PLATFORM GUIDE

Programme Insights Platform Overview

How AI-Powered Document Assessment Works

Executive Summary

Programme Insights is a document assessment platform built specifically for UK infrastructure programmes and the consultancies that support them. It reads programme documents against recognised criteria frameworks and produces traceable, evidence-cited assessments with board-ready outputs.

The platform was built by programme assurance practitioners who have sat in IPA gateway reviews, managed nuclear safety submissions, and evaluated major project business cases. Every module is aligned to the exact criteria that reviewers, regulators, and evaluation panels will use.

The Transformation

Programme teams and consultancies currently rely on weeks of manual document review that produces inconsistent findings dependent on the individual reviewer. Programme Insights replaces this with hours of automated, evidence-cited assessment where every finding is traceable to a specific page and paragraph in the source documents.

Senior experts currently reading documents can focus instead on strategic decisions, stakeholder management, and the judgement calls that require their experience. The platform handles the analytical workload.

6	218+	10+6	UK
ASSESSMENT MODULES	CRITERIA MAPPED	STEP AGENTIC PIPELINE	DATA RESIDENCY

Built for the Sector

Every module is aligned to recognised standards by programme assurance experts. This is not generic AI adapted for compliance — it is a purpose-built, multi-step agentic assessment pipeline with deterministic guardrails, engineered around the exact criteria your reviewers will assess you against.

The Platform Approach

Programme Insights is a single configurable engine with domain-specific modules. Every module shares the same core assessment infrastructure — what changes between modules is the criteria framework and the domain expertise encoded into it.

What You Get With Every Module

Regardless of which module you deploy, the platform provides the same core capabilities. This is the foundation that makes assessments repeatable, traceable, and defensible.

Criteria Decomposition

Complex criteria broken into binary sub-questions that can be objectively assessed. "Does the business case demonstrate value for money?" becomes 8–12 specific, testable questions.

Evidence Citation

Every finding links to the source document, page number, and relevant text. A reviewer can verify every claim independently — no "AI said it's fine."

Deterministic Scoring

Scores are based on evidence found, not probabilistic AI opinions. The same documents and criteria produce the same assessment every time.

Coverage Mapping

Clear visibility of which criteria have evidence, which are partially addressed, and which have gaps. No ambiguity about where you stand.

Dual RAG Ratings

Each criterion receives two independent ratings: **coverage** (how well sub-questions are addressed) and **delivery confidence** (quality and depth of evidence). Aggregated views by review area for executive-level reporting.

Board-Ready Reports

PDF reports with executive summaries, criterion-by-criterion findings, evidence citations, and prioritised recommendations. Suitable for SRO briefings and programme boards.

Incremental Assessment

Carry-forward logic copies valid results from previous runs. Only criteria affected by document changes or those that previously failed are re-assessed — making subsequent runs significantly faster.

Deterministic Guardrails

Rules-based overrides ensure consistency regardless of LLM variability. Coverage percentages calculated from sub-question answers override model conservatism with transparent, auditable logic.

Real-Time Progress

Server-Sent Events stream assessment progress to the UI in real time. Users see exactly which criterion is being assessed as it happens — no black box, no waiting without visibility.

Evidence Lineage

Every assessment finding connects to the specific document chunks used to produce it. Full traceability from finding to source text, with relevance scores and retrieval metadata.

Standard & Extended Tiers

Choose assessment depth: Standard tier (retrieval + scoring) for fast assessments, or Extended tier (adds corrective RAG, citation verification, self-critique, and adversarial debate) for deeper assurance.

SharePoint Sync

Connect to document libraries for continuous monitoring. As documents are updated, assessments refresh automatically — readiness tracked over time, not just at a point in time.

Configurable Criteria

Load your own assessment frameworks alongside standard modules. Proprietary methodologies, client-specific criteria, or bespoke evaluation matrices — configured without code changes.

Platform + Module Architecture

The platform provides the orchestration pipeline, hybrid evidence search, agentic per-criterion assessment, deterministic guardrails, and report generation. Modules provide the domain-specific criteria frameworks and analyst personas. Adding a new assessment domain means configuring a new module — not building new infrastructure.

The Assessment Pipeline

When a user clicks "Run Assessment", the platform validates the request, resolves the framework, selects the engine tier (Standard or Extended), and runs the assessment as a background task. Progress updates stream to the UI via Server-Sent Events in real time — users see exactly which criterion is being assessed as it happens.

