AI Literacy Marking Tool Technical Review Report

FINAL REPORT

August 2026

NZQA | Mana Tohu Mātauranga o Aotearoa
New Zealand Qualifications Authority

clearpoint

NZQA Management comment

Mana Tohu Mātauranga o Aotearoa
New Zealand Qualifications Authority

MANAGEMENT COMMENT

August 2026:

We thank ClearPoint for the analysis and insight in this report and are pleased at the finding that NZQA has developed a lean and adaptable platform that follows sound, industry-standard engineering practice.
We fully accept the recommendations outlined in this report and appreciate the independent assessment of the progress made against ClearPoint's recommendations from its review of NZQA's ATS model prior to production use. We also welcome the observation that overall, the platform is well suited to current production use and well positioned for the future, and we agree that addressing the automation and operational-documentation items before scaling should support repeatable, sustainable growth.

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, establishing a platform for scalable use.

Table of Contents

1. About this Review
2. Executive Summary
3. Approach & Methodology
4. Recommendations
5. Review
6. Findings
7. Appendices

(Links navigable)

Report disclaimer

This report relates to an independent technical review completed within the scope defined in the engagement SOW (NZQA-2101) between ClearPoint Limited (ClearPoint) and NZQA.

The review covered three areas:

- the design and as-built implementation of the ATS production pipeline
- the remediation of the recommendations from the 2025 review
- and the solution's scalability and future extensibility, considered at an options level

This technical review does not cover; the independent validation of marking results, model performance or statistical outcomes, including accuracy, equity and bias; assessment standards other than Literacy Writing (32405); implementation, remediation or build of any recommendations arising from this review; formal security assessment, penetration testing, or review of data handling and privacy; evaluation of, or commercial comparison with, third-party or vendor solutions; organisational change management, integration planning, procurement, and long-term maintenance strategy; and the design of future-state architecture or a delivery roadmap.

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

In preparing this report, ClearPoint has assumed that:

- there has been no material change in that information since the review;
- that the responses to ClearPoint's queries were true, correct, accurate and complete;
- and that the NZQA representatives who responded were competent to answer.

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

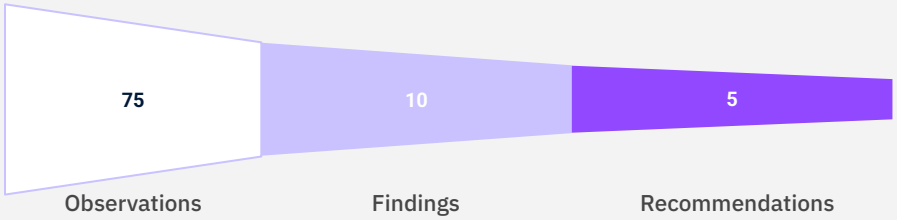
This report is addressed to, and may be relied on solely by, NZQA, for the sole purpose of providing an independent technical review of the ATS solution as implemented and operated in the live AE1 2026 environment, as defined in the SOW. To the extent NZQA makes it publicly available, it is provided for informational purposes only on an "as is" and "as available" basis; no other person may rely on it, and to the maximum extent permitted by law ClearPoint has no liability to any third party relying on it.

About this Review

Background

NZQA has implemented an end-to-end Automated Text Scoring (ATS) solution for the Co-requisite Literacy Writing Assessment (standard 32405), building on its 2025 AI marking pilots and supported by human marking and external assurance.

In late 2025, ClearPoint conducted an independent technical review of the ATS model. Its recommendations informed NZQA's preparation for live use, and the solution has since moved to production, completing its first live marking run in Assessment Event 1 (AE1) 2026.



ClearPoint has been engaged to undertake a follow-up review of the solution as operated in that live AE1 2026 environment.

The review covers three areas:

- the design and as-built implementation of the AWS production pipeline;
- the extent to which the 2025 recommendations have been addressed;
- and the solution's scalability and future extensibility, considered at an options level.

The engagement ran from 22 June to 24 July 2026, followed by a feedback window through to 26 August 2026 during which NZQA provided further evidence in response to the draft report.

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

Executive Summary

Executive Summary

Significant progress has been made since ClearPoint's 2025 review.

Working with AWS Professional Services, NZQA has moved from a proof-of-concept, piloted alongside human marking, to a compact, cloud-native production platform that aligns with the AWS Well-Architected Framework and follows sound, industry-standard engineering practice. The solution completed its first live marking run in AE1 2026.

The recommendations from the 2025 review have largely been addressed.

ClearPoint is of the opinion that pragmatic decisions have been taken, with effort focused where it matters most and other areas appropriately left to mature as the solution's requirements evolve and become clearer.

The result is a lean and adaptable platform. ClearPoint has identified some areas for development, ideally addressed before the platform scales to additional models or assessment events; these are set out in the findings and recommendations in this report.

Executive Summary _continued

Key strengths:

- A compact, cloud-native AWS platform aligned with industry-standard engineering practice, with strong observability, security and infrastructure-as-code.
- Substantial remediation of the 2025 recommendations, supported by thorough remediation documentation and model-registry traceability.
- A fundamentally scalable architecture suited to the asynchronous, event-driven workload.

Areas requiring attention:

- Automation gaps: source-data retrieval (an external-provider dependency), model training and testing, and inference orchestration and automated quality gates for code changes.
- Operational support documentation, such as runbooks, failure modes and incident response, is still maturing.
- Fairness monitoring: an in-house capability is emerging alongside NZCER's external evaluation, with the supporting monitoring framework yet to be documented.

Overall, the platform is well suited to current production use and well positioned for the future. Addressing the automation and operational-documentation items before scaling should support repeatable, sustainable growth.

Approach & Methodology

Approach & Methodology

ClearPoint follows a structured three-phase methodology that maintains traceability from evidence to recommendations. This post-implementation review applies the same method, drawing on the documentation, interviews and technical evidence provided by NZQA.

