KP	QWK	MCC
Male	10,713	0.801	0.197	0.998	0.811	0.452	0.528	0.771	0.535
Female	6,512	0.801	0.198	0.998	0.798	0.449	0.529	0.759	0.536
Gender U	15,138	0.790	0.208	0.998	0.788	0.459	0.528	0.750	0.534
Māori	9,223	0.807	0.191	0.998	0.798	0.440	0.526	0.757	0.533
Non-Māori	23,140	0.791	0.207	0.998	0.800	0.460	0.530	0.763	0.535
Pacific	5,690	0.796	0.202	0.998	0.807	0.455	0.537	0.764	0.544
Non-Pacific	26,673	0.796	0.202	0.998	0.798	0.455	0.527	0.761	0.533
More	6,159	0.704	0.292	0.996	0.829	0.558	0.462	0.780	0.469
Moderate	15,135	0.735	0.262	0.997	0.801	0.528	0.431	0.755	0.440
Fewer	6,827	0.726	0.270	0.996	0.783	0.540	0.427	0.737	0.436

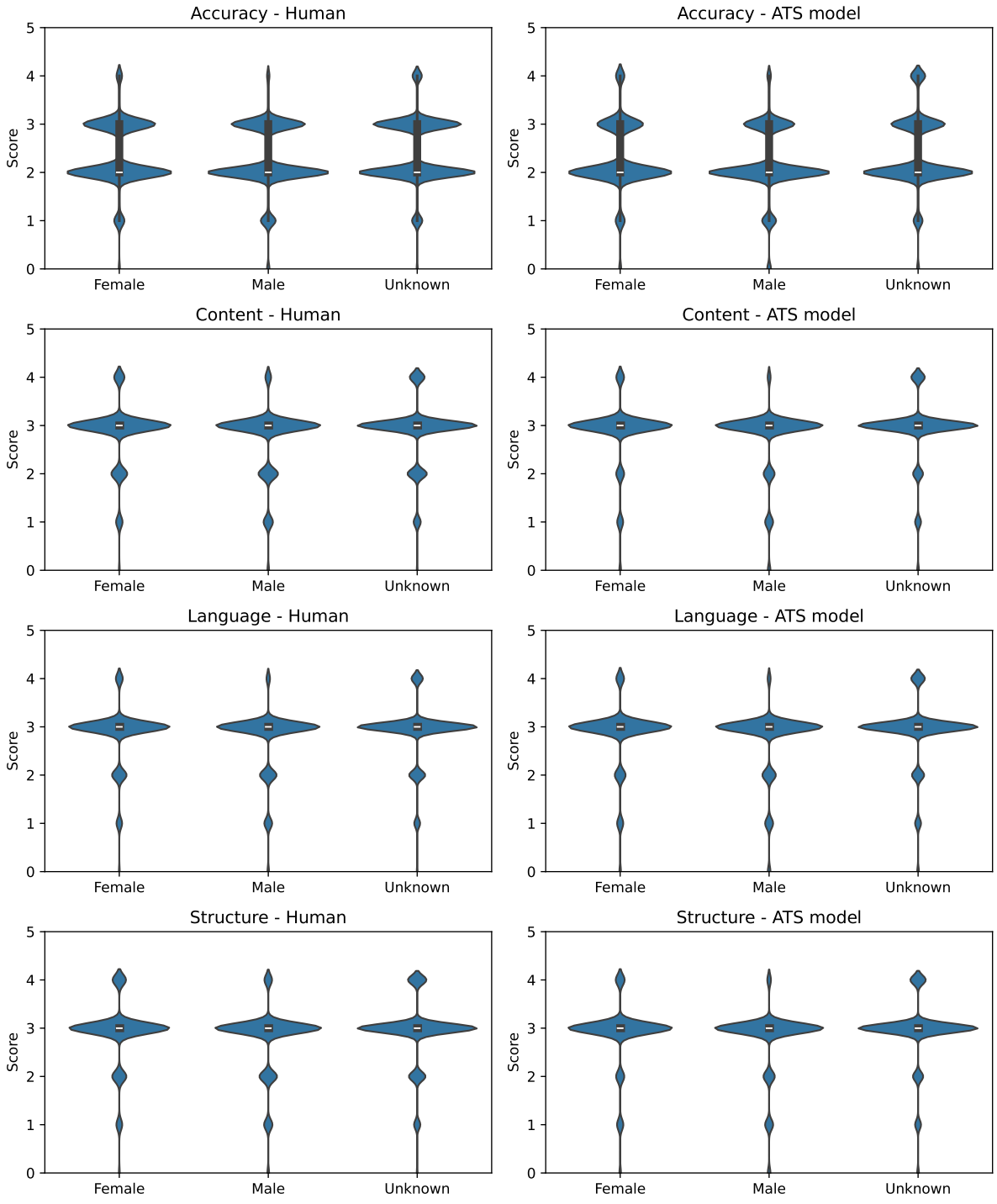


FIGURE 6 **Score distributions by gender for human and ATS models in AE1 2026**

TABLE 16 **Pairwise subgroup mean score differences for ATS and human marking by task in AE1 2026**