Orchestration Pipeline (10 Steps)

1

Duplicate Run Prevention

Checks if an assessment is already running (10-minute window). Prevents accidental duplicate runs that would waste compute and create confusing parallel results.

2

Stale Run Resumption

If a previous run stalled (e.g., server restart mid-assessment), the pipeline reclaims it rather than starting fresh. Crash recovery without user intervention — no lost progress.

3

Create Assessment Record

Inserts a new assessment run with an atomically-calculated version number. Records the engine tier and a snapshot of the pipeline configuration for full audit traceability.

4

Initialise Progress Tracker

Creates the real-time SSE progress stream so users see exactly which criterion is being assessed (e.g., "Assessing criterion 12 of 62: Client-Side Budget").

5

Load Criteria

The CriteriaLoader routes to the correct source: module pack criteria (DCO, CDM, Nuclear), user-defined criteria (Assure Anything), or framework criteria (Gate Ready, GreenBook). Evidence requirements are bulk-fetched, document scope resolved, and search configuration built — semantic weight 0.7, keyword weight 0.3, similarity threshold 0.35.

6 Per-Criterion Assessment Loop

The core of the pipeline. For each criterion (with timeouts of 300s standard / 480s extended), the system first checks for **carry-forward**: if this is not a forced re-assessment, valid results from the previous run are copied forward. Only criteria affected by document changes or those that previously failed/timed out are re-assessed. This makes subsequent runs significantly faster.

Criteria requiring assessment are then run through the agentic pipeline (Standard or Extended tier — detailed below).

7 Generate Summary

After all criteria are assessed: aggregates token usage, runs the Overall DCA Assessment (groups criteria by dimension, calculates per-dimension RAG, checks for showstoppers, combines rules-based preliminary rating with GPT-4o professional judgement), and generates the Executive Summary — verdict headline, key strengths, critical issues, and recommendation.

8 Finalise

Marks the run as completed, saves the project summary, resets project status, and queues a background consistency check. The assessment is now available for review, export, and board-ready reporting.

> **Incremental Intelligence**
>
> The carry-forward system means the first assessment of a document set runs every criterion from scratch, but subsequent assessments only re-assess what has changed. Updated a single document? Only the criteria that drew evidence from that document are re-run. This is not just faster — it provides a clear audit trail of what changed between assessment versions.

The Per-Criterion Agentic Pipeline

Each criterion is assessed by a multi-step agentic pipeline. The Standard tier is included with every assessment. The Extended tier adds four additional steps for deeper assurance on complex or high-stakes assessments.

Standard Tier — Included With Every Assessment

1 Retrieval

Hybrid document search combining semantic similarity (pgvector) and keyword matching. The criterion is expanded into 2–3 query variants, each run independently, then results are fused using Reciprocal Rank Fusion (RRF). Returns up to 15 ranked evidence chunks with source metadata and relevance scores.

Why it matters: Query expansion catches evidence described differently from the criterion language. RRF fusion ensures the best evidence rises regardless of which query found it.

2 Scoring

The core assessment intelligence. Formats retrieved chunks as numbered references, builds the assessment prompt using the framework's analyst persona, calibration guidance, criterion definition, and evidence requirements. Calls GPT-4o with structured output at temperature 0.0 for deterministic scoring.

Produces: **coverage rating**, **delivery confidence rating**, finding narrative, evidence citations, recommendation, key evidence quotes, confidence score, and per-requirement assessment.

Deterministic Guardrails — Rules Override LLM Conservatism

After the LLM scores each criterion, deterministic rules override the model where evidence contradicts its assessment. This is the mechanism that makes assessments **consistent and auditable** rather than probabilistic:

Coverage Override: Coverage percentage is calculated directly from sub-question answers (Yes=100%, Partial=70%, No=0%). Maps to GREEN ≥70%, AMBER 40–69%, RED <40%. This overrides the LLM's coverage rating with a transparent, deterministic calculation.

Guardrail A: If coverage is GREEN and negatives ≤20% but LLM said AMBER delivery → force GREEN.

Guardrail B: If coverage is GREEN but LLM said RED delivery → force AMBER.

Guardrail C: If coverage is RED but LLM said GREEN delivery → force AMBER.

Guardrail D: If retrieval quality was "poor" and LLM said GREEN delivery → force AMBER.

Extended Tier — Deeper Assurance (4 Additional Steps)

3 Corrective RAG (CRAG)

Classifies retrieval quality as good, ambiguous, or poor using GPT-4.1-mini. If ambiguous or poor, decomposes the criterion into sub-queries and re-retrieves in parallel. Merges results via RRF (ambiguous) or full replacement (poor). Up to 2 retrieval rounds ensure evidence gaps are from genuine absence, not search failure.