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.

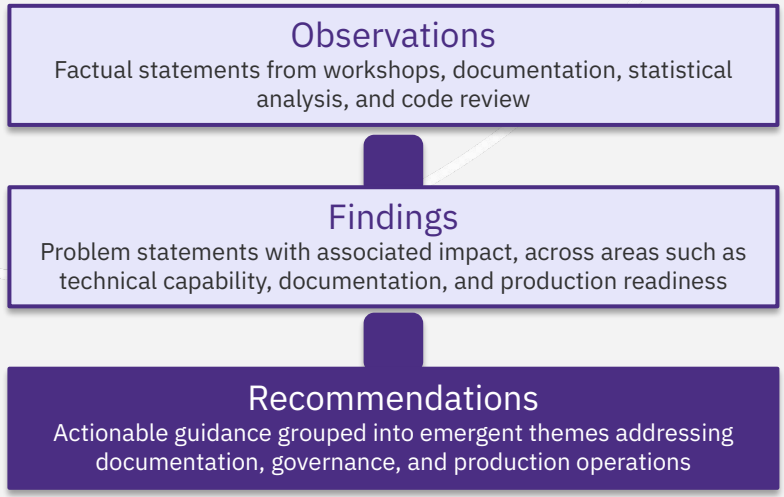
Phase 2: Findings

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.

NZQA ATS Review Application



Recommendations

R-001

Technical Foundation & Code Quality

Related findings:

IS-1: There is no automated code-quality scanning or quality gate on pull requests

RECOMMENDATION

Ensure all pull requests are automatically scanned and gated:

The production codebase is clean, concise and well-documented. Codebases evolve over time, however, and engineers will continue to modify and extend it as requirements develop.

ClearPoint recommends introducing automated checks in the CI/CD pipeline that run before a code contribution is merged and released, covering at a minimum:

- Automated code-quality scanning (for example, SonarQube)
- Automated security vulnerability scanning
- Automated tests covering critical logic and integrations

ClearPoint note that NZQA has expressed an intention to implement code-quality scanning.

R-002

Production Operations & Governance

Related findings:

IS-3: Learner responses cannot be retrieved programmatically from the external provider

IS-4: Once downloaded, learner responses require manual cleaning before they can be used

RECOMMENDATION

Engage the data provider to improve source-data delivery:

The ability to automatically retrieve source data that is ready for use by inference runs is important to the ongoing operation and likely scaling of the platform. At present these files must be downloaded manually, as the provider offers no alternative, and the source data must then be cleaned up in the early stages of the inference job.

The manual download step is a bottleneck that requires manual effort, and the need to clean the data introduces a risk of source data being altered, which can blur the lines of responsibility between provider and consumer.

ClearPoint recommend that NZQA engages with the data provider to develop a more robust, production-grade data service, for example:

- Data delivered directly to an S3 bucket (owned by either NZQA or the provider)
- **And/or** an HTTPS API from which NZQA can request the data
- With the delivered data pre-sanitised and requiring no reassembly

R-003

Production Operations & Governance

Related findings:

IS-6: Model approval depends on manual testing and oversight for each new or retrained model

IS-7: Model training and retraining are manually coordinated

RECOMMENDATION

Automate the model training and testing workflows:

The model training and testing workflows are documented, but largely manual at present. This is understandable at this early stage, but will become increasingly difficult to sustain as scaling pressures increase.

NZQA has expressed an intention to address this in the medium-to-long-term roadmap; ClearPoint includes this recommendation to underline its importance as the platform scales.

As with other aspects of the platform, it may be helpful to establish a code repository for each new model, holding the documentation, tests and test data that slot into the workflows.

R-004

Production Operations & Governance

Related findings:

IS-5: There is no facility to run an inference job using a previous version of a model

IS-9: Inference runs are manually orchestrated, including starting and configuring the step functions

IS-10: The naming scheme for inference runs contains only a date and a sequential number

RECOMMENDATION

Develop a front-end interface for inference runs:

Effort has been applied to provide a custom front-end for the model registry and approval workflow, though the inference workflow still depends on a degree of manual control and oversight.

Once inference is running it *is* orchestrated by the AWS serverless framework, but its instantiation and configuration still require a person to interact directly with the AWS console in production accounts.

This may not be a major concern at this early stage, but as the team and workloads grow there would be value in building automation around inference jobs to improve stability and reduce the risks associated with manual operation. Even a basic front-end interface would be beneficial, including:

- Automated invocation of inference jobs (dependent on automated source-data access or a catalogue of pre-downloaded data)
- Automated generation of unique inference-job IDs (date, event name, model, version)
- The ability to re-trigger historic runs using specific model versions
- An audit log of who ran each job and its progress
- Handling of failures, retries and delivery of results

R-005

Production Operations & Governance

Related findings:

IS-2: On-call and support arrangements are not clearly defined

IS-8: Operational documentation is incomplete

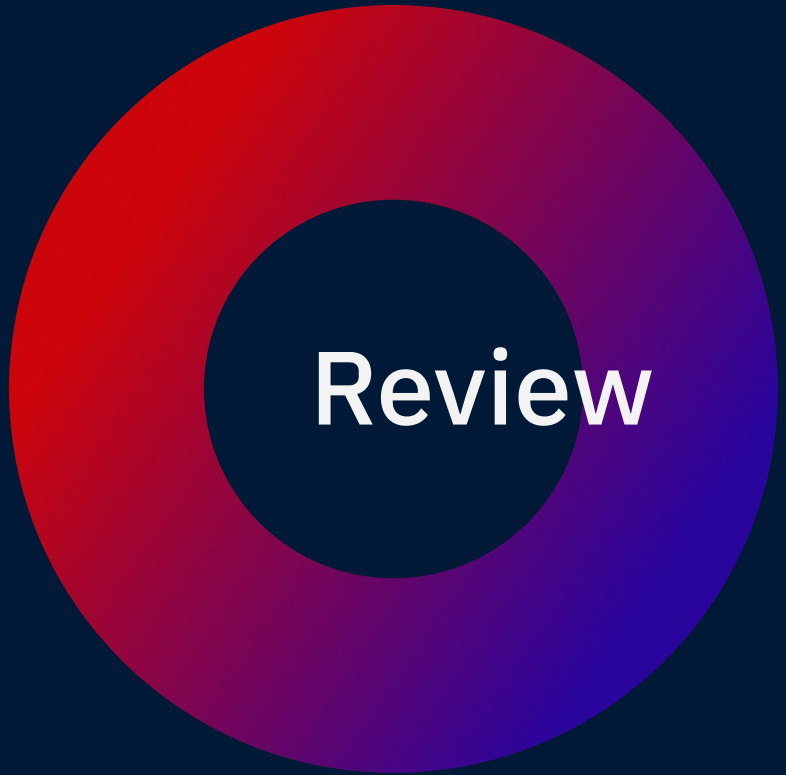
RECOMMENDATION

Continue to build out operational capabilities and documentation on a regular review cycle:

The platform's operational capabilities and supporting documentation are still developing, and will need to keep pace as the team grows and the number of assessment events and models increases.

ClearPoint recommends that NZQA build these out progressively, with clear ownership and a regular review cycle:

- Complete the operational support documentation, prioritising support runbooks, common and historic failure scenarios, fault-diagnosis procedures, incident response procedures, and a post-incident review process (with history).
- Develop the supporting operational capabilities: support rota and handover, including notification management, so platform alerts are routed to a named responder with clear responsibility for response.
- Review and update this material on a regular cadence, for example after each assessment event and following any incident, so it stays current as the platform evolves.



Part 1: As-built review

This section reviews the ATS solution as designed and built: the AWS production pipeline, its architecture and as-built documentation as run live in AE1 2026, and its alignment with sound, industry-standard engineering practice.

Part 1: Solution design - Overview

Overview:

The platform built to support the ATS model, and future automated text scoring models for the literacy exams, was designed and implemented by AWS Professional Services in line with requirements defined by NZQA following the early proof-of-concept and pilot phases.

NZQA has provided a high-level design document that gives a thorough overview of the solution ("**NZQA ATS HLSO**").

The architecture reflects the two core workflows:

- Developing, hosting, testing and approving models (*registry*)
- Using those models to evaluate responses to literacy assessment events (*inference*)

Part 1: Solution design _continued

Models and the model registry workflow:

Model development and training are carried out within Databricks, an environment the data science team is comfortable and familiar with. Data is brought into this environment for training and, once the models are ready, exported to S3 buckets. This triggers the serverless MLOps process, which brings the new model, associated metadata and performance testing results into the SageMaker model registry. The process is performed once for the dev/QA environments and again for non-prod/prod.

Notifications are issued once the model and its metadata are in the registry, which request approval from a list of four approvers. The approvers manually perform their tests, and their approvals are recorded against each model version in the registry. As sequential approvals are granted (through dev, QA, non-prod and prod), the new model version becomes the default for each environment.

The intention is to re-train the model(s) before each major event. The training, registry and approval workflow is designed to operate independently of any inference jobs running at the time, so new models can be brought in, tested and approved while inference jobs continue to run.

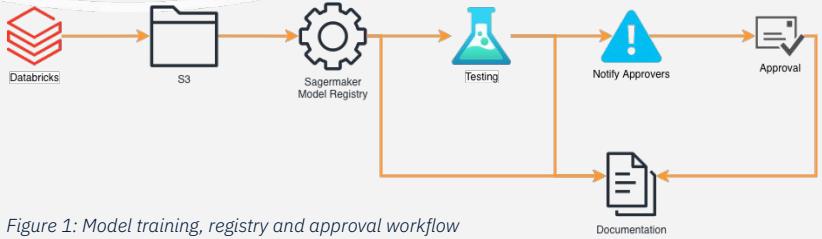


Figure 1: Model training, registry and approval workflow

Part 1: Solution design *_continued*

Inference Workflow:

The inference workflow for ATS begins with a manual process to download the source data. The external provider does not currently offer any programmatic means of retrieving this data, which is a known limitation; it means the process cannot yet be fully automated or re-run without human intervention. The invocation of inference runs is manually orchestrated at this stage. ClearPoint encourage NZQA to prioritise improvements in this particular area, bringing inference in-line with the model registry and approval workflow.

The inference job itself is coordinated by a step function that invokes Lambda functions for each stage. The Lambda functions in turn invoke SageMaker compute to perform the actual inference, with the results written to an S3 bucket. A final, intentionally manual, step checks the results before they are sent for consumption.

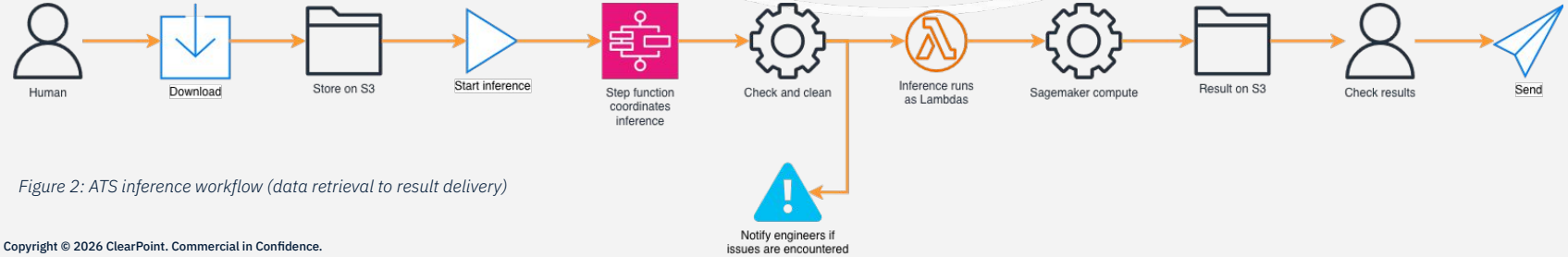


Figure 2: ATS inference workflow (data retrieval to result delivery)

Part 1: Solution design _continued

Infrastructure:

The platform was designed and built by AWS Professional Services and aligns with the AWS Well-Architected Framework, which sets out good practice across operational excellence, security, reliability, performance efficiency, cost optimisation and sustainability.