Assessment	Group_a	Group_b	n_a	n_b	ATS_ mean_diff	Human_ mean_diff	Gap_diff
Wk1Q1	Female	Male	1,754	2,771	0.128	0.116	0.012
Wk1Q2	Female	Male	1,698	2,652	0.101	0.101	−0.001
Wk2Q1	Female	Male	1,540	2,711	0.121	0.1	0.021
Wk2Q2	Female	Male	1,520	2,579	0.091	0.08	0.011
Wk1Q1	Māori	Non-Māori	2,379	6,296	−0.034	−0.032	−0.002
Wk1Q2	Māori	Non-Māori	2,272	6,142	−0.072	−0.055	−0.017
Wk2Q1	Māori	Non-Māori	2,331	5,432	−0.107	−0.097	−0.01
Wk2Q2	Māori	Non-Māori	2,241	5,270	−0.121	−0.093	−0.028
Wk1Q1	Pacific	Non-Pacific	1,337	7,338	−0.098	−0.071	−0.027
Wk1Q2	Pacific	Non-Pacific	1,276	7,138	−0.144	−0.12	−0.024
Wk2Q1	Pacific	Non-Pacific	1,562	6,201	−0.098	−0.072	−0.026
Wk2Q2	Pacific	Non-Pacific	1,515	5,996	−0.162	−0.124	−0.038
Wk1Q1	Fewer	Moderate	1,682	5,230	0.036	0.032	0.003
Wk1Q1	Fewer	More	1,682	1,520	0.174	0.155	0.018
Wk1Q1	Moderate	More	5,230	1,520	0.138	0.123	0.015
Wk1Q2	Fewer	Moderate	1,640	5,074	0.065	0.063	0.002
Wk1Q2	Fewer	More	1,640	1,463	0.231	0.203	0.028
Wk1Q2	Moderate	More	5,074	1,463	0.166	0.139	0.027
Wk2Q1	Fewer	Moderate	1,149	4,193	0.085	0.071	0.015
Wk2Q1	Fewer	More	1,149	2,052	0.252	0.212	0.04
Wk2Q1	Moderate	More	4,193	2,052	0.167	0.142	0.025
Wk2Q2	Fewer	Moderate	1,126	4,065	0.09	0.08	0.01
Wk2Q2	Fewer	More	1,126	1,958	0.276	0.221	0.055
Wk2Q2	Moderate	More	4,065	1,958	0.187	0.141	0.045

4. Quality assurance review findings

4.1 Quality assurance framework and selection methodology for human marking

The overall framework and selection methodology are thorough and well designed. The main quality assurance is done by top markers who mark a subset of the essays and compare human and ATS marking agreements. The selection of scripts for human marking includes three main cohorts:

- Cohort 1: All scripts that include the low_confidence or safety_concern flag. The number of scripts in this cohort is around 10% of the total, roughly 5,500.
- Cohort 2: A set of sample scripts that are marked as 1 or 4 by the ATS model. This cohort contains up to 1,000 scripts.
- Cohort 3: A cohort of scripts that are near the cut score. This cohort has around 10,000 scripts.

This approach leads to a total number of selected scripts for human marking of around 16,500, which is about 30% of the total number of scripts. These targeted cohorts are purposively selected and are appropriate for examining specific areas of risk.

Overall, the quality assurance framework and script selection methodology are sound. The three targeted cohorts address distinct risks: low-confidence or safety-related responses; less common extreme scores; and responses close to the achievement threshold. Together, they provide a strong basis for identifying cases where ATS errors would be most consequential. The refinements proposed below are intended to improve coverage and efficiency rather than to change the overall design.

4.2 Comments

Cohort 1: This cohort is based on the LLM guardrail component of the ATS model. It is a very important part of the quality assurance process that addresses the risk of off-topic scripts. However, it is unclear whether the off-topic prompt checking functionality detects content that is on-topic but not a valid piece of writing, such as prompt copy-pasting or rubric marking criteria copy-pasting.

Since this part is handled by an LLM, it is wise to be cautious about prompt injection attempts that manipulate the low-confidence flag, especially when students are aware that their writing is marked by AI models, though it is less likely that students know how the ATS guardrail works. However, this might be something worth considering.

My recommendations for this cohort are:

1. Consider enhancing the “low_confidence” flag by expanding off topic to a validity check against the prompt.
2. Consider removing empty responses from this cohort. There were 1,360 empty responses from AE2 2025 in this cohort, which is around 25% of the scripts. These scripts don’t need to be seen by ATS or human markers since they can be checked during the data cleaning or preprocessing. This will allow some space for other essays to be included in this cohort.

Cohort 2: This cohort selects scripts that received an ATS score of 1 or 4. This is designed to improve the accuracy of extreme scores, thereby enhancing fairness for students at different levels. It also addresses the imbalanced data distribution, where 1s and 4s are much less common than other scores.

The selection process is well designed. I would, however, recommend changing the selection for score 4 to 28<= total <= 32 instead of [30,32]. If a script received at least two 4s for each question, the minimum total score is 14+14=28. This subset is not just for all 4s but also helps to distinguish between 3s and 4s. I think expanding to [28,32] will provide a more balanced and richer dataset for retraining. This should provide more useful quality assurance and retraining evidence to distinguish between strong and very strong responses.

Cohort 3: This cohort selects scripts around the cut scores. This is particularly important because it affects whether a student will achieve. The overall design of this part is solid. I don't have any concerns about it.

However, given the high agreement between the human and ATS scores, it may be time to consider either reducing the size of this cohort or replacing blind marking with a human review of the ATS scores for this cohort.

5. Retraining approach review findings

The retraining process is well designed. It selects a subset of the human-marked scores and addresses the imbalanced data distribution by selecting a higher proportion (80%) from the lower and upper ends of the score range. The improvements in exact agreement, RMSE, underscoring, and assignment of score 4 in AE1 2026 are consistent with the approach having had a beneficial effect.

The main limitation is that the sampling method does not clearly ensure balanced representation of each assessment question. Q1 responses were sampled from a combined pool of Week 1 and Week 2 responses, even though Wk1Q1 and Wk2Q1 are different questions. Sampling the same proportion separately from Wk1Q1 and Wk2Q1 would ensure that both are adequately represented. The same approach should be applied to Wk1Q2 and Wk2Q2.

Based on the review findings, the following example illustrates how an alternative sampling strategy might be structured. It is not presented as a specific recommendation, but rather as one possible approach for consideration.

An illustrative sampling approach informed by a published study

Tan et al. (2022) found that training on a dataset with an approximately normal score distribution yielded better performance than training on an imbalanced dataset. This provides one example of how the distribution of training cases across score levels may influence model performance.

Using the reviewed retraining approach after AE2 2025 as an example, 5,899 responses were selected from Q1 and 6,088 from Q2. In both cases, sampling was based on the total score, measured on a 0–16 scale. As shown in Figure 3, the human-marked dataset itself is unevenly distributed. The reviewed retraining approach sought to address this by selecting 30% of responses from the middle score range and 80% from the extremes. However, when the resulting sample is considered at the rubric element level, some imbalance may remain.