Why it matters: If the first search missed relevant evidence, CRAG finds it. Assessments are only as good as the evidence they consider.

4 Citation Verification

Checks each evidence quote against actual source chunks to verify it is verbatim, not paraphrased or hallucinated. This catches the fabricated citations that plague other AI tools — when the report says "Benefits_Realisation_Strategy.docx, p.14," that page contains the referenced text.

Why it matters: No hallucinated citations. Every claim a reviewer checks will trace back to real evidence in real documents.

5 Self-Critique

If assessment confidence is below 70%, runs a 3-question critique: unsupported claims? rating consistency? generic recommendations? Can adjust ratings based on critique findings. Skipped for high-confidence assessments — compute is spent only where uncertainty exists.

Why it matters: Low-confidence assessments get an automatic second opinion before being presented as findings.

6 Adversarial Debate

Triggers only for AMBER-rated criteria (boundary cases). Runs a multi-persona adversarial debate using GPT-4o — IPA modules get 3 specialist personas, other modules get 2 general-purpose ones. The debate transcript is stored as additional signal, stress-testing the rating from multiple angles.

Why it matters: Boundary ratings get scrutinised by multiple perspectives. If a rating should be GREEN or RED, the debate surfaces the argument.

Model Architecture

Three model tiers are used throughout the pipeline, each selected for its specific role:

MODEL	ROLE	USED FOR
GPT-4o	Heavyweight assessment	Criterion scoring, overall summary, adversarial debate, DCA assessment
GPT-4.1-mini	Targeted analysis	CRAG classification, self-critique, citation verification
text-embedding-3-small	Semantic search	Document chunk vectors for hybrid retrieval (pgvector)

> **Reproducibility With Guardrails**
>
> The deterministic guardrails mean that even with inherent LLM variability, the same documents assessed against the same criteria produce consistent findings. The rules-based coverage calculation and override logic ensure that small variations in model output do not change the assessment outcome. This is a structured pipeline with deterministic controls — not a generative AI producing different opinions on each run.

Module Catalog

Six assessment modules covering the major frameworks used in UK infrastructure programme assurance. Each module uses the same agentic assessment pipeline with domain-specific criteria frameworks and analyst personas.

IPA Gateway Review

FRAMEWORK	Infrastructure and Projects Authority Gate Review criteria
CRITERIA COUNT	218 mapped from IPA workbook
TYPICAL USE	Pre-gateway readiness assessment — identify gaps and evidence weaknesses before the formal IPA review
WHO USES IT	Programme directors, PMO teams, and consultancies supporting gateway preparation on major UK infrastructure programmes
ASSESSMENT SCOPE	Business case documents, programme management plans, benefits realisation strategies, risk registers, resource plans, stakeholder engagement strategies, and commercial/procurement documentation

Example Finding

> `E.04 (Benefits Management): Evidence found in Benefits_Realisation_Strategy.docx (p.14–16) demonstrates benefit mapping to strategic objectives. However, no quantified baseline or measurement methodology identified. Benefits register references KPIs but does not define data collection approach. — AMBER`

Output

RAG-rated assessment against all 218 criteria with evidence citations, gap analysis grouped by review area (Strategic, Economic, Commercial, Financial, Management), prioritised improvement recommendations, and coverage heatmap showing evidence strength across the criteria framework.

HMT Green Book Five Case Model

FRAMEWORK	HM Treasury Green Book (2022), Five Case Model
CRITERIA COUNT	98 assessment points across 5 cases
TYPICAL USE	Business case quality assessment before Treasury approval — Strategic Outline Case, Outline Business Case, or Full Business Case submissions
WHO USES IT	Programme directors, economics teams, and consultancies preparing OBCs and FBCs for Treasury gateway
ASSESSMENT SCOPE	Strategic case, economic case, commercial case, financial case, and management case documents — including supporting evidence such as economic appraisals, procurement strategies, and benefits maps

Example Finding

> `SC-12 (Strategic Fit): Strong alignment demonstrated with departmental objectives (Strategic_Case_v4.docx, p.8). Cross-reference to National Infrastructure Strategy confirmed (p.23). Policy context is current and accurately cited. — GREEN with high confidence`

Output

Case-by-case RAG assessment with evidence citations, cross-case consistency check (identifying contradictions between cases), evidence quality scoring, and prioritised recommendations for strengthening weaker cases before submission.

