NZQA

AI Literacy Marking Tool Assessment

FINAL REPORT

March 2026

Mana Tohu Mātauranga o Aotearoa
New Zealand Qualifications Authority

clearpoint

NZQA Management comment

Mana Tohu Mātauranga o Aotearoa
New Zealand Qualifications Authority

We thank ClearPoint for the analysis and insight in this report and fully accept the recommendations.

We welcome the finding that NZQA has developed an accurate and consistent Automated Text Scoring (ATS) model, which has been our primary focus during the initial development and trialling period.

As we move towards deployment of the model, we are focused on production readiness, including finalising the processes and governance structures, and ensuring robust documentation.

We acknowledge that documentation is a critical enabler of production readiness and public confidence. We will prioritise the development of a centralised documentation repository and adopt standards aligned with software engineering best practice. This will ensure that key decisions, validation evidence, and operational procedures are clearly recorded and accessible, supporting transparency, auditability, and future scalability.

From a security standpoint, we can provide assurance that NZQA's ATS model is a closed model. All data provided to the model is held only by NZQA and cannot be accessed by other AI systems.

The analysis presented in this report, along with the implementation of its recommendations, will enable NZQA to continue embedding the use of ATS to support the marking of the Co-Requisite Literacy Writing assessment.

Table of Contents

1. About the Assessment →
2. Executive Summary →
3. Approach & Methodology →
4. Recommendations →
5. Findings →
6. Appendices →

(Links navigable)

Report disclaimer

This report relates to an independent technical review completed within the scope defined in the engagement SOW between ClearPoint Limited (ClearPoint) and NZQA.

This technical review does not cover: organisational change management; or security assessments (beyond guided technical walkthroughs); vendor solution evaluation or commercial comparisons; budget analysis, cost-benefit assessments, or procurement guidance; integration planning; long-term maintenance strategies; or compliance and regulatory framework development.

The information on which ClearPoint has relied for this technical review was sourced directly from NZQA via supplied documentation, workshops and interviews.

In preparing this technical review report, Clearpoint has assumed that:

- there has been no material change in that information since our investigation.
- the responses to our queries, including all information provided, were true, correct, accurate and complete, and did not contain any omission.
- the representatives of NZQA who responded to our queries were competent to answer.

If any assumption is incorrect, there may be issues that are not identified in this technical review report. ClearPoint has no responsibility or liability for errors or omissions in this technical review report arising from an assumption being incorrect.

This technical review report is addressed to, and may be relied on solely by, NZQA. Without limiting the previous sentence, this technical review report is provided for the sole purpose of assisting NZQA to validate the robustness of its internal AI marking model architecture, assess implementation readiness for potential production deployment, and provide technical assurance for the NZQA AI marking approach of the Writing Co-Requisite, as defined in the SOW between ClearPoint and NZQA.

To the extent that NZQA makes publicly available this technical review report, it is for informational purposes only and on an “as is” and “as available” basis. No person other than NZQA (a third party) may use or rely on this technical review report. To the maximum extent permitted by law, ClearPoint has no responsibility or liability to, or in respect of, any third party relating to their use and/or reliance on this technical review report.

About this Assessment

Background

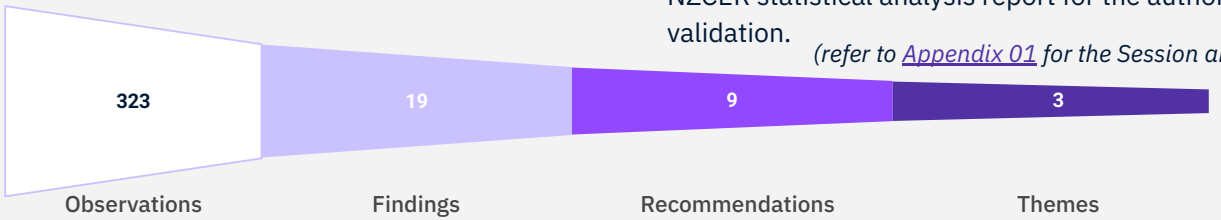
ClearPoint has been engaged by NZQA to conduct an independent technical review of the internally developed Automated Text Scoring (ATS) system for literacy assessments. This review follows NZQA's development of an in-house AI marking model as a proof-of-concept using the DeBERTa-v3-large architecture combined with LLM-based guardrails for safety detection.

The model has been used in pilot assessments (May 2025) but has not yet been deployed to full production, making this the ideal time to identify considerations for future development and production readiness.

The engagement ran from 24 September through to mid-October 2025, during which time the ClearPoint team conducted two workshops with NZQA's technical team (24 September and 30 September), reviewed documentation including technical specifications, architecture diagrams, statistical analyses, and code repositories, and examined performance data comparing the model against established benchmarks.

Note: Subsequent to ClearPoint's independent technical review, NZQA commissioned the New Zealand Council for Educational Research (NZCER) to conduct a detailed statistical analysis of the ATS model's marking performance. Where this report references statistical validation, readers should refer to the subsequent NZCER statistical analysis report for the authoritative statistical validation.

(refer to Appendix 01 for the Session and Sources Register)



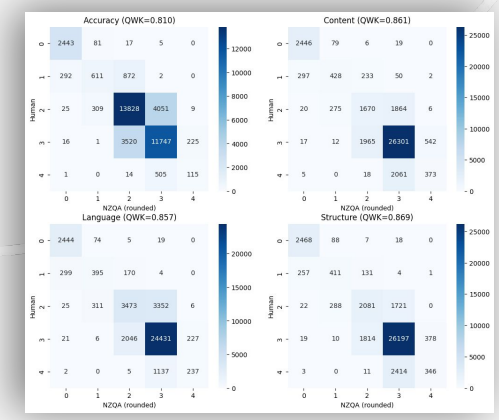
Executive Summary

Executive Summary

The review focused on architectural soundness, process validation, and production readiness from an independent technical perspective, covering model architecture and technical assessment, guardrails and safety framework evaluation, performance and validation assessment, documentation and knowledge management practices.

Statistical analysis demonstrates performance comparable to established benchmarks and substantial agreement with expert human markers. The model shows particular strength in Structure, Language, and Content, with the Accuracy rubric demonstrating slightly different scoring patterns that warrant further investigation*. However, statistical validation confirms the model's scoring patterns align with expected assessment outcomes, supporting its technical viability.

Safety mechanisms are sound: confidence thresholds trigger human review, LLM guardrails screen for content safety, and boundary score review protocols provide additional oversight.



(refer to Appendix 02 for independent analysis details)

* As noted in the background section, please refer to the subsequent NZCER statistical analysis report for the authoritative statistical validation.

Executive Summary _continued

Issues found were mainly on documentation and production readiness rather than the solution capability. NZQA has built an effective proof of concept; the path to production requires work in five areas.

Requirements Maturity

Requirements exist but lack production-level depth in operational procedures, edge case handling, and fail-safe protocols.

Development Traceability

Model selection rationale, hyperparameter decisions, training iterations, and challenges encountered lack documentation.

Governance Framework

Decision-making processes and accountability structures need formal definition to support production deployment.

Production Readiness

Solution design documentation, operational runbooks, monitoring protocols, and incident response procedures are not yet established.

Code Documentation

The codebase functions well but lacks inline explanation of implementation decisions and architectural choices.

Approach & Methodology

Approach & Methodology

ClearPoint follows a structured three-phase methodology that maintains traceability from evidence to recommendations:

Phase 1: Observations

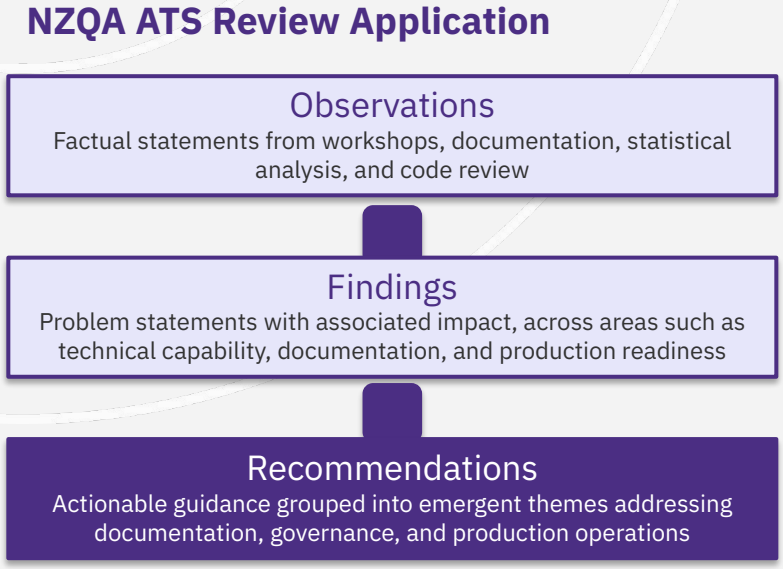
Capture factual findings from workshops, document reviews, technical analysis, and code repositories. Each observation is classified as either 'Issue' or 'Non-Issue' based on whether it represents a concern requiring attention in the context of the assessment.