There are two related considerations that could be explored:

1. Representation across tasks

One consideration is to have approximately equal numbers of responses represented across questions and weeks. For example, if the target were approximately 3,000 responses for each question in each week, an illustrative sampling structure could be:

ATS model	Question	Week	Illustrative target
Model 1: 150 words	Q1	Week 1	3,000
Model 1: 150 words	Q1	Week 2	3,000
Model 2: 250–300 words	Q2	Week 1	3,000
Model 2: 250–300 words	Q2	Week 2	3,000

2. A score-aware stratified sampling approach

The current sampling approach uses the total score on a 0–16 scale, whereas the individual rubric elements are scored from 0 to 4. Sampling only on the total score may obscure variation at the rubric element level, because different combinations of element scores can produce the same total score.

Stratifying across all rubric element score combinations would provide the most detailed coverage of different score combinations, but this is unlikely to be practical. With four rubric elements scored from 0 to 4, there are potentially (5^4 = 625) possible score profiles, many of which may contain very few or no responses.

One possible simplification would be to calculate a rounded, average score on a 0–4 scale aligned with the individual rubric elements. This average score could then be used as the primary stratification variable. Where there is evidence that a particular rubric element is more difficult for the ATS, the stratification could also incorporate one additional rubric element, such as "Accuracy". This would provide a manageable way of introducing rubric-level information into the sampling process.

Within each score stratum, the number of responses selected could then be determined using a chosen allocation strategy. The main objective would be to ensure adequate representation across the score range while avoiding excessive concentration in already well-represented categories.

For illustration, one possible allocation, inspired by Tan et al. (2022), would be an approximately normal distribution across the 0–4 holistic score scale. Assuming a target of approximately 3,000 responses per question, with a mean of 2.5 and a standard deviation of approximately 1.0, the resulting allocation might be:

Score	Illustrative target	Percentage
0	68	2.3
1	408	13.6
2	1,024	34.1
3	1,024	34.1
4	476	15.9
Total	3,000	100

This distribution provides one possible approach, and other allocations could also be considered, including more even representation across score categories or additional sampling of some score categories that perform relatively poorly.

In practice, the number of available responses at certain score levels, particularly 0, 1, or 4, may limit the extent to which a target can be achieved. The sampling proportions could therefore be adjusted according to data availability while maintaining broad representation across the score range. The key is to make the sampling score-aware, especially at the rubric element level. The most suitable sampling approach should be determined empirically by model performance after retraining.

6. Discussion

The combined findings indicate that ATS marking performance has remained strong and stable since AE1 2025, with slight improvements in some areas. Rubric-level QWK improved between AE1 and AE2 2025 and was maintained in AE1 2026. Between AE2 2025 and AE1 2026, exact agreement increased, RMSE decreased, underscoring reduced, and total score agreement improved. These results support continued use of ATS with human quality assurance.

The remaining weaknesses are relatively specific. ATS continues to score slightly conservatively overall, and the "Accuracy" rubric element is weaker than the other rubric elements. The model also appears to differentiate less strongly among "Content", "Language", and "Structure" than human markers. These issues may affect the diagnostic quality of rubric scores, even when total score agreement is high, and should guide future retraining and monitoring.

The descriptive fairness results are reassuring: no large additional subgroup bias was evident, and subgroup ATS–human scoring gaps differed by less than 0.05. However, this conclusion is limited by unequal subgroup sizes, the large unknown-gender category, and reliance on aggregate descriptive measures.

A further limitation is that the quality assurance sample is purposively concentrated around identified low-confidence marking and responses around the cut score. Performance calculated from that sample may not represent performance across all responses. A random or stratified-random sample therefore might be beneficial for future retraining and evaluation. Finally, the improvements are consistent with successful retraining but should not be attributed to retraining alone without additional evidence. Differences in questions, student populations, human marking, and dataset selection may all contribute to the observed changes in performance.

7. Recommendations

Fitness for the purpose of marking Literacy Writing 32405 assessments

Based on the findings of this review, the ATS model is considered fit-for-purpose for marking the Literacy Writing assessment. Its performance has remained stable across three live assessment events (AE1 2025, AE2 2025, and AE1 2026) with consistently high agreement between ATS and human scores. The model demonstrates reliable and consistent marking behaviour, supported by a well-designed assurance plan with appropriate human-in-the-loop oversight. The retraining approach is also sound, and performance stayed stable with some improvement following each retraining cycle.

Transferability to other assessment standards (beyond Literacy Writing)

The findings of this review provide evidence that the ATS model can maintain strong performance across different assessment events within the Literacy Writing standard. However, transferability to other assessment standards beyond Literacy Writing cannot be assessed without more information about those assessments.

The broader ATS approach, including its model architecture, training and retraining process, assurance framework, and human-in-the-loop arrangements, may potentially be transferable. However, for each new assessment standard, the model would require redesign, appropriate training data, rubric alignment, standard-specific validation, fairness and robustness testing, and ongoing human oversight before it could be considered fit for operational use.

Retraining and human marking assurance approach

The current retraining and human-marking assurance approach is well designed and has been effectively implemented. Based on the findings of this review, the following recommendations may further strengthen current practice:

Human marking

- Consider automating the detection and review of empty essays currently included in Cohort 1. This process is primarily a validation check rather than a marking exercise, so it may not require human marker resources.
- Consider expanding the Cohort 2 high-score selection range from total scores of 30–32 to 28–32. If this cohort is used for retraining, it would provide a richer dataset to help the model distinguish between scores of 3 and 4.

- Consider including a random sample of approximately 500–1,000 scripts in future quality assurance reviews to provide a broader indication of model performance across the full range of responses. The sample size could be determined based on the desired level of precision and available resources. Double marking this sample may also be considered for quality assurance.
- The findings support consideration of a reduction in the size of Cohort 3 or replacing blind marking with human review of the scores produced by the ATS for Cohort 3.

Retraining

Use a more balanced sampling approach when selecting human-marked data for retraining. In particular:

- Sample retraining data separately for Wk1Q1, Wk2Q1, Wk1Q2, and Wk2Q2 rather than combining responses by question number.
- Consider using a score-aware stratified sampling approach to ensure representation across a diverse range of rubric element scores, rather than applying a fixed sampling proportion based on the total score.

The 1,000 scripts in Cohort 2 were marked by top human markers to provide more accurate reference scores for scripts that the ATS assigned a score of 1 or 4. Consider including all scripts from this cohort in the retraining dataset. However, this is subject to any changes in the selection of the cohorts for human marking.