The multi-account layout for ATS sits within NZQA's existing AWS organisation, with the foundational controls you would expect: CloudTrail forwarding to a log-archive account, SSO for access, account creation and base networking handled by external teams. Periodic external security scanning is performed using CrowdStrike.

The ATS solution currently occupies four separate AWS accounts in the ap-southeast-2 region: dev, QA, non-prod and prod. A fifth account hosts the production model registry, which is intended to be shared not only by ATS but also by additional future models.

Part 1: Solution design _continued

Observability:

Observability is provided through the AWS-native CloudWatch service, which allows alerting and notifications to be defined based on metric thresholds. Given the AWS-native nature of the solution, meaningful metrics are readily available, allowing the core workflows to be monitored with little additional effort. Core metrics are derived from sources such as Lambda function invocations and step-function status.

Security:

AWS account-level isolation has been employed, with the principle of least privilege applied to user-level access in each environment. Organisation-level guardrails are in place, along with processes that provide an audit trail. The cloud-native nature of the ATS solution makes infrastructure-level audit possible, and regular security scanning/auditing of these resources is being performed by third parties. There is room for further measures, some of which are raised as issues and follow-up recommendations as a part of this report.

Good practise has been observed across the board, but ideally NZQA will continue to work towards a scenario where direct AWS console access is no longer required for the day-to-day operation of the core inference workflow.

Part 1: Solution design _continued

Resilience:

The platform is not currently built for high availability in the classic sense. This reflects the operating model rather than a limitation; there is no requirement for real-time 24/7/365 availability, there is no critical customer-facing element, and the workloads are asynchronous. The event-driven, asynchronous architecture also supports resilience, as processes can be retried in the event of failure.

Automation:

The infrastructure and cloud resources are managed through Python/CDK stacks registered in AWS Service Catalog within a shared ("tooling") account, satisfying an infrastructure-as-code capability. The approach adds some complexity and maintenance overhead, but keeps the solution within a single provider (AWS) and allows fine-grained customisation.

Some aspects *are* not yet fully automated; source-data retrieval is currently manual, and test automation is still under development. At this stage in the platform's evolution these are not considered major concerns, but they should be addressed before scaling, whether increasing volume within the current scope or extending beyond ATS marking to additional standards.

Part 1: Solution design _continued

As noted previously, this platform was built in collaboration with AWS Professional Services and reflects sound, industry-standard engineering practice, aligned with the AWS Well-Architected Framework, which focuses on simplicity, reliability, scalability, and cost-effectiveness. The result is a compact cloud-native solution with a bare minimum of moving parts. The model registry pattern is not unique to NZQA, and is likely in use in a number of different organisations around the world. This puts the platform in a good position to be able to benefit from the learnings of others, and to avoid obscure custom solutions likely to become an ownership burden.

Assessing the platform and its processes against a relevant selection of ClearPoint's foundational engineering capabilities (F25)
* indicates the level of maturity and best-practise in the technology stack:

- Code Maintainability
- Test Automation
- Immutable Infrastructure
- Continuous Delivery
- Security Automation
- Pipelines as Code
- Observability
- Infrastructure Automation
- Test Data Management
- Resiliency Engineering
- Modern Architecture

** for more information please see Appendix 02*

Part 2: Remediation of the 2025 Recommendations

NZQA have provided a snapshot of pertinent documentation extracted from their Confluence spaces. This archive gives detail to a catalogue of remediations undertaken since the previous review in 2025. These documents were sufficient to immediately satisfy many of the requirements.

In addition to the documentary evidence, a series of focused miniature workshops was conducted. These allowed us to follow up on any outstanding detail required to satisfy the remaining recommendations.A full list of these documents and workshops is provided in the appendix section of this report.

Following issue of the draft report on 6 August 2026, NZQA provided further evidence in response to review comments. This remediation window closed on 25 August 2026, and the statuses recorded in this section reflect the evidence available at that point.

Part 2 : Remediation of the 2025 Recommendations

Recommendation	Status	Description
LR-001	●	The original scripts have been superseded by a production Lambda and Step Functions codebase, held in a structured monorepo.
LR-002	●	The guardrail solution (Bedrock batch, Sonnet 4.6) is sound and its resilience is documented; the approach is consistent with NZQA's architectural principles.
LR-003	●	Model validation is evidenced by NZCER's rubric-level analysis, and the LLM Explainer has been assessed and retired with documented rationale.
LR-004	●	A centralised Confluence repository is in place, with the 2025 retrospective items (provenance and licensing, cleansing, decisions) documented.
LR-005	●	Model and ML documentation follows recognised frameworks (model card, dataset Datasheet, ADRs), with registry-based traceability.
LR-006	◐	Model governance is well documented; operational support documentation (runbooks, failure modes) is still maturing.
LR-007	●	The requirements have been completed and updated, aligned with NZISM and supported by security, privacy and CAB evidence.
LR-008	●	Production readiness is achieved (live AE1 2026 run); the residual manual data-handling is externally constrained.
LR-009	◐	Ongoing fairness evaluation is performed externally through NZCER at each event to-date; the outstanding element is developing NZQA's internal capability to perform this function in-house.