Phase 2: Findings

Analyse observations to develop finding statements with associated impact. Multiple observations may combine into a single finding, or be split into distinct findings. Every observation marked as 'Issue' traces to at least one finding statement.

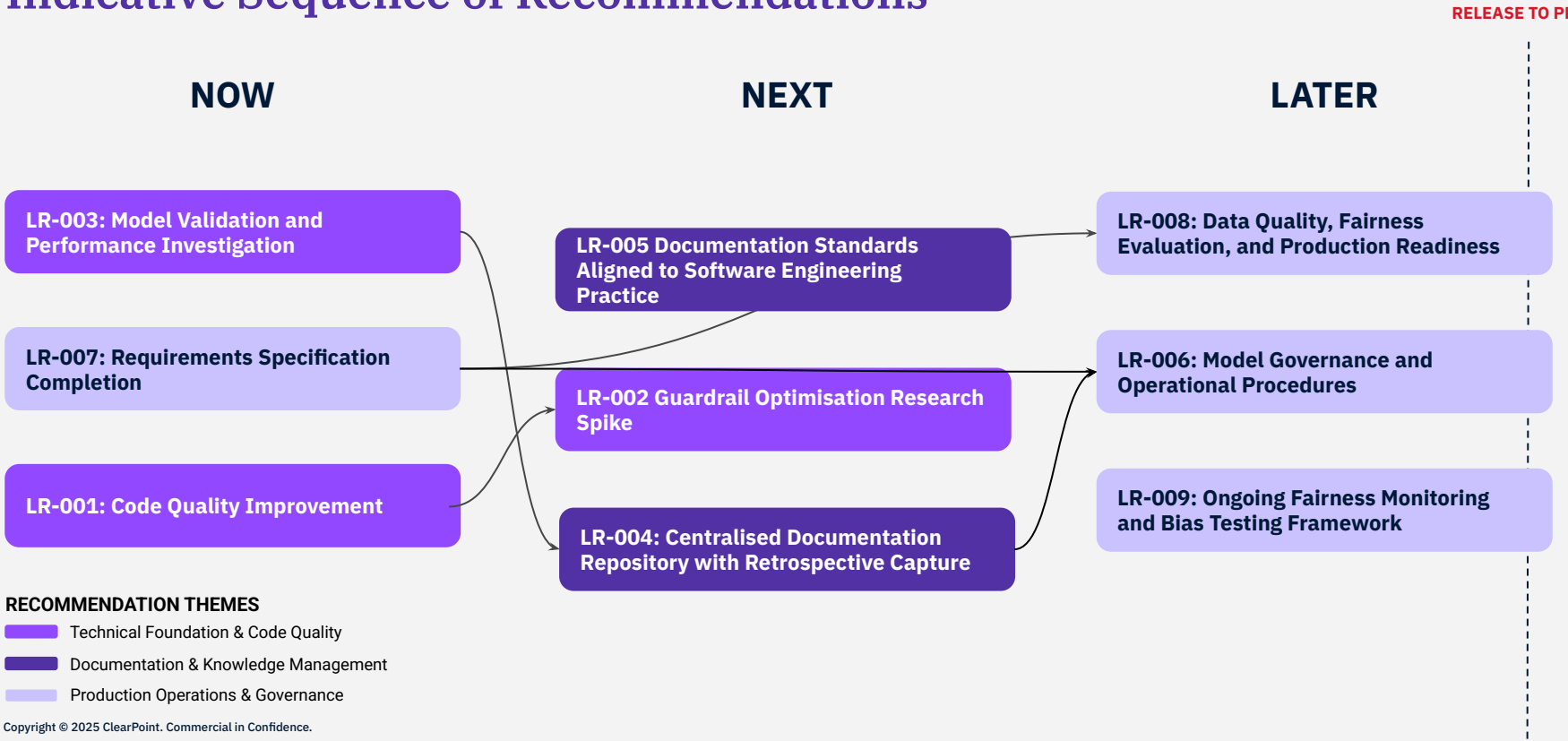
Phase 3: Recommendations

Develop actionable recommendations that address identified findings, grouped into emergent themes. Every finding maps to at least one recommendation, ensuring appropriate coverage and traceability.



Recommendations

Indicative Sequence of Recommendations



LR-001

Technical Foundation & Code Quality

Related findings:

IS-13: Code Quality and Documentation Standards

RECOMMENDATION

Code Quality Improvement:

NZQA should undertake a focused code review and refactoring pass to enhance code maintainability for production deployment. This includes adding inline comments to explain decision rationale, documenting parameter selections, removing unused code, and standardising configuration naming conventions.

Whilst these are minor technical debt items individually, addressing them collectively will improve the team's ability to troubleshoot production issues, onboard new team members efficiently, and maintain code quality standards appropriate for a government assessment system.

LR-002

Technical Foundation & Code Quality

Related findings:

IS-13: Code Quality and Documentation Standards
IS-14: Guardrail Optimisation and Error Handling

RECOMMENDATION

Guardrail Optimisation Research Spike:

We recommend NZQA initiate a time-boxed research spike to evaluate guardrail optimisation opportunities. First, investigate whether fine-tuning smaller, more cost-effective language models could deliver comparable guardrail performance without compromising safety detection capabilities. Second, redesign the fallback logic to handle response-level errors rather than solely endpoint availability, creating a more intelligent and resilient safety mechanism. This investigation should be completed before production deployment to inform the final guardrail architecture.

LR-003

Technical Foundation & Code Quality

Related findings:

IS-4: LLM Guardrail Documentation
IS-10: LLM Explainer Validation
IS-11: Accuracy Rubric Performance Patterns

RECOMMENDATION

Model Validation and Performance Investigation:

NZQA should undertake further investigations to strengthen model validation before production deployment. First, investigate the Accuracy rubric's scoring patterns, which, whilst demonstrating a very strong QWK score of 0.81 (almost perfect agreement), is lower than the other rubrics (0.83-0.87) and exhibits different scoring patterns, particularly a tendency toward more conservative predictions. This investigation should determine whether the performance differences stem from rubric complexity, training data characteristics, model architecture factors, or inherent characteristics of accuracy assessment. Second, conduct thorough validation of the LLM Explainer component to assess its reliability, accuracy, and consistency in providing score explanations.

Given the Explainer's role as a transparency and assurance layer, validation should include testing explanation quality across diverse response types and establishing quality metrics for explanation outputs. Both investigations should be documented with clear findings and any remediation actions required before production deployment.

Refer to the subsequent NZCER statistical analysis report for the authoritative statistical validation.

LR-004

Documentation & Knowledge Management

Related findings:

IS-4: LLM Guardrail Documentation
IS-5: Kaggle Provenance Documentation
IS-9: Architectural Decision Traceability
IS-12: Testing Evidence Consolidation
IS-1: Testing Methodology Documentation
IS-3: Fairness Evaluation and Bias Testing
IS-6: Requirements Production Readiness
IS-8: Performance Monitoring Capabilities
IS-17: Evaluation Methodology and Metrics Framework
IS-18: Documentation Metadata and Knowledge Management

RECOMMENDATION

Centralised Documentation Repository with Retrospective Capture:

NZQA should establish a centralised documentation repository (such as a Confluence space or equivalent knowledge management platform) to consolidate all ATS model artefacts, serving as the single source of truth. Each model should have a dedicated "landing page" (similar to a system runbook) providing overview, key performance metrics, and links to detailed supporting documentation.

Given the model's development as a proof-of-concept, NZQA should undertake retrospective documentation to capture missing information, with priority given to:

- Kaggle competition provenance and licensing verification,
- architectural decision rationale,
- consolidated testing and evaluation evidence and validation methodologies,
- guardrail performance metrics, and
- data cleansing processes.