Ongoing monitoring and improvement

- Continue subgroup performance monitoring across gender, Māori/non-Māori, Pacific/non-Pacific, and EQI groups as part of regular quality assurance.
- Consider investigating cross-rubric correlations and rubric-specific differentiation, particularly among “Content”, “Language”, and “Structure”.
- Consider enhancing the low_confidence flag by expanding the off_topic check to a validity check against the prompt.

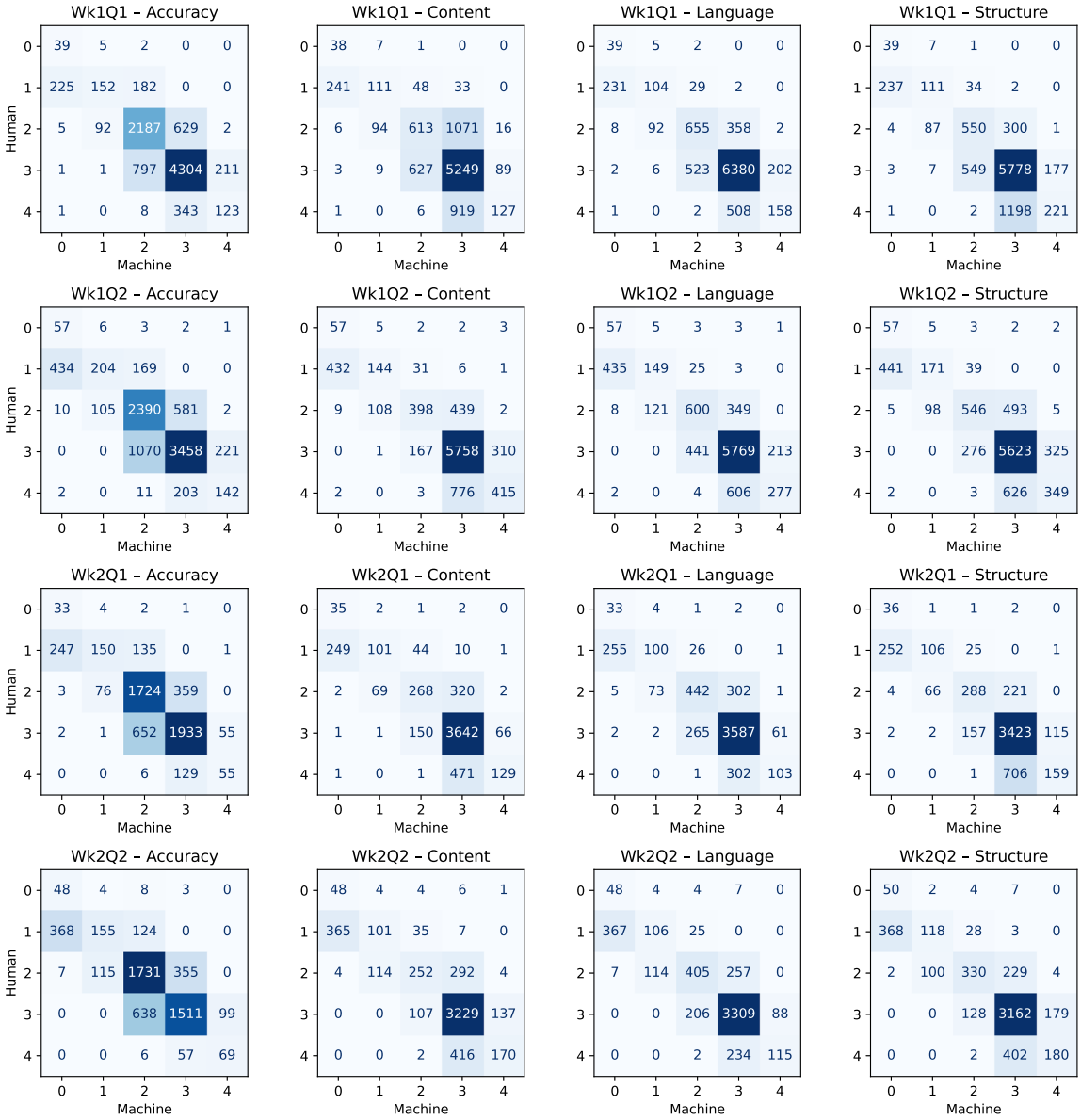
References

New Zealand Qualifications Authority. (2026). *Assessment schedule—Week 1, Term 2, 2026: Literacy: Write texts to communicate ideas and information (32405).* https://www2.nzqa.govt.nz/assets/Uploads/Assessment-Schedule-32405-Week-1-Term-2-2026.pdf

Tan, J. S., Tan, I. K. T., Soon, L. K., & Ong, H. F. (2022). *Improved automated essay scoring using Gaussian multi-class SMOTE for dataset sampling. p.647.* https://doi.org/10.5281/zenodo.6853024