Tender & ITT Evaluation

FRAMEWORK	Client-defined evaluation criteria (configurable per tender)
CRITERIA COUNT	Variable — typically 40–120 per tender
TYPICAL USE	Consistent, traceable tender evaluation across multiple submissions — ensuring every response is assessed against the same criteria with the same rigour
WHO USES IT	Procurement teams, bid evaluation panels, and consultancies managing tender processes for infrastructure and public sector contracts
ASSESSMENT SCOPE	Tender response documents, method statements, CVs, programme proposals, pricing schedules, and supplementary evidence across all submissions

Example Finding

> `Q3.2 (Methodology): Response describes phased approach but lacks detail on risk mitigation during mobilisation period. Evidence: Method_Statement.docx p.7–9. Competing submission (Bidder C) provides more specific risk treatment with named mitigation actions. — AMBER`

Output

Per-criterion scoring with evidence citations for each bidder, cross-bidder comparison matrix, score distribution analysis, and moderation-ready assessment pack with audit trail for procurement governance.

Nuclear Safety (ONR)

FRAMEWORK	ONR Licence Conditions, Safety Assessment Principles (SAPs)
CRITERIA COUNT	36 licence conditions + SAP assessment points
TYPICAL USE	Nuclear site licence compliance assessment, periodic safety review support, and pre-submission readiness checks for ONR regulatory submissions
WHO USES IT	Nuclear licensees, nuclear safety teams, and consultancies supporting ONR regulatory submissions on new build and operational sites
ASSESSMENT SCOPE	Safety cases, periodic safety reviews, environmental impact assessments, emergency preparedness documents, ALARP demonstrations, and supporting technical evidence

Example Finding

> `LC22 (Modification or Experiment): Modification proposal references ALARP assessment (Safety_Case_Amendment_v2.docx, p.12) but does not demonstrate systematic hazard identification per SAP EKP.3. Hazard identification methodology is referenced but not evidenced in the submission pack. — RED — requires remediation before submission`

Output

Licence condition compliance matrix with evidence status, SAP alignment assessment, regulatory submission readiness rating, and prioritised remediation actions with severity classification.

CDM Compliance

FRAMEWORK	Construction (Design and Management) Regulations 2015
CRITERIA COUNT	48 assessment points across CDM duty holder requirements
TYPICAL USE	CDM compliance audit, pre-construction information assessment, and health & safety file review against regulatory requirements
WHO USES IT	CDM coordinators, principal designers, principal contractors, and clients with duties under CDM 2015
ASSESSMENT SCOPE	Pre-construction information packs, construction phase plans, health and safety files, designer risk assessments, and duty holder appointment documentation

Example Finding

> `Reg 4 (Client Duties): Pre-construction information pack identified (PCI_Pack_v3.pdf). Contains hazard register and existing site conditions survey. Missing: asbestos survey reference required by Reg 4(4) — document referenced in contents page but not included in submission. — AMBER`

Output

Duty holder compliance matrix showing evidence status per regulation, gap analysis by CDM duty holder role, priority remediation actions, and document completeness assessment for the pre-construction information pack.

NEC Contract Assessment

FRAMEWORK	NEC4 Engineering and Construction Contract
CRITERIA COUNT	62 assessment points across NEC clauses and procedures
TYPICAL USE	NEC compliance audit, compensation event assessment, programme review, and early warning process evaluation
WHO USES IT	Project managers, commercial teams, and NEC-accredited consultancies managing or auditing NEC contracts
ASSESSMENT SCOPE	Contracts, compensation event notices and quotations, programme submissions, risk registers, early warning registers, and project manager communications

Example Finding

> `Clause 61.3 (Compensation Event Notification): CE-047 notification issued within 8 weeks of awareness (CE_Register.xlsx, row 47; Notification_CE047.pdf, p.3). However, quotation does not include Defined Cost forecast per Clause 63.1 — lump sum stated without cost breakdown. — AMBER — quotation incomplete`

Output

Clause-by-clause compliance assessment with evidence trail, compensation event procedural audit (timeliness, completeness, notification compliance), programme assessment findings, and early warning register effectiveness analysis.

Custom Modules

The configurable criteria engine means new modules can be created for any assessment framework — proprietary consultancy methodologies, client-specific evaluation criteria, or emerging regulatory standards. Module configuration does not require code changes: define the criteria, map the decomposition rules, and the platform handles the rest.

Security & Data Handling

Programme documents contain sensitive information — commercial data, safety cases, procurement evaluations, and strategic programme decisions. The platform is designed for this reality from the ground up.