This centralised approach supports knowledge transfer, facilitates future model iterations, and provides a scalable template for potential expansion to other assessments.

LR-005

Documentation & Knowledge Management

Related findings:

IS-1: Testing Methodology Documentation
IS-4: LLM Guardrail Documentation
IS-5: Kaggle Provenance Documentation
IS-9: Architectural Decision Traceability
IS-15: Model Transparency and Validation Documentation
IS-17: Evaluation Methodology and Metrics Framework
IS-18: Documentation Metadata and Knowledge Management

RECOMMENDATION

Documentation Standards Aligned to Software Engineering Practice:

NZQA should develop and implement documentation standards aligned with software engineering best practices for machine learning systems. These standards should define expectations for:

- Model specifications and architecture decision records (ADRs)
- ML-specific testing and evaluation frameworks, including:
  - Test and evaluation approach and methodology
  - Training, validation, and test dataset specifications (including stratification strategies, size, and provenance)
  - Performance metrics selection rationale and thresholds (e.g., QWK, accuracy by rubric, confidence scoring)
  - Evaluation protocols across different assessment types and student cohorts
  - Model performance monitoring and drift detection approaches
- Operational procedures and deployment protocols
- Requirements traceability matrices

The standards should draw from established frameworks such as Model Cards, Datasheets for Datasets, and IEEE/ISO documentation standards, adapting patterns from New Zealand Government ICT assurance frameworks where applicable. *Please refer to Appendix 03 - Suggested Documentation Catalogue*

LR-006

Production Operations & Governance

Related findings:

IS-2: Model Oversight and Operational Procedures
IS-3: Fairness Evaluation and Bias Testing
IS-8: Performance Monitoring Capabilities

RECOMMENDATION

Model Governance and Operational Procedures:

NZQA should document model governance and operational procedures to support sustained production operation.

This should include: model versioning and change management processes, performance monitoring frameworks (including data drift detection, model performance degradation tracking, and ongoing fairness evaluation), incident response procedures, and rollback capabilities. These procedures should define clear roles and responsibilities, escalation paths, and decision-making authority for operational scenarios.

Establishing these governance frameworks will enable NZQA to demonstrate appropriate AI system stewardship expected of a government organisation operating in the assessment space.

LR-007

Production Operations & Governance

Related findings:

IS-6: Requirements Production Readiness
IS-19: Requirements Dependencies and Logical Consistency

RECOMMENDATION

Requirements Specification Completion:

NZQA should complete the requirements specification to address gaps identified during this review, with particular focus on operational requirements, security specifications, privacy documentation, and production acceptance criteria. The requirements framework should be reviewed to resolve logical inconsistencies where must-have requirements depend on should-have or could-have capabilities.

Clear production readiness criteria should be established and documented, including specific metrics, thresholds, and acceptance gates that will inform the deployment decision. This work should reference the New Zealand Information Security Manual (NZISM) for security and privacy requirements, ensuring alignment with government ICT standards.

LR-008

Production Operations & Governance

Related findings:

IS-6: Requirements Production Readiness
IS-7: Data Quality and Provenance
IS-16: Data Pipeline Operations and Manual Interventions
IS-19: Requirements Dependencies and Logical Consistency

RECOMMENDATION

Data Quality, Fairness Evaluation, and Production Readiness:

NZQA should address production readiness gaps across data quality and operational scalability.

First, resolve data quality and provenance risks by obtaining definitive metadata indicating final student submissions and documenting the data cleansing process with quality controls. Second, reduce manual intervention requirements throughout the data pipeline, including manual assessment status corrections, manual transfer of low-confidence flags, and manual data cleansing procedures. Third, address any remaining inhibitors to production deployment identified during requirements completion (LR-007) and operational procedure development (LR-006).

These foundational elements are essential for NZQA to demonstrate production readiness and ensure assessment integrity at scale.

LR-009

Production Operations & Governance

Related findings:

IS-3: Fairness Evaluation and Bias Testing
IS-8: Performance Monitoring Capabilities

RECOMMENDATION - IN PROGRESS

Ongoing Fairness Monitoring and Bias Testing Framework:

Building on the positive NZCER fairness evaluation findings, NZQA should establish internal frameworks for ongoing fairness monitoring and bias testing.

This should include documented protocols for periodic fairness re-evaluation across demographic groups, disaggregated performance analysis methodologies, procedures for detecting emerging fairness issues over time, and bias mitigation strategies for sustained operational use. Whilst initial external validation confirms equitable operation, production deployment requires internal capability for continuous fairness assurance as student populations and assessment patterns evolve.

These frameworks are essential for NZQA to provide ongoing public assurance of equitable assessment and respond appropriately to Official Information Act requests or public scrutiny about sustained algorithmic fairness.

Findings

Model Architecture & Technical Assessment

IS-9	**Architectural Decision Traceability:** The literature review documenting model architecture alternatives exists as a standalone document, but the connection between this research and the final selection of DeBERTa-v3-large is not explicitly documented as part of a traceable technical decision-making process	Whilst substantive alternatives analysis was conducted via literature review covering BERT variants, LLM approaches, and hybrid systems, the absence of explicit documentation linking research findings to the architectural decision limits ability to demonstrate decision-making traceability to stakeholders. Without clear documentation showing how the literature review informed the choice of DeBERTa-v3-large over alternatives (and why other approaches were deemed unsuitable for NZQA's specific use case), new team members and external reviewers cannot trace the rationale from exploration through to implementation.
IS-5	**Kaggle Provenance Documentation:** The provenance and extent of adaptation from the Kaggle competition entry remains unclear, with limited documentation of which specific competition entry was used, what portions were adapted versus developed from scratch, and the rationale for architectural decisions based on competition results	Without clear documentation of code provenance and adaptation decisions, NZQA cannot demonstrate appropriate due diligence if usage rights are questioned. Whilst the DeBERTa base model uses MIT licence and technical risk of licensing violation is low, the documentation gap means NZQA cannot demonstrate that appropriate verification was undertaken—particularly important given public sector context and potential Official Information Act requests about AI development provenance. This gap also makes it difficult for new team members to understand architectural decisions and limits ability to clearly communicate development approach to stakeholders.

Model Architecture & Technical Assessment _continued

IS-13	**Code Quality and Documentation Standards:** Review of the code identified several minor opportunities to improve overall code quality, including adding comments, explaining parameter selection, removing dead code, and refactoring inconsistent naming conventions for configuration parameters.	Code quality improvements would support smoother production operations and team scalability. The absence of parameter documentation and explanatory comments makes it harder for the team to validate model behaviour, troubleshoot unexpected scoring outcomes, or efficiently onboard new team members who need to understand decision rationale. Inconsistent naming conventions and areas of unmaintained code reduce code review efficiency and increase the potential for configuration errors during system updates or incident response. For a system intended to support literacy assessments at scale, addressing these engineering practice gaps will improve the maintainability and auditability of the codebase as the ATS system transitions from PoC to sustained operational use.
IS-11	**Accuracy Rubric Performance Patterns:** The Accuracy rubric demonstrates QWK 0.81 ('almost perfect agreement' per Landis & Koch standards) but exhibits a different scoring pattern compared to other rubrics (Content QWK 0.86, Language QWK 0.86, Structure QWK 0.87), with more balanced distribution between scores 2 and 3 and the model scoring more conservatively than humans in this rubric **Refer to the subsequent NZCER statistical analysis report for the authoritative statistical validation.**	This performance pattern difference, whilst within acceptable statistical ranges for deployment, warrants investigation to understand whether it reflects genuine rubric-specific assessment characteristics (e.g., inherent difficulty in assessing accuracy versus structure), training data distribution patterns, or rubric definition considerations. Understanding and documenting the rationale for these patterns is important for explaining model behaviour to stakeholders, particularly if questioned about consistency in scoring approaches across all rubrics, and for informing future rubric development or training data collection strategies.