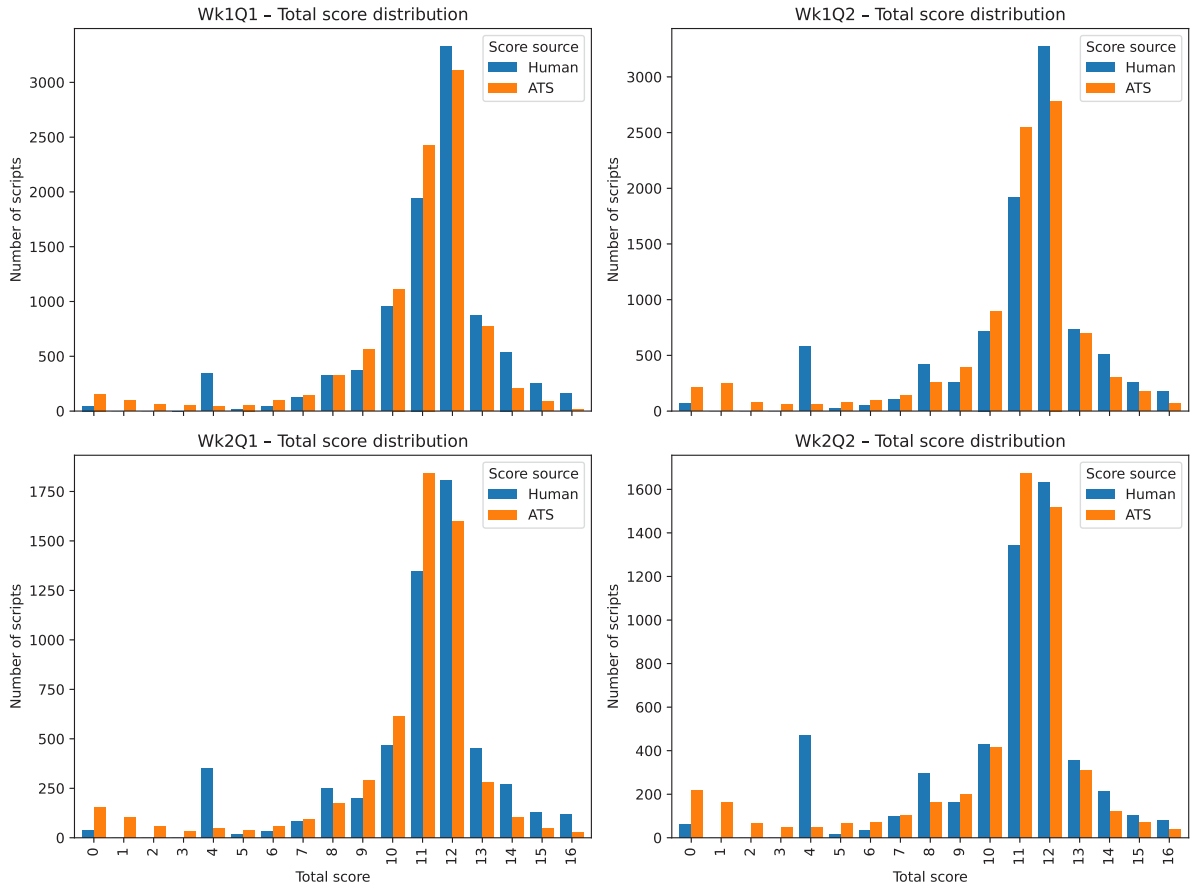
Appendix 1

Confusion matrices for the ATS model score vs. human marks by rubric element across all tasks in AE2 2025



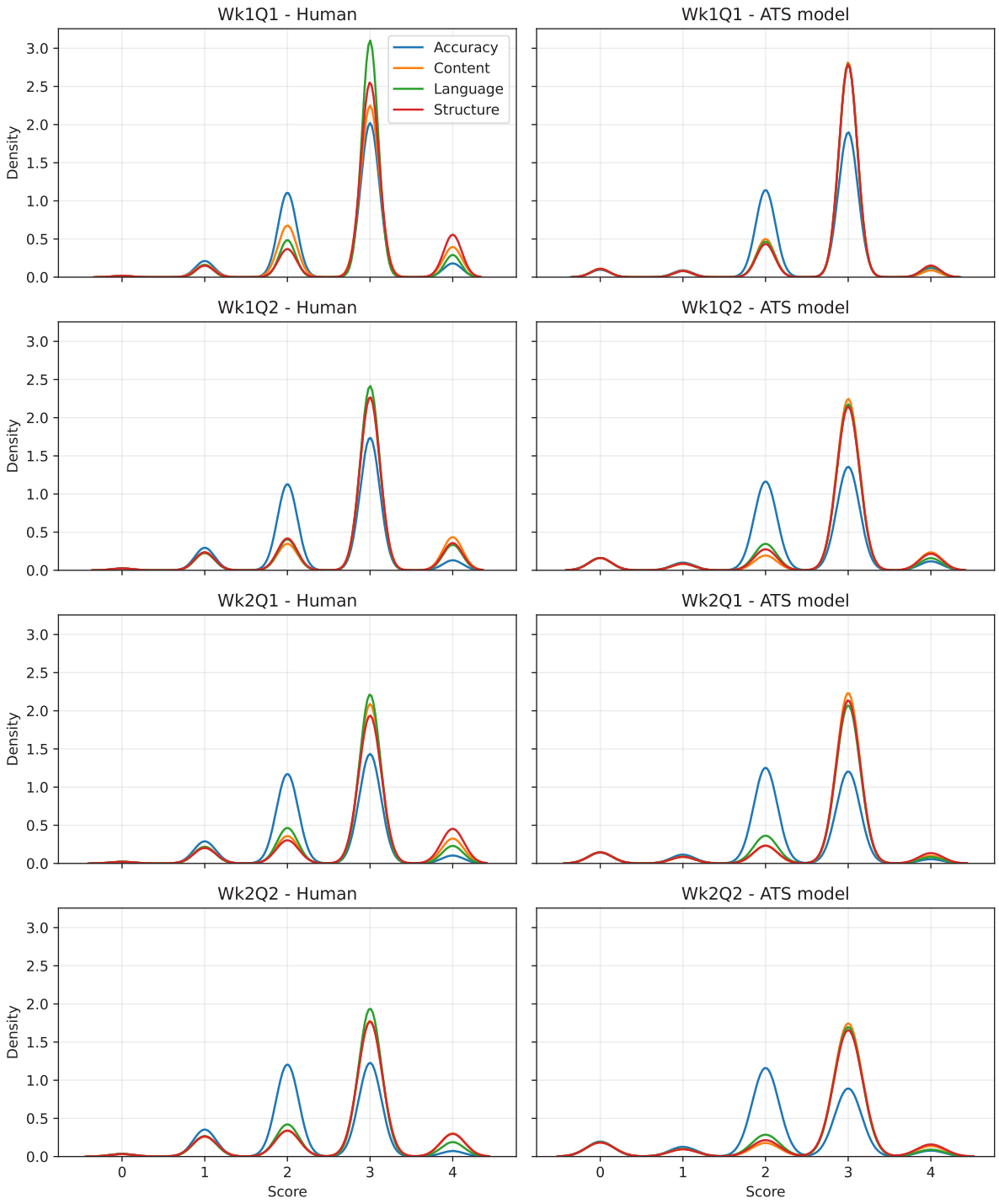
Appendix 2

Total score distribution comparison between human marks and ATS scores in AE2 2025