LR-001: Code quality improvement

Addressed

Since ClearPoint's review late 2025, NZQA has undertaken significant redevelopment of the ATS platform and the workflows it supports. What was considered to be a proof-of-concept or pilot platform has evolved into a production-ready solution, detailed in the as-built review. Recommendation LR-001 identified the quality of the two main scripts that, at the time, handled the training and inference workflows. Those scripts no longer exist; they have been superseded by a suite of Lambda functions coordinated by AWS Step Functions and asynchronous workflows built on AWS's serverless framework.

The code is now held in AWS CodeCommit, in a monorepo named ats-mlops that covers inference scripts, Lambda functions, CDK (infrastructure-as-code) and supporting content such as documentation and scripts. Comments and doc-strings were observed throughout the repository, which was organised in a clearly-defined directory structure .

Although not part of the original ClearPoint recommendations, it is worth noting that there is currently no automated review or quality gate for pull requests in the CI/CD pipeline. Code quality therefore relies on manual review alone, which increases the risk of defects, inconsistencies or security issues reaching the main branch, and provides less assurance as the codebase, team and range of assessment events grows.

LR-002: Guardrail Optimisation Research Spike

Addressed

The ATS inference workflow includes steps to measure additional qualities in each response

- whether the response is on-topic (relevance)?
- whether it signals harmful intent?

These checks are performed using Amazon Bedrock's batch LLM service with Claude Sonnet 4.6. NZQA has assessed the solution as fit for purpose and is satisfied with its accuracy, runtime and cost. On this basis, NZQA decided not to proceed with the spike to investigate alternative approaches such as fine-tuning or smaller models. The "**Guardrails**" document describes the checks themselves, and the "**Low Level Design - Inference Pipeline**" document sets out how they are executed within the inference workflow.

The decision to provide this functionality with Bedrock is deemed to be in line with NZQA’s architectural principles. The absence of a documented rationale for discounting non-conforming options is reasonable in this instance.

LR-003: Model validation and performance investigation

Addressed

Recommendation LR-003 covered two main aspects. The first concerned validation of the ATS model, in particular the scoring patterns of the Accuracy rubric, which had shown a slightly more conservative pattern than the other rubrics at the time of the 2025 review.

NZQA engaged NZCER to carry out an in-depth analysis, documented in the "**NZQA ATS model Review Report**". This examines the model's performance at the rubric level, including the Accuracy rubric, and concludes that its scoring patterns are in line with expectations and consistent with human marking, including across both "achieved" and "not achieved" responses.

The second aspect concerned an "LLM Explainer" component that existed at the time of the 2025 review. NZQA assessed this under its transparency and explainability approach and determined that AI marking falls under a medium-risk classification, with no business requirement for explainability at an individual student level. On this basis the component was not progressed and is no longer part of the ATS platform; NZQA instead provides transparency through information published on its website, covering a high-level overview of the solution and how the model applies the assessment schedules.

LR-004: Centralised Documentation Repository with Retrospective Capture

Addressed

There are several distinct documentation requirements for the platform, each of which has been addressed. LR-004 sought a centralised documentation repository with clear rationale for key decisions, together with retrospective capture of information missing from the earlier proof-of-concept; both have been met.

Documentation is held across two Confluence spaces, with version-specific model documentation managed separately by the model registry.

The retrospective items identified in 2025 have since been documented - provenance and licensing, architectural decision rationale and data cleansing, with guardrail performance metrics and testing and evaluation evidence covered through the functional-testing documentation and NZCER's review.

Collectively, this provides coverage across the platform's infrastructure, design, decisions, model, operations and end users, satisfying the intent of LR-004.

- **Platform infrastructure:** the AWS multi-account environment and networking.
- **Low-level design:** component-level design of the training, inference and approval workflows.
- **Architectural decisions:** rationale for key technical choices, including the scoring model.
- **Model details:** model card, training data and provenance.
- **Operational docs:** MLOps, runbooks and approval/inference procedures.
- **End-user guides:** batch inference and model-approval workflow guides.

Optional future enhancement: consolidate the two Confluence spaces into one with a clear hierarchy.

LR-005: Documentation Standards Aligned to Software Engineering Practice

Addressed

This recommendation was to develop and implement documentation standards for the model and ML aspects of the platform, aligned with software-engineering practice, allowing future audiences to follow the design decisions, testing, development and real-world performance which make up the history of each model.

NZQA's response to this comes in two forms, the first being the general Confluence documentation repository covering:

- Testing, training, validation methodologies
- Performance review reports
- Evaluation protocols
- Operational procedures
- A model card and a dataset Datasheet

The second part to the solution is the AWS MLOps model-registry (discussed previously in this document). This captures the history of each model version in the system, including its testing and performance status, lineage, and approval chain.

The model registry and associated workflow detailed can be found in the following documents:

- **“Low Level Design - Model Approval Workflow”**
- **“Low Level Design - Model Ingestion”**

These artefacts adopt recognised frameworks such as Model Cards and Datasheets for Datasets. ClearPoint is satisfied that appropriate documentation is recorded and will continue to be captured by these measures. As further models are introduced, consolidating these practices into a defined documentation standard would help maintain consistency at scale.

LR-006: Model Governance and Operational Procedures

In Progress

Model governance is well documented. The "**ATS MLOps Governance and Operations**" document, together with the **model approval low-level design**, sets out model versioning and stage-based promotion (Development, QA, Pre-production and Production), change management through the NZQA Change Approval Board, defined approver personas and decision-making authority at each stage, performance monitoring (including data drift detection, performance-degradation tracking and ongoing fairness evaluation via the re-training approach), rollback (an unsatisfactory re-trained model is discarded and the previous production model retained), and incident response through the Assessment Division's incident-management policy, register and templates.

Operational documentation has been prepared, which provides a foundation for understanding how the various system components and workflows tie together.

A good amount of detail is given in the low-level system documentation around how the observability suite and notification mechanisms are designed. Procedural detail is provided in documents such as "**End User Guide - Model Approval Workflow**" and "**End User Guide - Batch Inference Workflow**", which cover the two main recurring tasks.