Guardrails & Safety Framework

IS-4	**LLM Guardrail Documentation:** LLM-based guardrails have been implemented and tested (including 1,000 sample validation achieving 100% precision for off-topic detection), however documentation of model selection rationale, validation methodology, and performance metrics is not consolidated or formally captured	Without consolidated documentation of guardrail performance, validation methodology, and model selection rationale, it is difficult to assess safety control effectiveness across different scenarios, demonstrate that guardrails reliably detect off-topic or concerning content, or provide assurance to the Ministry, school leaders, and parents about the robustness of safety mechanisms protecting students. This limits transparency about how potentially concerning student responses are identified and handled, makes it harder to troubleshoot or refine guardrails for future assessment cycles, and affects ability to demonstrate appropriate safeguarding in a high-stakes student assessment context.
IS-10	**LLM Explainer Validation:** The LLM Explainer component was recently developed and integrated into the scoring pipeline, but has not yet undergone thorough evaluation, testing, or validation to assess its reliability, accuracy, and consistency in providing score explanations	Without thorough evaluation of the Explainer component, there is no documented evidence of its reliability, accuracy, and consistency in providing explanations for scores. Since the Explainer serves as an additional assurance layer and contributes to the transparency and explainability requirements, unvalidated performance creates risk that explanations may be inconsistent, inaccurate, or misleading, potentially undermining confidence in the overall marking solution.

Guardrails & Safety Framework _continued

IS-14	**Guardrail Optimisation and Error Handling:** The current guardrail system, which leverages prompt engineering and a multi-model fallback primarily for outage resilience, presents several opportunities for optimisation. There is scope to enhance response quality, reduce costs, and improve performance by exploring the fine-tuning of more cost-effective models. Additionally, the fallback mechanism could be evolved to handle response-level errors rather than just endpoint availability, creating a more robust and intelligent guardrail.	The current guardrail configuration fulfils its primary safety objectives but potentially operates at a higher cost and complexity than necessary for sustained production use. Reliance on prompt engineering alone without exploring model fine-tuning opportunities means the system incurs ongoing operational costs that could be optimised whilst maintaining or improving response quality. The outage-focused fallback mechanism provides infrastructure resilience; however, the approach of switching model endpoints in response to any error represents an overly simplistic error-handling strategy that could be refined.

Performance & Validation

IS-1	**Testing Methodology Documentation:** Testing activities including guardrail validation and statistical validation against human scoring have been conducted, but formal testing approaches, acceptance criteria, and validation methodologies have not been documented	Without documented testing procedures and acceptance criteria, NZQA cannot demonstrate the rigour and completeness of validation activities to stakeholders, cannot replicate validation approaches for future model iterations, or provide assurance about testing coverage. This creates particular risk when responding to public scrutiny about AI assessment safety and limits the ability to demonstrate due diligence in model validation.
IS-17	**Evaluation Methodology and Metrics Framework:** Evaluation metrics and performance assessment methodologies lack extensive documentation of calculation approaches, filter assumptions, edge case handling, and acceptable performance thresholds, with different possible interpretations for metrics depending on data inclusion/exclusion decisions	Without documented evaluation methodology and agreed metrics framework, it is difficult to demonstrate consistent performance assessment across assessment cycles, compare results meaningfully between events, or provide transparent explanations of model performance to stakeholders. Ambiguity in metrics calculation (e.g., whether to include low confidence data, how to handle zero scores, which filters to apply) creates risk of inconsistent reporting and limits ability to establish clear production readiness criteria.

Performance & Validation _continued

IS-12	**Testing Evidence Consolidation:** Test reports, evaluation analyses, and performance validation results exist but are distributed across multiple locations, documents, and organisational divisions rather than being consolidated into an accessible testing evidence repository	Distribution of testing evidence across multiple locations makes it difficult to obtain a single view of validation activities, assess testing completeness, support knowledge transfer to new team members, or provide consolidated evidence to stakeholders. This fragmentation increases the effort required to demonstrate production readiness and creates risk that important testing evidence may be overlooked when needed for assurance purposes.
IS-15	**Model Transparency and Validation Documentation:** Experience with AI assessment systems in education has demonstrated how model performance issues such as overfitting and poor generalisation on unseen data can be exacerbated by a lack of transparency into development and validation processes. To ensure stakeholders and the public have faith in the in-house developed model, it is critical that NZQA establish a higher standard of practice through transparency and robust validation documentation, thereby building a foundation of trust and reliability from the outset.	Without transparent and well-documented model development and validation practices, NZQA cannot adequately defend scoring decisions to students, educators, or regulatory bodies when outcomes are challenged. For a system scoring important literacy assessments that affect student qualifications, insufficient transparency in methodology and performance validation undermines stakeholder confidence and operational readiness for scaled deployment. Critically, NZQA's internal stakeholders and governance bodies require complete visibility into how the system works, why it makes particular decisions, and how to validate its outputs. This transparency is essential to avoid situations where AI systems appear to perform well but operate as 'black boxes' without sufficient understanding of their operation. This visibility enables NZQA to demonstrate the robustness and reliability required for public trust in an AI-powered assessment capability, and ensures the organisation's capacity to maintain, improve, and extend the solution over time.

Performance & Validation _continued

IS-3	**Fairness Evaluation and Bias Testing:** NZQA has engaged NZCER to conduct independent fairness and bias analysis, with preliminary findings (October 2025) indicating no evidence of demographic bias across Gender, Māori/Non-Māori, Pacific/Non-Pacific, and Equity Index groups. However, internal frameworks for ongoing fairness monitoring, periodic evaluation protocols, and documented bias mitigation strategies have not been established for sustained operational use.	Whilst the NZCER external review provides initial validation that the model performs equitably across demographic groups, NZQA requires documented internal methodologies for periodic fairness re-evaluation, disaggregated performance monitoring over time, and bias mitigation protocols. The absence of an established ongoing monitoring framework means NZQA cannot systematically track fairness metrics across future assessment cycles, detect emerging fairness issues as student populations or assessment patterns evolve, or respond appropriately to Official Information Act requests or public scrutiny about sustained algorithmic fairness. This particularly affects confidence in maintaining equitable assessment as the solution scales across diverse student populations nationally.
IS-8	**Performance Monitoring Capabilities:** Performance monitoring capabilities for detecting data drift, model performance degradation, and ongoing fairness evaluation have not been implemented, and NZQA has not yet developed an internal framework for this monitoring.	Without ongoing monitoring for data drift, model performance degradation, and fairness metrics, NZQA cannot detect when model performance begins to degrade, when input data characteristics change in ways affecting model validity, or when fairness issues emerge over time. This limits ability to proactively identify and address issues before they impact assessment outcomes, reduces confidence in sustained model performance across assessment cycles, and affects ability to provide ongoing assurance to stakeholders that the model continues to perform as validated. The absence of monitoring frameworks particularly affects confidence in scaling the solution across broader student populations where data characteristics may differ from training data.

Documentation & Governance

IS-6	**Requirements Production Readiness:** Requirements specification has not been developed to include the completeness and detail expected for production deployment, with multiple critical components marked as 'TBC' (to be confirmed) or 'TBD' (to be determined), incomplete operational requirements, and undefined acceptance criteria for assessing production readiness	Incomplete requirements specification creates uncertainty for production deployment planning, makes it difficult to assess when the solution is ready for scaled deployment, and limits the ability to provide clear acceptance criteria to stakeholders. This particularly affects operational components such as monitoring, error handling, and data validation where architecture documentation does not yet specify production requirements. Whilst the retrospective, adaptive approach to requirements development was appropriate for PoC/pilot phases and intentional flexibility in specific operational parameters is good practice, production deployment requires resolution of outstanding items and documented clarity for ongoing requirements management.
IS-19	**Requirements Dependencies and Logical Consistency:** Requirements contain logical inconsistencies where must-have requirements depend on should-have or could-have capabilities, and critical performance metrics depend on optional functionality	Inconsistent requirement prioritisation creates ambiguity about true production readiness criteria and affects ability to assess solution completeness against defined needs. The dependency of must-have performance metrics on optional capabilities particularly creates risk that core requirements cannot be verified if optional features are not implemented. This affects contract management if procuring solutions, internal development prioritisation decisions, and the ability to provide clear production readiness criteria to stakeholders.

Documentation & Governance _continued

IS-2	**Model Oversight and Operational Procedures:** Model oversight and operational procedures have not been documented, including model versioning, change management, monitoring, and rollback capabilities	NZQA does not yet have established processes for managing model changes, tracking versions, monitoring production performance, and responding to issues. As the solution moves towards production deployment, formalising these procedures will be essential to support operational risk management during high-stakes assessment cycles and enable effective incident response. The absence of formal change management and rollback procedures particularly affects confidence in deploying model updates during assessment windows and responding to unexpected performance issues. Establishing these governance frameworks will enable NZQA to demonstrate the appropriate AI system stewardship expected of a government organisation.
IS-18	**Documentation Metadata and Knowledge Management:** Technical documentation including the ATS Model Approaches Review lacks document metadata (authorship, versioning, dates, purpose statements, intended audience), limiting traceability, ownership clarity, and document lifecycle management	Without document metadata and ownership attribution, NZQA cannot demonstrate document authority, track revisions over time, or establish clear accountability for technical documentation currency. This affects knowledge transfer to new team members, creates ambiguity about which documents represent current authoritative guidance versus historical analysis, and limits ability to demonstrate documentation governance to auditors or in response to Official Information Act requests.