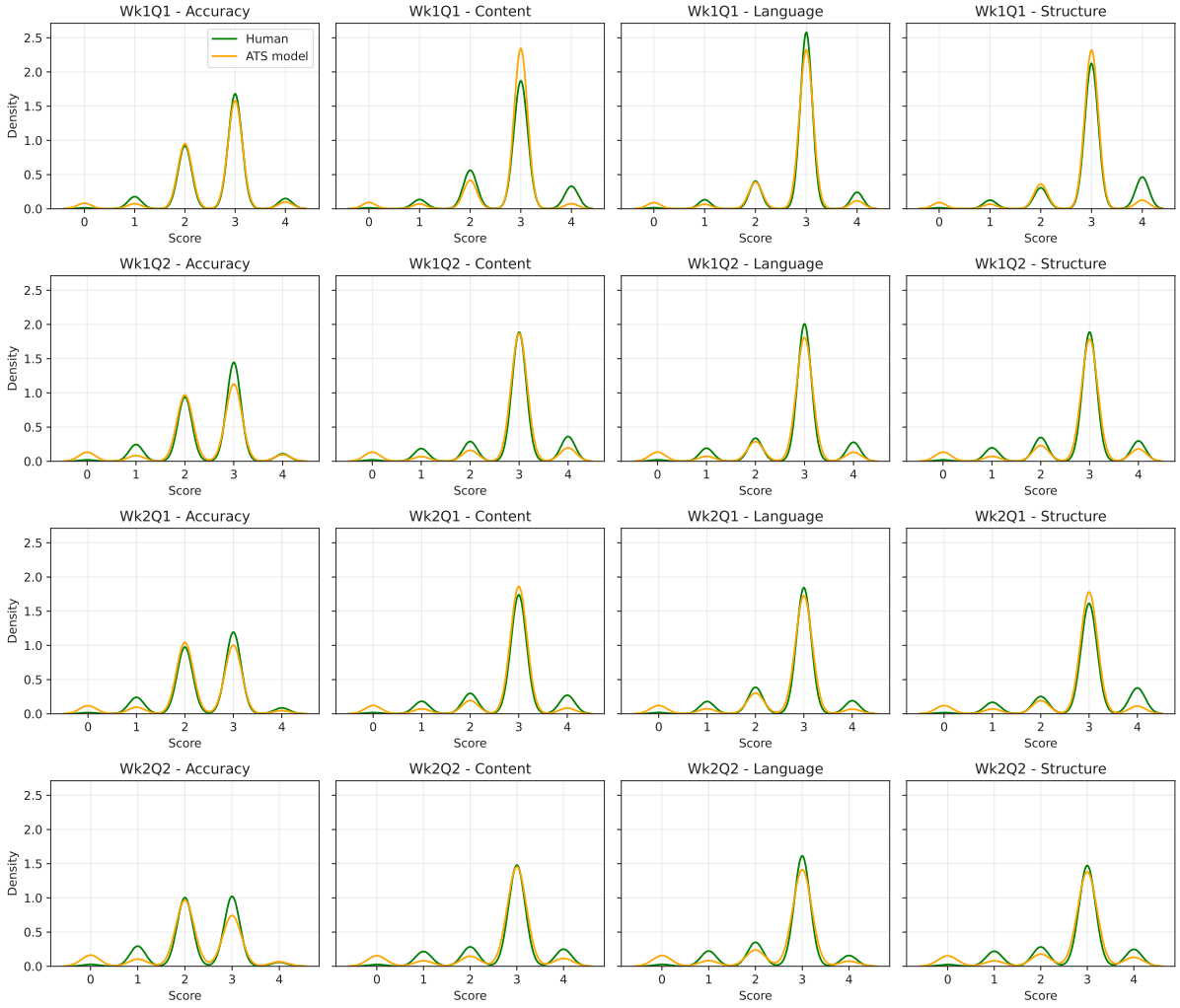


Appendix 3

a. Rubric element score distribution comparison between human marks and ATS scores across assessment tasks in AE2 2025



b. Comparison of score density distribution from human marks and the ATS model by rubric element in AE2 2025



Appendix 4

Spearman correlation between rubric element scores marked by human in AE2 2025

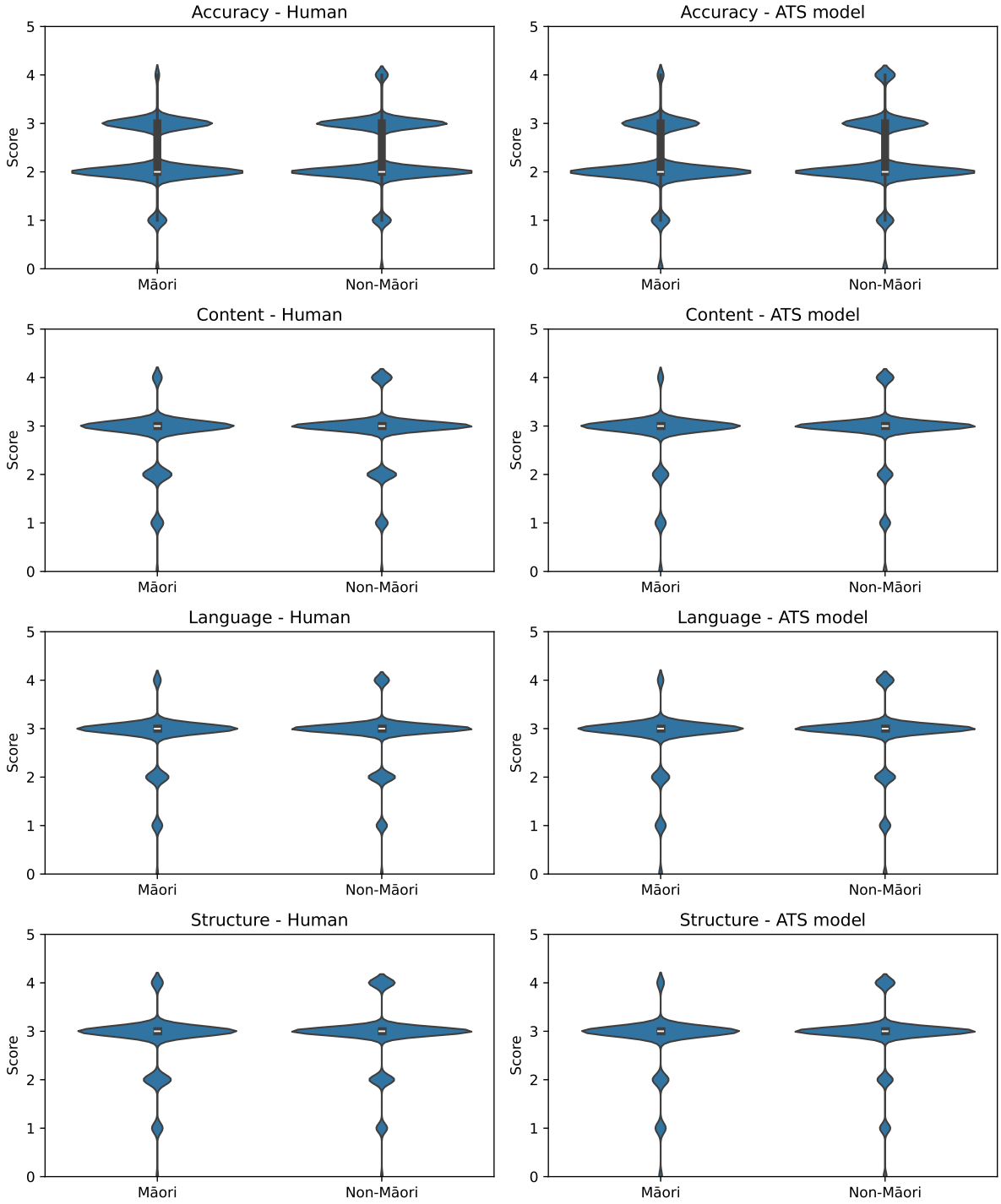
	Accuracy_human	Content_human	Language_human	Structure_human
Accuracy_human	1.00	0.48	0.57	0.53
Content_human	0.48	1.00	0.56	0.56
Language_human	0.57	0.56	1.00	0.58
structure_human	0.53	0.56	0.58	1.00