AREA	DETAIL
Hosting Location	Azure UK South — all data processed and stored within the United Kingdom. Data never leaves UK jurisdiction.
Regulatory Registration	ICO registered data controller. Compliant with UK GDPR and Data Protection Act 2018.
Certification	Cyber Essentials certified. Security controls independently assessed.
Document Processing	Documents processed in isolated compute environments. No data shared between client tenants. No document content used for model training or improvement.
Access Control	Role-based access: Administrator, Reviewer, Viewer. Granular permissions per assessment and module. SSO integration available (SAML 2.0).
Audit Trail	Every assessment action logged with timestamp, user, and action detail. Complete provenance from document upload through to finding generation. Exportable for governance review.
Data Retention	Configurable retention policies per client. Clients control the data lifecycle including deletion. Retention periods can align with organisational records management policy.
Encryption	Data encrypted at rest (AES-256) and in transit (TLS 1.2+). Encryption keys managed per tenant.
Model Training	Client documents and assessment data are never used for AI model training. Explicit contractual commitment. No data leakage between tenants.
Penetration Testing	Regular third-party penetration testing. Findings remediated on published schedule. Results available under NDA.

Built for UK Government Data

The platform is designed to handle OFFICIAL and OFFICIAL-SENSITIVE classified documents in compliance with UK Government Security Classifications. Azure UK South hosting, ICO registration, and Cyber Essentials certification provide the baseline. Additional controls are available for specific classification requirements.

Integration & Deployment

The platform connects to existing document management infrastructure and authentication systems. No custom development required from the client.

Document Connectivity

SharePoint Sync

Primary document connection method. Connect to SharePoint Online document libraries for continuous monitoring. Documents are automatically ingested when added or updated. Assessment findings refresh as documents change — providing continuous readiness visibility rather than point-in-time snapshots.

File Upload

Direct upload of PDF, Word (.docx), and Excel (.xlsx) files. Suitable for ad-hoc assessments or environments where SharePoint access is not available. Drag-and-drop interface with batch upload support for large document sets.

Enterprise Integration

API Access

RESTful API for programmatic integration with existing programme management tools, dashboards, and reporting systems. Trigger assessments, retrieve findings, and export reports without using the web interface.

SSO / SAML

SAML 2.0 integration for enterprise single sign-on. Connect to Azure AD, Okta, or other SAML-compatible identity providers. Users authenticate with existing corporate credentials — no separate platform accounts to manage.

Deployment Model

ASPECT	DETAIL
Delivery	SaaS only — hosted on Azure UK South. No on-premises deployment.
Availability	99.9% uptime SLA. Scheduled maintenance windows communicated in advance.
Scalability	Multi-tenant architecture scales with document volume. No client-side infrastructure requirements.
Browser Support	Chrome, Edge, Firefox, Safari (current and previous major version).
Pilot Timeline	Typically 2–4 weeks from agreement to first assessment using real documents and criteria.

Getting Started

A structured pilot process that gets real output from real documents within weeks — not months of procurement. Designed so your team can evaluate the platform with minimal commitment.

The Pilot Process

1. **Initial Demo** (30 minutes)
   See the platform running with your domain — IPA gateway, Green Book, tender evaluation, or whichever module is relevant. Not a slide deck. Not a marketing video. The actual platform assessing real framework criteria.

2. **Pilot Configuration**
   We configure the platform for your specific context: your criteria framework (standard or custom), your document types, your reporting requirements. If you have a proprietary methodology, we map it into the criteria engine.

3. **Assessment Run**
   Real output from real documents. Upload programme documentation or connect a SharePoint library, and the platform produces a full assessment — evidence citations, RAG ratings, gap analysis, and a board-ready report.

4. **Validate Findings**
   Your subject matter experts review the assessment output. Check the evidence citations against source documents. Verify the gap identification matches their own understanding. This is where trust is built — or not.

5. **Decision**
   Based on validated pilot results, decide whether to deploy across engagements or programmes. Scale at the pace that suits your organisation. No lock-in beyond the subscription period.

Pricing

Per-module subscription with pricing based on assessment volume and number of concurrent users. Pilot pricing available to reduce evaluation risk. Contact us for a tailored proposal based on your deployment scope.

Request a Demo

See Programme Insights running against the criteria framework that matters to your organisation.

Web	programmeinsights.com
Email	hello@programmeinsights.com
LinkedIn	Programme Insights

Programme Insights

Objective, evidence-based assessment you can trust — so your expertise goes where it matters most.