Data Quality & Operations

IS-7	**Data Quality and Provenance:** NZQA lacks direct access to final submitted student responses, instead relying on approximately 30-second to 5-minute progress snapshots throughout the assessment period that are reconstructed into learner responses, without metadata definitively indicating which snapshot represents the student's final intended answer	Without direct access to final submitted responses, there is risk that the reconstruction process may not accurately capture the student's intended final answer, particularly in scenarios where assessments are reopened after initial submission or where snapshot timing doesn't align with student submission actions. This data provenance uncertainty affects the foundation of all scoring (both human and automated), creates quality risk for students near pass/fail boundaries, and limits NZQA's ability to provide definitive assurance that marks are assigned to the correct version of student work.
IS-16	**Data Pipeline Operations and Manual Interventions:** Production data pipeline requires manual interventions at multiple stages, including manual assessment status corrections before ingestion, manual transfer of low-confidence flags due to API limitations, and manual data cleansing procedures, creating operational risk and limiting scalability	Manual intervention requirements throughout the data pipeline create operational bottlenecks, introduce human error risk during high-pressure assessment cycles, and limit the ability to scale the solution beyond current volumes. Manual status corrections and workarounds for API limitations increase operational complexity and reduce confidence in end-to-end automation. This affects production readiness assessment and creates risk during assessment cycles where timing is critical for results delivery.

Appendix

Appendix 01: Session and Sources Register

Date	Description	Attendees/Source
24 September 2025	Workshop #1	Various ClearPoint and NZQA Stakeholders
30 September 2025	Workshop #2	Various ClearPoint and NZQA Stakeholders
24 September 2025	High-level technical overview of NZQA's in-house Automated Text Scoring model using DeBERTa-v3-large architecture with supervised learning, multi-objective training across four rubrics, multilabel stratified 5-fold cross-validation, LLM guardrails for safety/off-topic detection, and word count moderation.	NZQA ATS in-house high-level description.pdf
29 September 2025	System architecture diagrams showing the ATS workflow including: entry guardrails, marker components, explainer module, exit guardrails, data acquisition from Assessment Master, machine marking pipeline, and integration with RM Assessor for targeted human marking.	Diagrams - ATS V2 - 26092025.pdf
29 September 2025	Code repository archive containing NZQA's internal ATS implementation codebase (dated 25 September 2025). Includes manifest file and framework directory with training/inference code for the DeBERTa-based scoring model, LLM guardrails (safety/off-topic detection), word count moderation, and explainer components referenced in workshop discussions. This is the actual implementation provided to ClearPoint for technical review.	repo-ats-framework.zip

Appendix 01: Session and Sources Register *_continued*

Date	Description	Attendees/Source
29 September 2025	Analysis undertaken by ClearPoint based on information contained in (data_litnum_may_2025.csv) + NZQA Report based on data_litnum_may_2025.csv (quadratic_weighted_kappa.pdf) which contains: Statistical analysis notebook showing QWK performance metrics (0.8103-0.8693, "almost perfect agreement") across four rubrics, confusion matrices comparing human vs NZQA internal model scores for 38,689 May 2025 assessment records.	Results Analysis (data_litnum_may_2025.csv) + quadratic_weighted_kappa.pdf
29 September 2025	Requirements specification document (September 2025) defining functional (F1-F18) and non-functional (N1-N10) requirements for ATS implementation including criteria-based scoring, guardrails, explainability, performance thresholds (60K students/3 days), and NZISM alignment.	Automated Text Scoring Requirements_202509
3 October 2025	Spreadsheet consolidating the answers to our follow up questions	Answers to follow up Questions (Excel spreadsheet)
1 October 2025	Sample dataset (18 records) of cleaned student responses for Week 1 Question 1, showing NSN, date/time, exam details, response text after HTML artifact removal, and word counts.	Cleansed_Responses_W1Q1_20250820_n18.csv

Appendix 01: Session and Sources Register *_continued*

Date	Description	Attendees/Source
1 October 2025	Flagged records (3 shown) where NZQA internal model indicated low confidence in assigned scores, containing NSN, week/question numbers, and low confidence indicators for targeted human review.	Low_confidence_scores_from_In-house_W1Q1_20250912_n30045.csv
30 September 2025	Comparative performance visualisation showing NZQA Internal Model performance against an external benchmark, across accuracy metrics (Exact, Adj., Exact+Adj., PCC, CK, QWK, MCC) for May 2025 assessment event covering ~38K records.	model_metrics_may_2025.png
1 October 2025	LLM guardrail outputs (3 records shown) flagging responses with potential safety concerns (self-harm, violence, suicidal ideation) requiring immediate triage and welfare intervention.	Safety_concern_from_In-house_W1Q1_20250912_n30045.csv
7 October 2025	Stakeholder update with Alastair Holmes	Various ClearPoint and NZQA Stakeholders
7 October 2025	NZQA internal planning document (updated 28 May 2025) outlining the operational approach for implementing automated text scoring for Co-requisite Writing Assessment Event 1, May 2025. Captures information relating to the research methodology, business decisions, implementation processes, contingency planning, and statistical analysis from the September 2024 pilot. Includes targeted marketing strategies and treatment of te reo Māori responses	COPY ATS Operating Plan - Research business decisions and implementation.pdf

Appendix 01: Session and Sources Register _continued

Date	Description	Attendees/Source
17 October 2025	The ATS Model Approaches Review is a technical research document that surveys automated text scoring methodologies, covering both supervised and unsupervised approaches, with particular focus on transformer-based models (BERT, RoBERTa, GPT) and ensemble methods using large language models. The document includes a literature review of existing ATS systems, open source implementations, performance analysis, explainability considerations, and extensibility options for scaling to other exams, providing the theoretical and methodological foundation that informed NZQA's technical decisions for their in-house 2025 ATS system implementation.	ATS Model Approaches Review.pdf
24 October 2025	Summary of initial findings from NZCER's independent fairness and bias evaluation of NZQA's in-house ATS model. This external review was commissioned by NZQA in parallel to ClearPoint's technical assessment to validate equitable performance across demographic groups (Gender, Māori/Non-Māori, Pacific/Non-Pacific, and Equity Index).	Excerpt of NZCER DRAFT report - Bias and Fairness.pdf
24 October 2025	ClearPoint/NZQA: draft report playback	Various ClearPoint and NZQA Stakeholders

Appendix 02: Independent Analysis Overview

ClearPoint conducted lightweight independent analysis to validate NZQA's reported performance findings. This preliminary analysis was subsequently investigated in greater detail by the statistical analysis completed by NZCER.

Refer to the NZCER subsequent statistical analysis report for the authoritative statistical validation.

Analysis Approach:

- Performed independent statistical analysis using the comparative scoring dataset provided by NZQA (May 2025, ~38,689 student responses comparing human markers and NZQA's in-house model scores).
- Calculated Quadratic Weighted Kappa agreement metrics between human markers and NZQA's in-house model across all four rubrics
- Generated confusion matrices to visualise scoring patterns and identify systematic differences

Key Findings:

- **Analysis correlates with NZQA's reported findings**, providing independent validation of model performance
- Confusion matrices reveal scoring patterns across rubrics, highlighting areas of strong agreement and conservative scoring tendencies
- Dataset analysis confirms the model's capability to assess student writing at scale