Spearman correlation between rounded rubric element scores marked by ATS in AE2 2025

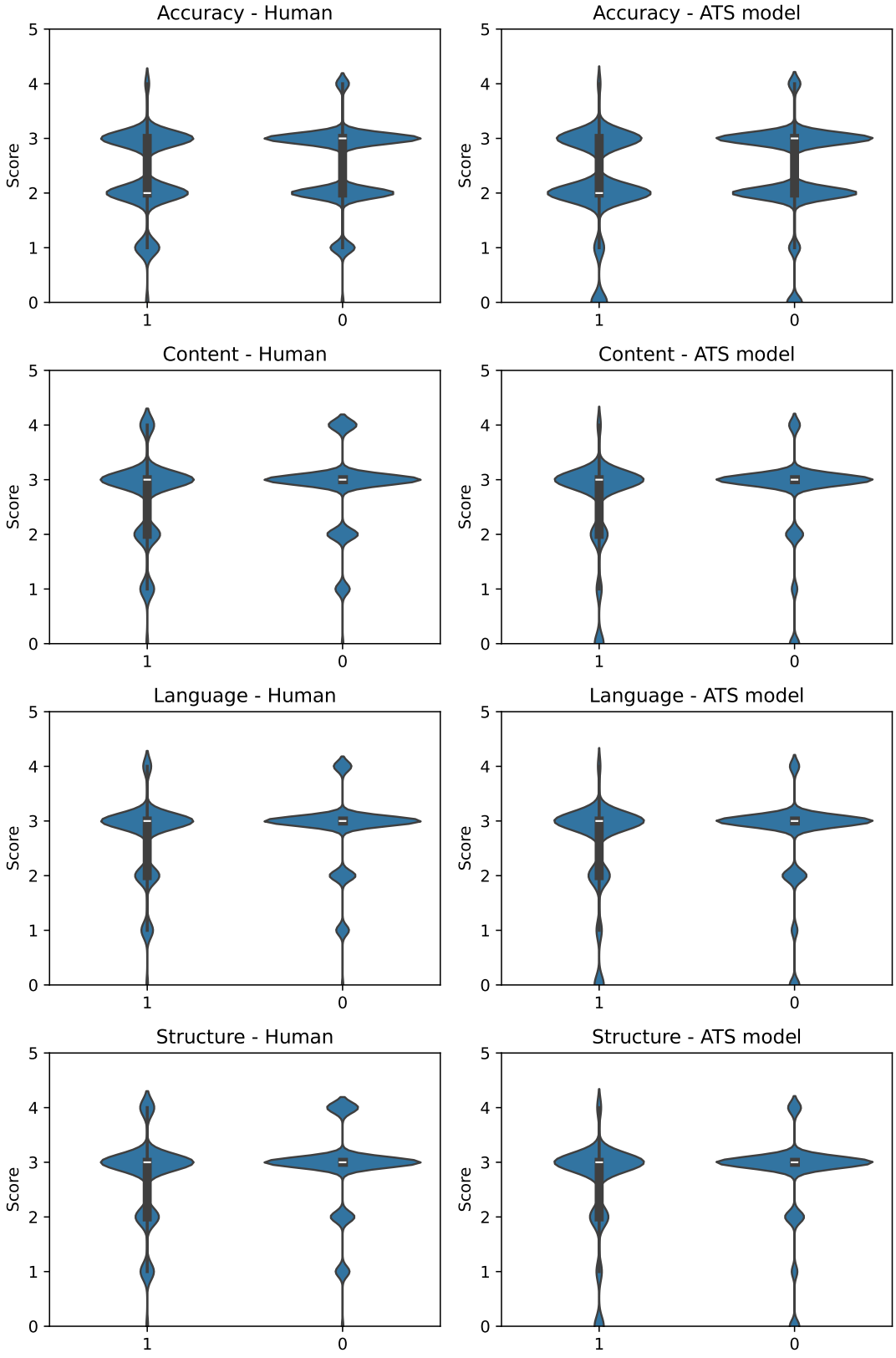
	Accuracy_ats	Content_ats	Language_ats	Structure_ats
Accuracy_ats	1.00	0.64	0.67	0.64
Content_ats	0.64	1.00	0.83	0.89
Language_ats	0.67	0.83	1.00	0.82
Structure_ats	0.64	0.89	0.82	1.00