Governance is well documented and fully addressed; the in progress status reflects only the operational support documentation, which is still maturing (see next slide).

LR-006: Model Governance and Operational Procedures

_continued

In Progress

ClearPoint has seen evidence of support documentation, which will continue to grow as the platform develops and operational experience is gained.

Several elements are already in place:

- High-level design (the "NZQA ATS HLSO")
- End-user guides for the main recurring tasks (model approval and batch inference)
- Incident response and escalation (the Assessment Division's incident-management policy, register and templates)
- Post-incident review archive (the Assessment Division's PIR register)

Others are still to develop as the platform matures:

- Runbooks (restart and repair procedures)
- Common failure modes
- Fault-diagnosis procedures

As the platform is newly in production and is not a real-time 24/7 service, these are maturation items rather than production blockers.

It is expected that these sections will grow as production issues are encountered and added to the operational documentation sections.

LR-007: Requirements

Addressed

NZQA formally documented the requirements for the ATS platform, to the point where they were issued to prospective service providers for design proposals. These requirements were subsequently used to procure and build the platform in partnership with AWS Professional Services.

The requirements are captured in the **“Automated Text Scoring Requirements”** document.

The updated document closes the gaps identified in LR-007. It aligns explicitly with the NZ Information Security Manual (NZISM) across authentication, logging, incident response and risk management; adds privacy documentation and a required Privacy Impact Assessment; and defines production acceptance criteria with specific metrics and acceptance gates. The earlier priority inconsistencies have also been resolved, with dependent requirements promoted to "must have". These are further supported by an independent security risk assessment, a privacy threshold analysis, and Change Approval Board sign-off for the production deployment (CR-986).

Footnote: Formal security and privacy assessment is outside the scope of this review; these documents are referenced as evidence of remediation rather than independently validated by ClearPoint.

LR-008: Data Quality, Fairness Evaluation, and Production Readiness

Addressed

Considerable development effort has gone into the production deployment of the ATS platform since the proof-of-concept seen in 2025. The pipeline has been automated wherever practical. The principal exception is the retrieval and pre-processing of source data, which cannot currently be automated due to the external provider's limitations; some further steps (model training and testing) also remain manual by design at this stage, as discussed in Part 3.

Data quality and provenance are documented. The data-cleansing process (including preservation of Te Reo Māori macrons) is set out in the "Essay text cleansing" material, and model provenance and licensing are covered separately. Fairness evaluation is addressed through NZQA's engagement with NZCER and its developing in-house monitoring (see LR-009), and sits outside the scope of this review.

A more detailed analysis is provided in the as-built review (Part 1), where the technology stack and processes are assessed against a relevant selection of ClearPoint's foundational technology capabilities.

Details can be found in the following documents provided to us by NZQA:

- **"AE1-2026 Data flow with automations"**
- **"End User Guide - Batch Inference Workflow"**
- **"Essay text cleansing"**

The platform successfully completed a live production marking run in AE1 2026, and NZQA has addressed the elements within its control. The remaining manual data-handling steps depend on the external provider offering programmatic access; these are outside NZQA's sole control and are carried forward as findings and a recommendation.

LR-009: Ongoing Fairness Monitoring and Bias Testing

In Progress

LR-009 recommended that NZQA establish internal frameworks for ongoing fairness monitoring and bias testing, building on NZCER's initial fairness evaluation. NZQA has begun to develop an in-house capability, with a first analysis of AE1 2026 output provided in "**AIML-Fairness analysis of ATS model output**", prepared using demographic data supplied independently of the team running the analysis.

The stated intention is to perform internal fairness monitoring alongside NZCER's external testing at every assessment event, until the internal solution is deemed to be of sufficient maturity. This establishes the ongoing, periodic monitoring cadence that LR-009 envisaged, and maintaining the NZCER arrangement in the interim is a pragmatic position.

The outstanding element is the framework rather than the analysis. Documented protocols, the thresholds at which a result becomes a concern, cadence and ownership, and the response path when a threshold is exceeded. As a consideration, NZQA may wish to complement its in-house capability with periodic independent external review, to reinforce public assurance of sustained algorithmic fairness.

Part 3: Scalability & Future Extensibility

The NZQA platform is built on a scalable technology stack that should be capable of scaling in several ways as required, without carrying excessive overhead.

The asynchronous nature of the workloads suits the AWS serverless suite, which provisions compute capacity in response to event triggers and shuts it down again once the work is complete.

The following section explores the likely scaling scenarios for the platform, along with their implications and potential mitigations.

Part 3: Scalability & Future Extensibility

Increased number of responses per ATS event:

The current platform could reasonably be expected to handle an increase in the number of responses without significant modification.

Some steps would simply take longer to execute, such as downloading data and exporting results, while the SageMaker compute can be scaled vertically and horizontally if required.

Even without scaling the compute, the existing inference workflow could simply be allowed to run for longer.

Increased frequency of ATS events per year:

There are currently two ATS events each year, processed using ephemeral compute resources.

The only real limit on their frequency would be the point at which the retraining and inference workflows take longer than the interval between events allows (for example, running every day rather than every six months).

This is not considered a constraint at present.

Part 3: Scalability & Future Extensibility *_continued*

Introduction of additional subject assessments:

Additional subject assessments can be supported by hosting further models, which the platform is designed to accommodate.

This is not expected to cause difficulties, though some scripts may need updating to handle the input and output data for new, as-yet-unknown models.

The core workflow scripts are built to work with different models, and infrastructure deployments are not expected to be required for each new model.

Increased frequency of training new models:

NZQA expects to re-train models ahead of each assessment event, which is currently a manually controlled process.

The manual approach is not in itself a concern at this stage; it is sensible to run manually until the scope and requirements of an automated solution are fully understood.

The process does take time, however, and would ideally be automated before too many additional models are introduced, given the limited capacity of a small data science team.