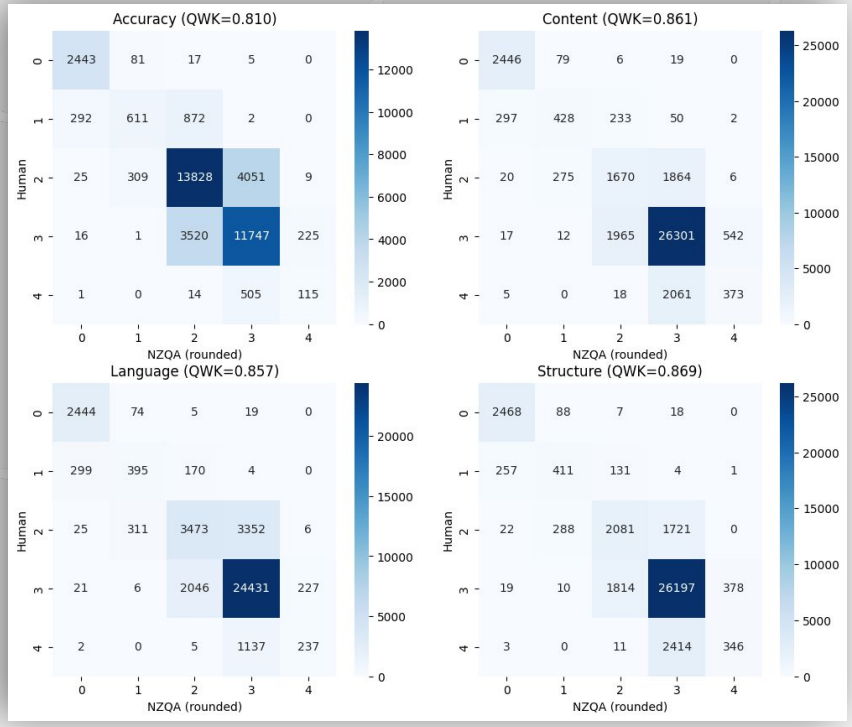
Appendix 02: Independent Analysis Overview _continued

Confusion matrices reveal scoring patterns and agreement levels across the four assessment rubrics

Key Observations:

- **Strongest agreement:** Structure (QWK=0.869), Language (QWK=0.857), and Content (QWK=0.861) rubrics demonstrate high alignment with human marking
- **Lower agreement:** Accuracy rubric (QWK=0.810) shows a slightly lower agreement to other rubrics, but still shows a very high QWK score, with evidence of more conservative scoring by the model (clustering in scores 2-3)

Refer to the subsequent NZCER statistical analysis report for the authoritative statistical validation.

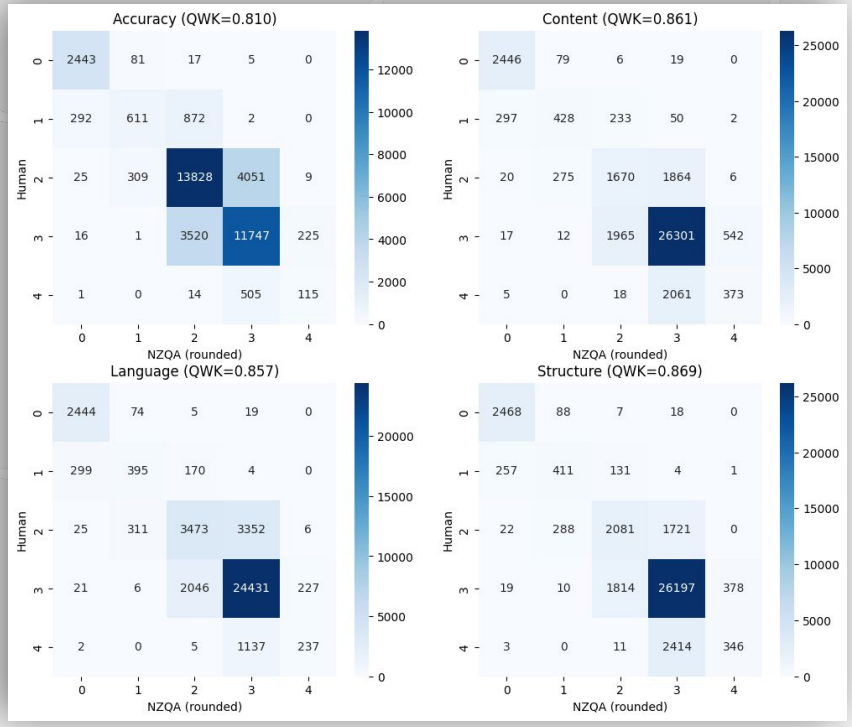


Appendix 02: Independent Analysis Overview _continued

Confusion matrices reveal scoring patterns and agreement levels across the four assessment rubrics

Key Observations:

- **Scoring patterns:** Darker diagonal bands indicate strong exact agreement between human and model scores across all rubrics
- **Conservative tendency:** The model demonstrates a tendency toward middle scores, particularly visible in the Accuracy rubric where higher human scores (3-4) are sometimes assigned lower model scores (2-3)



Refer to the subsequent NZCER statistical analysis report for the authoritative statistical validation.

Appendix 02: Independent Analysis Overview *_continued*

Interpretation:

- Overall performance across all rubrics supports the model's technical capability for literacy assessment
- The Accuracy rubric variance warrants further investigation to understand whether this reflects genuine assessment difficulty, training data characteristics, or rubric definition considerations *(as per observation IS-11)*

Caveats:

- This analysis represents a lightweight validation exercise and has not been fully vetted for production-level accuracy
- Findings are intended to provide assurance that ClearPoint independently reviewed the data rather than relying solely on reported metrics
- More comprehensive analysis would be required to support production deployment decisions

Appendix 03: Suggested Documentation Catalogue - Priority Summary

Priority Category	Count	Description	Colour Code
Critical - Retrospective	6	Must be completed retrospectively before production - documents what has been done during PoC phase	Light Red
Critical - Prospective	8	Must be developed prospectively before production - new processes/standards for production operation	Light Orange
High - Retrospective	3	Should be completed retrospectively to support production readiness	Light Green
High - Prospective	4	Should be developed prospectively to support production operation	Light Blue

Appendix 03: Suggested Documentation Catalogue - Requirements

Document Type	Purpose	Key Contents	Audience	Related To (Traceability)	Industry Standards / Frameworks	NZQA Priority	Notes
Requirements Specification	Define a comprehensive set of functional, non-functional, and operational requirements for the ATS model and system to support production deployment decision-making	Functional requirements (scoring capabilities across 4 rubrics, rubric definitions, input/output specifications, guardrail behaviour), Non-functional requirements (performance thresholds, scalability, reliability, maintainability), Operational requirements (monitoring, alerting, incident response, data validation, backup and recovery), Model performance requirements (QWK thresholds by rubric, confidence scoring criteria, fairness metrics), Security and privacy requirements (data protection, access control, audit logging), Acceptance criteria for production readiness, Requirement prioritisation (must-have, should-have, could-have with logical consistency), Dependencies mapped and validated	Product owner, technical team, stakeholders, QA team, auditors, architecture review board	Model Specification, Requirements Traceability Matrix, Model Validation and Acceptance Report, Test Strategy and Test Plans, Security Documentation, Privacy Impact Assessment, Model Governance Framework, Architecture Decision Records	IEEE 29148-2018 Requirements engineering; ISO/IEC 23053:2022 Framework for AI systems using ML (requirements considerations); NZISM (NZ Information Security Manual) for government system requirements; ISO 9001 Quality management for requirements traceability	Critical - Prospective	Current specification contains gaps with critical components marked 'TBC'/'TBD', incomplete operational requirements, undefined production acceptance criteria, and logical inconsistencies where must-have requirements depend on should-have capabilities (IS-6, IS-19). Completion essential before production deployment to provide clear acceptance criteria and resolve requirement dependencies. Must reference NZISM for security and privacy requirements. Directly addressed by LR-007.
Requirements Traceability Matrix	Track relationship between requirements, design decisions, implementation, and test coverage	Mapping showing: Requirement ID, Requirement description, Design artefact reference, Implementation reference, Test case reference, Validation status, Notes on implementation decisions	Product owner, technical team, QA team, auditors	Requirements Specification, Test Documentation, Model Specification	IEEE 29148-2018 Requirements engineering; ISO 9001 Quality management for traceability requirements	High - Prospective	Essential for demonstrating that all must-have requirements are met and tested before production deployment.