Appendix 5

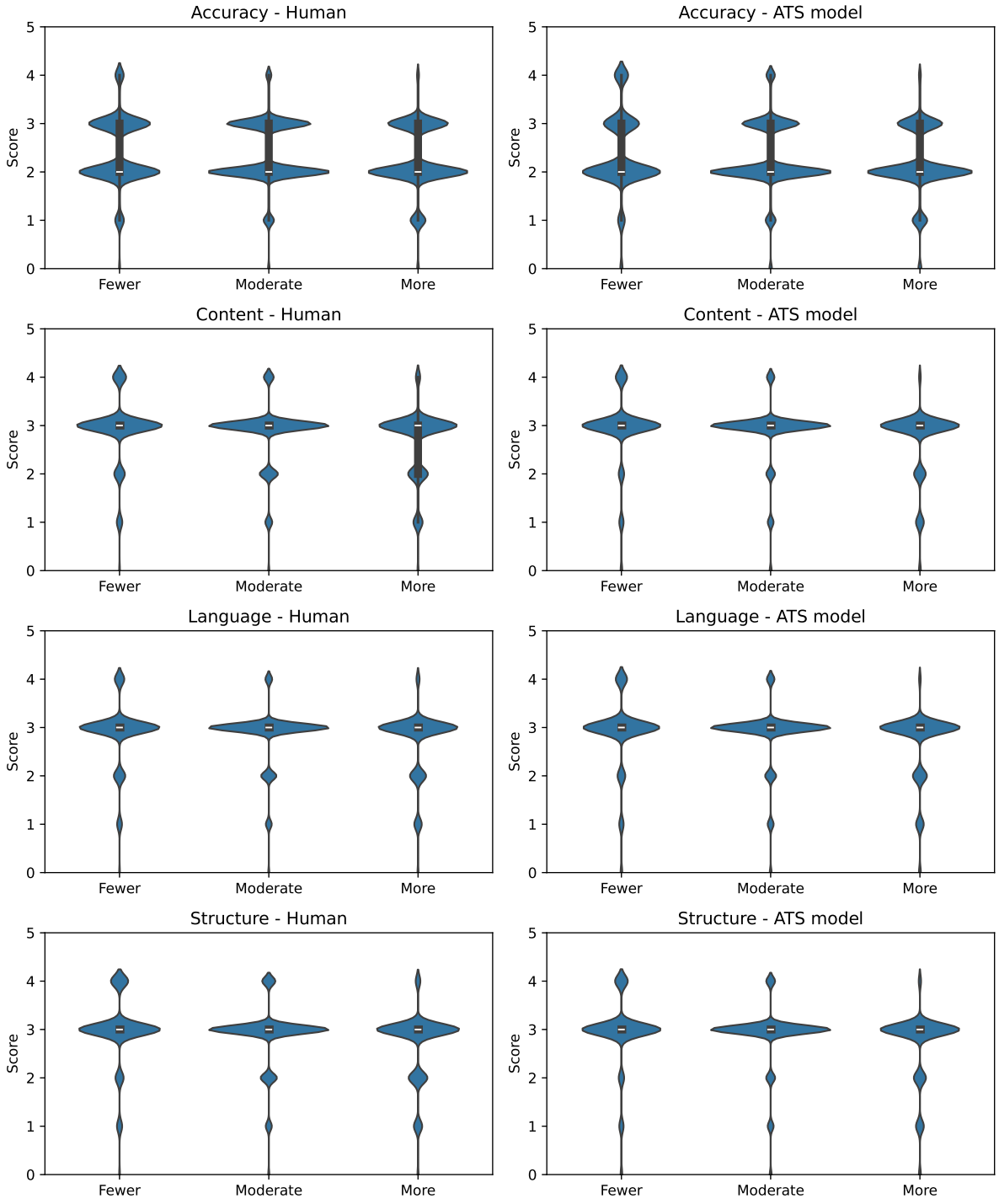
a. Score distribution by Māori/Non-Māori for human and ATS models in AE1 2026



b. Score distribution by Pacific/Non-Pacific for human and ATS models in AE1 2026



c. Score distribution by EQI Groups for human and ATS models in AE1 2026



Appendix 6

Pairwise summary statistics for subgroup bias analysis in AE2 2025

Assessment	Group_a	Group_b	n_a	n_b	ATS_ mean_diff	Human_ mean_diff	Gap_diff
Wk1Q1	Female	Male	3,928	5,357	0.263	0.269	−0.006
Wk1Q2	Female	Male	3,860	5,187	0.237	0.229	0.008
Wk2Q1	Female	Male	2,046	3,513	0.189	0.172	0.017
Wk2Q2	Female	Male	1,983	3,306	0.211	0.168	0.043
Wk1Q1	Māori	Non-Māori	2,343	6,966	−0.137	−0.149	0.012
Wk1Q2	Māori	Non-Māori	2,268	6,803	−0.244	−0.217	−0.027
Wk2Q1	Māori	Non-Māori	1,676	3,892	−0.052	−0.077	0.024
Wk2Q2	Māori	Non-Māori	1,574	3,724	−0.131	−0.109	−0.022
Wk1Q1	Pacific	Non-Pacific	1,376	7,933	−0.223	−0.165	−0.058
Wk1Q2	Pacific	Non-Pacific	1,317	7,754	−0.399	−0.311	−0.088
Wk2Q1	Pacific	Non-Pacific	887	4,681	−0.345	−0.284	−0.061
Wk2Q2	Pacific	Non-Pacific	809	4,489	−0.28	−0.257	−0.024
Wk1Q1	Fewer	Moderate	2,356	4,785	0.232	0.226	0.007
Wk1Q1	Fewer	More	2,356	1,768	0.534	0.49	0.044
Wk1Q1	Moderate	More	4,785	1,768	0.302	0.265	0.037
Wk1Q2	Fewer	Moderate	2,330	4,649	0.318	0.271	0.047
Wk1Q2	Fewer	More	2,330	1,691	0.752	0.623	0.13
Wk1Q2	Moderate	More	4,649	1,691	0.434	0.352	0.082
Wk2Q1	Fewer	Moderate	1,086	2,923	0.072	0.05	0.022
Wk2Q1	Fewer	More	1,086	1,396	0.387	0.38	0.007
Wk2Q1	Moderate	More	2,923	1,396	0.315	0.329	-0.014
Wk2Q2	Fewer	Moderate	1,055	2,778	0.106	0.084	0.021
Wk2Q2	Fewer	More	1,055	1,304	0.503	0.421	0.082
Wk2Q2	Moderate	More	2,778	1,304	0.397	0.337	0.06
Correlation					0.995		

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