Part 3: Scalability & Future Extensibility *_continued*

Increased requirement to test and approve new models:

Each new or retrained model must also be tested and approved.

It is reasonable to keep people in the approval loop indefinitely, particularly at this stage.

Over time, however, testing could be automated and trusted to the point where reviewers act more as a release gate than spending significant time on each approval, allowing approvals to occur more frequently without becoming onerous.

Part 3: Scalability & Future Extensibility *_continued*

Potential scaling difficulties:

The platform is well positioned for scaling overall. A few points are worth noting:

- Source data for ATS must be downloaded manually, which limits the ability to run inference automatically.
- The manual training and re-training workflow is adequate for a single model running infrequently, but scaling to more models and more frequent events would benefit from an automated training workflow.
- Automated training, more frequent events and additional subjects will increase demand for model approvals. Manual steps such as testing would then make approval a bottleneck, so the process should be streamlined while retaining human oversight.

Part 3: Scalability & Future Extensibility *_continued*

Future Aspirations:

NZQA advised that plans are already in place to develop new tooling and processes to improve the outlook for scaling, particularly around automating the training and testing of new models.

The existing workflows will need to run more frequently as new models and assessment events are introduced. Any remaining manual processes will become increasingly difficult to maintain and a potential source of error.

The inference workflow is of particular interest: it could be wrapped in a custom front-end, or built as an extension of the model registry and approval interface. This would reduce the need for engineers to access the AWS console directly, and allow orchestration of features such as a list of approved model versions, automatic inference-run-ID generation, and end-to-end logging, audit and observability. This depends on having programmatic access to source data.

Findings

Consolidated Findings

ID	Findings	Impact Statement
IS-1	There is no automated code-quality scanning or quality gate on pull requests	Code quality depends on manual review alone, increasing the risk that defects or inconsistencies reach the main branch and accumulate as technical debt
IS-2	Incidents are identified and escalated through NZQA's standard service desk (FirstCall) and the Assessment Division's incident-management process, but ATS-specific support arrangements, including who responds to platform alerts and how handover occurs, are not yet defined.	Platform alerts may not reach a named responder, so responsibility for acting on them is diffuse and issues may go unactioned, particularly between assessment events.
IS-3	Learner responses cannot be retrieved programmatically from the external provider	The download must be performed manually for each run, which introduces the potential for human error and prevents fully automated inference
IS-4	Once downloaded, learner responses require manual cleaning before they can be used	This adds delay and potential inconsistency, and modifying source data creates a risk of altering it in ways that blur responsibility between the provider and NZQA
IS-5	There is no facility to run an inference job using a previous version of a model	A historic inference run cannot be reproduced once that model version has been superseded

Consolidated Findings _continued

ID	Findings	Impact Statement
IS-6	Model approval depends on manual testing and oversight for each new or retrained model	The effort required grows with each additional model and retraining cycle, which will not scale efficiently as the number of models increases
IS-7	Model training and retraining are manually coordinated	The coordination effort grows with each additional model and event, placing increasing demand on a small data science team as the platform scales
IS-8	Operational documentation is incomplete, including runbooks, incident-response procedures, common failure modes and fault-diagnosis procedures	On-call engineers may lack the guidance needed to support the platform effectively, particularly during incidents
IS-9	Inference runs are manually orchestrated, including starting and configuring the step functions	This adds manual effort, increases the potential for error, and requires direct AWS console access to production accounts
IS-10	The naming scheme for inference runs contains only a date and a sequential number	Run history will become difficult to track as the number of models and the frequency of runs increase

Appendix

Appendix 01: Session and Sources Register

Source ID	Date	Description
S-1	2 July 2026	ClearPoint / NZQA ATS discovery workshop
S-2	9 July 2026	ClearPoint - NZQA AWS/ATS Operational documentation review
S-3	9 July 2026	ClearPoint - AWS/ATS Infrastructure discussion
S-4	13 July 2026	ClearPoint - NZQA ATS model codebase discussion
S-5	15 July 2026	ClearPoint/NZQA AWS/ML Ops scalability discussion
S-6	22 July 2026	ATS Model registry for ClearPoint review
S-7	10-24 August 2026	ClearPoint / NZQA follow-up correspondence on draft report findings

Appendix 01: Session and Sources Register *_continued*

Ref	Category	Document	Description
D-1	DSAT (Data Science team remediation documentation)	DSAT-Guardrails-240626-013300	Guardrail design: off-topic and safety prompts, categories and evaluation
D-2	DSAT	DSAT-Fallback logic of LLM guardrails-240626-013728	Fallback / retry logic for guardrail LLM calls
D-3	DSAT	DSAT-ATS MLOps Governance and Operations-240626-013858	Model governance and operations: versioning, CAB change management, approver personas, monitoring, rollback, incident response
D-4	DSAT	DSAT-Architectural Decision-240626-012800	Architectural decision record: DeBERTa-v3-large model choice rationale
D-5	DSAT	DSAT-Automated Text Scoring (Literacy Writing): Model framework overview-240626-012440	Overview of the ATS model framework
D-6	DSAT	DSAT-Data set for model training-240626-012955	Training-data specification (dataset Datasheet)
D-7	DSAT	DSAT-Essay text cleansing-240626-012855	Source-data cleansing process (SRH extraction; preserves Te Reo Maori macrons)
D-8	DSAT	DSAT-Model card for NZQA AI marking model-240626-013058	Model card for the ATS marking model
D-9	DSAT	DSAT-Quality assurance framework for ATS-240626-014245	QA framework: script selection for human quality assurance