Appendix 03: Suggested Documentation Catalogue - Model Architecture & Design Decisions

Document Type	Purpose	Key Contents	Audience	Related To (Traceability)	Industry Standards / Frameworks	NZQA Priority	Notes
Model Specification Document	Extensive technical specification of the ATS model architecture, training approach, and configuration	Model architecture (DeBERTa-v3-large), hyperparameters, training data specifications (volume, sources, preprocessing), rubric definitions (Accuracy, Content, Language, Structure), scoring methodology, confidence thresholds, performance baselines	Technical team, auditors, architecture review board, future developers	Architecture Decision Records, Training Documentation, Test Reports, Requirements Specification	Model Cards for Model Reporting - https://arxiv.org/abs/1810.03993; ISO/IEC 23053:2022 Framework for AI systems using ML	Critical - Retrospective	Should document Kaggle competition provenance, DeBERTa-v3-large selection rationale, and any adaptations made from the original competition entry.
Model Decision Records (MDRs)	Document significant ML model architecture and design decisions made during model development	For each decision: Context, Decision made, Rationale, Consequences (positive/negative), Alternatives considered, Status (accepted/superseded). Key decisions to document: Base model architecture selection (DeBERTa-v3-large), Multi-objective scoring approach, Hyperparameter choices, Training methodology, Confidence threshold methodology, Kaggle competition adaptation rationale	Data science team, technical team, future developers, architecture review board	Model Specification, Requirements Specification, Training Documentation	Adapted from ADR format (adr.github.io) but focused on ML-specific decisions; Aligned with Model Cards recommendations for documenting model development decisions	Critical - Retrospective	Particularly important for documenting DeBERTa-v3-large selection over alternatives, guardrail LLM selection, and any trade-off decisions made during PoC development. Distinguishes ML model decisions from broader system architecture decisions.

Appendix 03: Suggested Documentation Catalogue - Model Architecture & Design Decisions _continued

Document Type	Purpose	Key Contents	Audience	Related To (Traceability)	Industry Standards / Frameworks	NZQA Priority	Notes
Training Process Documentation	Comprehensive documentation of model training methodology, data sources, and training outcomes	Training data sources and volumes (38,689 responses documented), Data preprocessing and cleansing procedures, Data splits (training/validation/test), Training methodology and hyperparameters, Training duration and computational resources, Model performance metrics during training, Training data characteristics and limitations	Data science team, auditors, researchers, MLOps team	Model Specification, Data Quality Documentation, Test Reports	Model Cards for Model Reporting - training data section; ISO/IEC 23053:2022 ML framework; Dataset Datasheets (Gebru et al., 2021) - https://arxiv.org/abs/1803.09010	Critical - Retrospective	Training process (data inputs and outputs, sample volumes etc) should document 38,689 responses across 6-month assessment cycles, human marker agreement baselines.
Data Quality and Provenance Documentation	Document data sources, quality controls, cleansing processes, and provenance assurance	Data source specifications (assessment platform snapshots), Snapshot capture methodology (~30 sec to 5 min intervals), Data cleansing and reconstruction procedures, Quality validation checks, Known data limitations, Metadata specifications, Provenance tracking from raw snapshots to model inputs	Data science team, operations team, auditors, quality assurance	Training Documentation, Requirements Specification, Operational Procedures	Dataset Datasheets for Datasets (Gebru et al., 2021); ISO 8000 Data quality standards; Data Management Body of Knowledge (DAMA-DMBOK)	Critical - Retrospective	Must address snapshot-based data collection vs final submission provenance. Document cleansing process that reconstructs learner responses from progress snapshots.

Appendix 03: Suggested Documentation Catalogue - System Architecture & Design Decisions _continued

Document Type	Purpose	Key Contents	Audience	Related To (Traceability)	Industry Standards / Frameworks	NZQA Priority	Notes
Architecture Decision Records (ADRs) - System Level	Document significant architectural and design decisions made for the broader ATS system architecture and integration approach	For each decision: Context, Decision made, Rationale, Consequences (positive/negative), Alternatives considered, Status (accepted/superseded). Key decisions to document: Assessment platform integration approach, Data storage and retrieval architecture, Component interaction patterns (DeBERTa model, LLM guardrails, Explainer component), Deployment architecture, API design decisions	Technical team, operations team, enterprise architects, architecture review board	System Architecture Diagram, Model Decision Records, Operational Procedures, Requirements Specification	ADR format widely adopted in software engineering - see https://adr.github.io/ for templates and tooling; Aligned with ISO/IEC/IEEE 42010:2022 Architecture description	High - Retrospective	Complements Model Decision Records by capturing system-level architectural decisions. Together, MDRs and ADRs provide complete traceability of design decisions from ML model through to system integration. Essential for understanding how the ATS system components interact and why specific integration patterns were chosen.
System Architecture Diagram	Visual representation of system components, data flows, and integration points	Component diagram showing: DeBERTa model, LLM guardrails, Explainer component, scoring pipeline, data inputs/outputs, integration points with assessment platform. Include both logical and physical architecture views	Technical team, operations team, stakeholders, architecture review board	Model Specification, Architecture Decision Records, Operational Procedures	C4 Model for software architecture diagrams - https://c4model.com/; ISO/IEC/IEEE 42010:2022 Architecture description	High - Retrospective	Should show integration between DeBERTa model, LLM guardrails, and Explainer component. Consider including sequence diagrams for scoring workflow.

Appendix 03: Suggested Documentation Catalogue - Component-Specific Documentation

Document Type	Purpose	Key Contents	Audience	Related To (Traceability)	Industry Standards / Frameworks	NZQA Priority	Notes
Guardrail Specification and Validation	Document LLM-based guardrail implementation, validation methodology, and performance	Guardrail architecture and LLM model selection rationale, Safety detection capabilities (off-topic, concerning content), Integration with scoring pipeline, Validation methodology and sample size, Performance metrics (precision, recall, false positive/negative rates), Multi-model fallback logic, Edge case handling, Limitations and monitoring plan	Technical team, safety reviewers, stakeholders, operations team	Model Specification, Architecture Decision Records, Test Reports, Operational Procedures	NIST AI Risk Management Framework - Measure function; ISO/IEC 24028:2020 Trustworthiness concepts; Responsible AI practices from major AI providers	Critical - Retrospective	Document 1,000 sample validation achieving 100% precision for off-topic detection. Include validation of safety concern flagging for concerning content (e.g., self-harm indicators).
LLM Explainer Component Documentation	Document the LLM Explainer implementation, validation, and explanation quality assurance	Explainer architecture and integration, Explanation generation methodology, Validation approach and results, Explanation quality metrics, Explanation-score alignment testing, Sample explanations across score ranges, Known limitations, Operational monitoring plan	Technical team, educational assessment specialists, stakeholders	Model Specification, Test Reports, Requirements Specification	XAI (Explainable AI) frameworks; LIME/SHAP approaches for ML explainability; EU AI Act explainability requirements	High - Prospective	Component recently developed but not yet thoroughly validated. Validation should assess reliability, accuracy, and consistency of explanations across diverse response types.

Appendix 03: Suggested Documentation Catalogue - Testing, Validation & Acceptance

Document Type	Purpose	Key Contents	Audience	Related To (Traceability)	Industry Standards / Frameworks	NZQA Priority	Notes
Test Strategy and Test Plans	Define testing approach, methodologies, and acceptance criteria for model validation	Testing scope and objectives, Test methodologies (statistical validation, guardrail testing, performance testing), Test environments, Test data specifications, Acceptance criteria, Test schedule, Roles and responsibilities	QA team, technical team, stakeholders	Requirements Specification, Test Reports, Model Specification	ISO/IEC/IEEE 29119 Software Testing standards; Special considerations for ML testing per ISO/IEC 23053:2022	High - Retrospective	Should document testing methodologies used for QWK analysis, guardrail validation (1,000 sample validation), and performance benchmarking against human markers.
Test Reports and Validation Results	Document test execution results, validation outcomes, and performance evidence	Test execution summaries, Statistical validation results (QWK scores by rubric: Accuracy 0.81, Content 0.92, Language 0.93, Structure 0.93), Guardrail performance metrics (1,000 sample validation, 100% precision for off-topic detection), Performance comparisons with human markers, Edge case testing results, Test coverage metrics, Known issues and limitations	Technical team, stakeholders, auditors, quality assurance	Test Strategy, Requirements Specification, Model Specification	ISO/IEC/IEEE 29119 Software Testing standards; ISO/IEC 25040 Evaluation process; Quadratic Weighted Kappa for inter-rater reliability	Critical - Retrospective	Consolidate distributed test evidence across documents, notebooks, and teams. Document QWK validation methodology and confusion matrix analysis performed.

Appendix 03: Suggested Documentation Catalogue - Testing, Validation & Acceptance _continued

Document Type	Purpose	Key Contents	Audience	Related To (Traceability)	Industry Standards / Frameworks	NZQA Priority	Notes
Model Validation and Acceptance Report	Document model performance against defined acceptance criteria with empirical evidence demonstrating production readiness and supporting the deployment decision	Acceptance criteria definitions and thresholds (QWK by rubric, confidence scoring, guardrail performance, throughput), Validation methodology and test dataset specifications, Performance results with statistical confidence intervals, Benchmark comparisons (human markers, baselines), Edge case and boundary condition testing results, Fairness assessment summary with reference to detailed Fairness Evaluation Report, Production readiness decision (pass/fail against criteria, deployment recommendation), Known limitations and monitoring recommendations	Senior leadership, stakeholders, architecture review board, decision-makers for production deployment	Requirements Specification, Test Strategy and Test Plans, Test Reports and Validation Results, Model Specification, Training Process Documentation, Fairness Evaluation and Bias Testing Report, Requirements Traceability Matrix	ISO/IEC 25040 Evaluation process; Model Cards for Model Reporting (performance metrics section) - https://arxiv.org/abs/1810.03993; ML Test Score rubric for production ML systems; IEEE 29119 Software Testing standards	Critical - Retrospective	Notes: Consolidates distributed test evidence to provide definitive production readiness assessment. Essential for demonstrating due diligence and providing clear audit trail for deployment decision. Directly addresses IS-12 (Testing Evidence Consolidation) and supports LR-007 (Requirements Specification Completion). Should document QWK scores by rubric (Accuracy 0.81, Content 0.86, Language 0.86, Structure 0.87), guardrail validation results (1,000 sample validation, 100% precision for off-topic detection), and overall pass/fail deployment recommendation.
Fairness Evaluation and Bias Testing Report	Document fairness evaluation methodology and results across student demographic groups	Fairness evaluation framework and methodology, Student demographic group definitions, Disaggregated performance analysis by group, Bias metrics and fairness indicators, Bias mitigation strategies implemented, Limitations and recommendations, Ongoing monitoring plan	Data science team, stakeholders, Ministry of Education, public (potentially)	Test Reports, Requirements Specification, Model Specification, Operational Procedures	NIST AI Risk Management Framework - https://www.nist.gov/itl/ai-risk-management-framework; EU AI Act fairness requirements; Aequitas bias audit toolkit methodology	Critical - Prospective	Framework under development. Essential for public sector accountability and demonstrating equitable assessment across all student populations, including considerations for Māori and Pasifika students.

Appendix 03: Suggested Documentation Catalogue - Production Readiness & Operations

Document Type	Purpose	Key Contents	Audience	Related To (Traceability)	Industry Standards / Frameworks	NZQA Priority	Notes
Operational Procedures and Runbook	Define operational procedures for production deployment, monitoring, and incident response	Deployment procedures, Monitoring and alerting configuration, Performance monitoring (data drift, model degradation, fairness metrics), Incident response procedures, Escalation paths and decision authority, Rollback procedures, Model versioning and change management, Routine maintenance procedures, Troubleshooting guides	Operations team, technical team, support staff, incident responders	Model Specification, Requirements Specification, System Architecture	ITIL 4 for IT service management; ISO/IEC 20000 Service management; Site Reliability Engineering (SRE) practices; MLOps best practices	Critical - Prospective	Build upon existing operational manual templates to define in-house model versioning, monitoring, and troubleshooting frameworks, and rollback capabilities.
Security Documentation	Document security controls, threat modelling, and security testing for the ATS system	Threat model and risk assessment, Security controls implemented, Data encryption (in transit and at rest), Access controls and authentication, Vulnerability assessment results, Security testing outcomes, Compliance with NZISM requirements, Incident response procedures for security events	Security team, operations team, auditors, compliance officers	Requirements Specification, System Architecture, Privacy Documentation, Operational Procedures	NZISM (NZ Information Security Manual) - https://www.nzism.gcsb.govt.nz/; ISO/IEC 27001:2022 Information security management; OWASP for application security; NZ Privacy Act 2020 security requirements	Critical - Prospective	Must align with NZISM requirements for New Zealand government organisations. Consider cloud security if using cloud-based infrastructure.

Appendix 03: Suggested Documentation Catalogue - Production Readiness & Operations _continued

Document Type	Purpose	Key Contents	Audience	Related To (Traceability)	Industry Standards / Frameworks	NZQA Priority	Notes
Privacy Impact Assessment and Documentation	Assess and document privacy risks, controls, and compliance with Privacy Act 2020	Privacy Impact Assessment (PIA), Personal information collected and processed, Privacy risks identified and mitigations, Data retention and disposal procedures, Consent and transparency measures, Privacy controls implemented, Compliance with Privacy Act 2020 and NZISM, Privacy breach response procedures	Privacy officer, legal team, stakeholders, Office of the Privacy Commissioner	Requirements Specification, Security Documentation, Data Quality Documentation	Privacy Act 2020 (NZ) - https://www.privacy.org.nz/; NZISM privacy requirements; ISO/IEC 29134:2017 Privacy impact assessment; Office of the Privacy Commissioner PIA guidance	Critical - Prospective	This should address student response data, assessment data, and any personal information processed by the ATS system. Consider Māori data sovereignty principles.
Model Governance Framework	Define governance structure, roles, responsibilities, and decision-making authority for the ATS model	Governance structure and committees, Roles and responsibilities (model owner, data science team, operations, stakeholders), Model lifecycle management, Change management processes, Risk management framework, Compliance and audit procedures, Stakeholder communication plan, Decision-making authority matrix	Senior leadership, model owner, technical team, stakeholders, auditors	Operational Procedures, Requirements Specification, All documentation	COBIT 2019 for governance; ISO/IEC 38500:2024 Governance of IT; AI governance frameworks from OECD, IEEE, ISO/IEC	Critical - Prospective	Essential for demonstrating AI system stewardship expected of government organisation. Should define clear accountability for production operation, change management, and ongoing assurance.

Appendix 03: Suggested Documentation Catalogue - Navigation & Maintenance

Document Type	Purpose	Key Contents	Audience	Related To (Traceability)	Industry Standards / Frameworks	NZQA Priority	Notes
Model Landing Page / Runbook	Central entry point providing overview and links to all model documentation (one page per model)	Model overview and purpose, Current version and status, Key performance metrics (QWK scores, confidence levels, throughput), Quick links to all supporting documentation, Responsible contacts, Recent changes and updates, Known issues and limitations, Quick start guide for new team members	All stakeholders - technical team, operations, management, auditors	All documentation - serves as navigation hub	System runbook best practices from SRE; Knowledge management approaches; Model registry concepts from MLOps	Critical - Prospective	Essential for scaling to multiple models - provides consistent entry point. Serves as the primary navigation hub for all model documentation; without this landing page, documentation remains fragmented even if individual documents exist. First point of reference for operations team, new team members, and stakeholders.
Code Documentation and Comments	In-code documentation explaining implementation decisions, parameters, and logic	Inline code comments explaining decision rationale, Parameter selection documentation, Configuration variable explanations, Complex logic explanations, API documentation, Dependency specifications, Setup and installation instructions	Developers, technical team, code reviewers, future maintainers	Model Specification, Architecture Decision Records	PEP 257 Python docstring conventions; Clean Code principles (Robert Martin); IEEE 1063-2001 Software User Documentation	High - Retrospective	Code review identified opportunities to add comments, explain parameter selection, and document configuration. Essential for maintainability and onboarding.
Version Control and Change Log	Track model versions, changes, and evolution over time	Version numbering scheme, Change log for each version, Model artefacts per version, Performance metrics by version, Breaking changes documented, Rollback procedures, Deprecated features	Technical team, operations team, auditors	Model Specification, Operational Procedures, Model Governance	Semantic Versioning - https://semver.org/; MLflow Model Registry concepts; Git version control best practices	High - Prospective	Essential for production operations. Must be established before deployment to track model iterations, performance changes, and enable rollback if needed.

Diligence Statement

In the creation of this document, we used Enterprise Generative AI to assist in text creation and refinement. We affirm that all AI-generated content underwent thorough vetting, editing, and curation by the human co-authors.

The final document accurately reflects our understanding, expertise, and intended meaning. While AI tools were included in the writing process, we maintain full responsibility for the content, its accuracy, and its presentation.

This disclosure is made in the spirit of transparency and to acknowledge the evolving role of AI in content creation and other intellectual work.

AI-generated content may include information such as metadata, digital signatures, or watermarks to identify that it was generated using a generative artificial intelligence model (provenance markers).

clearpoint.

Thank you.