Appendix 01: Session and Sources Register *_continued*

Ref	Category	Document	Description
D-10	DSAT	DSAT-Re-training ATS model: augmented data selection and results-240626-014337	Model re-training approach, augmented data selection and results
D-11	DSAT	DSAT-Supervised model provenance and licensing-240626-012626	Model provenance and licensing (DeBERTa MIT; Kaggle notebooks Apache 2.0)
D-12	AWS implementation - Low Level Design	AIML-Low Level Design - Inference Workflow-230626-212206	Inference workflow low-level design
D-13	AWS implementation - Low Level Design	AIML-Low Level Design - Model Approval Workflow-230626-212555	Model approval workflow low-level design
D-14	AWS implementation - Low Level Design	AIML-Low Level Design - Model Ingestion-230626-212717	Model ingestion low-level design
D-15	AWS implementation - Low Level Design	AIML-Low Level Design - Observability-230626-212823	Observability low-level design
D-16	AWS implementation - Functional Testing	AIML-Testing - Model Ingestion-230626-213953	Functional testing: model ingestion
D-17	AWS implementation - Functional Testing	AIML-Testing - Model Inference-230626-220054	Functional testing: model inference

Appendix 01: Session and Sources Register *_continued*

Ref	Category	Document	Description
D-18	AWS implementation - Functional Testing	AIML-Testing - Model Approval-230626-213826	Functional testing: model approval
D-19	AWS implementation - Functional Testing	AIML-Testing - Batch Inference Full Load-230626-223605	Functional testing: batch inference full load
D-20	AWS implementation - Other	AIML-Model Card-230626-224322	AWS model card
D-21	AWS implementation - Other	AIML-End User Guide - Model Approval Workflow-230626-223757	End-user guide: model approval workflow
D-22	AWS implementation - Other	AIML-End User Guide - Batch Inference Workflow-230626-224138	End-user guide: batch inference workflow
D-23	AWS implementation - Other	AIML-NZQA CodePipeline - Deploy CDK Code-230626-230602	CodePipeline / CDK deployment reference
D-24	AWS implementation - Other	AIML-Pipeline File Format Reference-230626-213602	Pipeline file-format reference
D-25	AWS implementation - Other	AWS CodeCommit monorepo	Inference scripts, Lambda functions, CDK (infrastructure-as-code), and supporting documentation and scripts

Appendix 01: Session and Sources Register *_continued*

Ref	Category	Document	Description
D-26	Governance, security & requirements	Automated Text Scoring Requirements_Updated_202601	Requirements specification (updated Jan 2026): NZISM, privacy, security, production acceptance criteria
D-27	Governance, security & requirements	AE1-2026 Data flow with automations	Business process / data flow with automations
D-28	Governance, security & requirements	BSG-NZQ-13777-Bastion-NZQA ATS Security Risk Assessment Report v0.4	Independent security risk assessment (Bastion)
D-29	Governance, security & requirements	Copy of Privacy Threshold Analysis NCE110 ATS Implementation	Privacy Threshold Analysis (PTA)
D-30	Governance, security & requirements	CAB Approval Change Request for deployment - CR-986	Change Approval Board sign-off for production deployment
D-31	Governance, security & requirements	ClearPoint NZQA ATS Review 202510 Recommendation Reconciliation	Reconciliation of the 2025 review recommendations
D-32	NZQA / NZCER reports	NZQA ATS HLSO	High-Level Solution Overview
D-33	NZQA / NZCER reports	NZQA ATS model Review Report	NZCER statistical and fairness analysis of the ATS model
D-34	AWS implementation Low Level Design	Low Level Design - Inference Pipeline (Approval Workflow via API Gateway)	Production inference pipeline LLD (Bedrock batch guardrails, SageMaker scoring, Step Functions). Provided during follow-up (LR-002 clarification)

Appendix 01: Session and Sources Register *_continued*

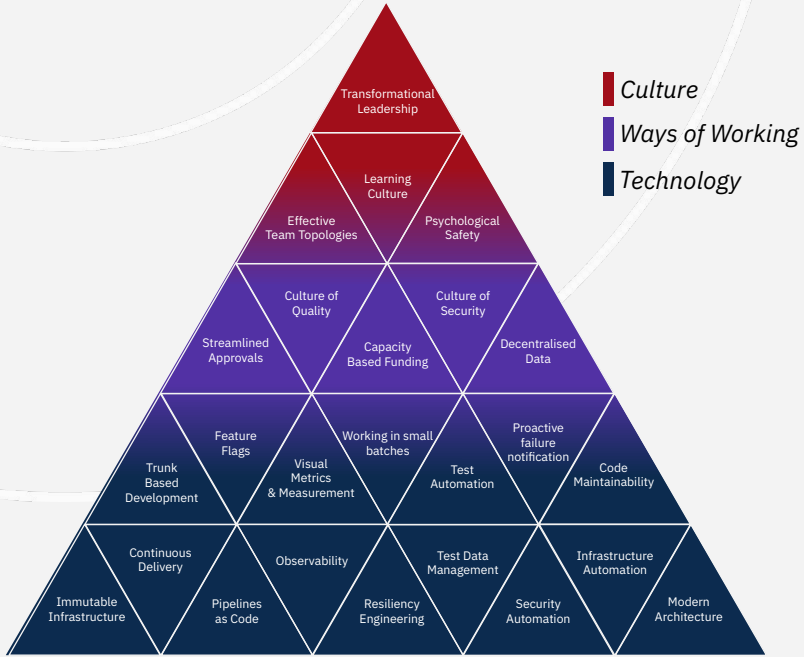
Ref	Category	Document	Description
D-35	NZQA / NZCER reports	AIML-Fairness analysis of ATS model output	In-house fairness analysis of AE1 2026 model output, covering the supervised model and the LLM approach. Provided during follow-up (LR-009 clarification)

Appendix 02: ClearPoint's F25 capabilities

ClearPoint have identified 25 areas which can be used to assess a team or organisation's foundational delivery and development capabilities.

They are spread across 3 wider themes, which can point to general areas where attention and guidance may be required.

ClearPoint have referenced some of the lower-down technology sections in this review to highlight some of the areas considered when appraising NZQA's platform.



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.

clearpoint.

Thank you.