

AQR KPI2.1 AI

Documentation & User Guide

Release date: September 2026 | Application version: 1.0 | Rubric version 2.1.0

Platform: Universal HTML5 / JavaScript — client-side · Google Gemini-powered · single-file deployment

Central principle

Assessment quality review should be consistent, transparent and evidence-based. AQR KPI2.1 AI applies a structured, documented check to every assessment element and records the evidence behind each score. It supports professional academic judgement; it does not replace the reviewer or institutional authority.

1. Introduction

AQR KPI2.1 AI is a browser-based quality assurance tool built for the College of Pharmacy at Gulf Medical University. It supports the internal assessment review carried out by the Assessment Committee against **CAA OBEF KPI 2.1**, producing a scored, evidence-referenced dossier for each course under review.

The application evaluates every assessment element of a course against nine rubric criteria grouped into four weighted domains, and computes a composite score and a KPI percentage. Each criterion score is accompanied by the evidence that produced it, what was found, what was missing, and — where relevant — the exact measurement behind the judgement.

Two engines are available. The **local rule engine** applies deterministic, auditable rules entirely in the browser with no network calls. The **AI engine** additionally sends de-identified evidence to Google Gemini for criteria that require reading and interpretation. Criterion 2c is always deterministic and cannot be overridden by the model.

Key features

- **Nine-criterion OBEF rubric** across four weighted domains — alignment with learning outcomes, assessment design, grading clarity, and feedback — producing a composite of 1.00–5.00 and a KPI of 20–100%.
- **Genuine current-versus-previous comparison** for item reuse: questions are extracted from both papers and matched question by question, with whole-document comparison as a second basis for assessments that are not question papers.
- **Automatic model selection** — the Gemini model catalogue is ranked live and the newest stable model is chosen, with automatic step-down through up to eight candidates when one is unavailable.
- **Browser-side de-identification** — student names, identifiers, emails and phone numbers are stripped from grade files at upload, before any analysis and before any network call.
- **PDF export** that is an exact copy of the on-screen report, stamped PROVISIONAL until the review is signed off.
- **Runs anywhere** — one self-contained HTML file, no backend, no installation; works standalone or embedded in Wix or Cloudflare Pages.
-

Professional review required

The scores and defects produced by this application are structured decision support, not a validated psychometric instrument. Every criterion score and every reported defect must be reviewed and confirmed by a qualified academic before it informs any decision or is submitted to the Assessment Office.

2. System Requirements

Browser and device

- Modern desktop browser: Chrome 110+, Edge 110+, Firefox 115+, Safari 16.4+
- Tablet and mobile: Chrome for Android, Safari for iOS — tablet or larger recommended for the evidence and report stages
- JavaScript must be enabled; minimum usable screen width 320 px
- An internet connection is required to load the CDN libraries, and for the AI engine. The local rule engine needs no network access once the page has loaded.

Google Gemini API key

Required only for the AI engine

The local rule engine runs with no key and no network calls. To use the AI engine you need a free Gemini API key. The application never ships with a key, and never stores one on a server.

Get your API Key: i. Go to Google AI Studio (aistudio.google.com/app/apikey) ii. Log in to your PERSONAL account, and click Create API Key. iii. Copy and paste the key below.

The key is sent only to generativelanguage.googleapis.com, always in the `x-goog-api-key` request header and never in a URL. A Content-Security-Policy in the page restricts outbound connections to that endpoint and the two script CDNs.

Supported file formats

Every attachment slot accepts all of the formats below. Maximum 25 MB per file, and several files may be attached to a single slot.

Format	Extensions	Notes
Excel spreadsheet	<code>.xlsx</code> <code>.xls</code> <code>.xlsm</code>	Best format for blueprints and grade reports. Table structure is preserved, so header-aware column selection works.
CSV / plain text	<code>.csv</code> <code>.txt</code> <code>.md</code> <code>.rtf</code>	Read directly. Good for exported question banks.
PDF	<code>.pdf</code>	Text-based PDFs only. A scanned image yields no text and item-level checks cannot run.
Word document	<code>.docx</code>	Read via Mammoth. Legacy <code>.doc</code> is not supported — save as <code>.docx</code> first.
Images	<code>.png</code> <code>.jpg</code> <code>.jpeg</code> <code>.webp</code>	Screenshots of Moodle pages. Routed to the multimodal model; readable only under the AI engine.

Scanned documents

A scanned exam paper contains no machine-readable text. Item segmentation, defect detection and similarity analysis all depend on text, so a scan will produce an "insufficient evidence" result rather than a score. Export from Examsoft or Moodle as DOCX or a text-based PDF.

3. Getting Started — the Four-Stage Workflow

The application opens fresh every time. If a previous session was saved on the device, a banner offers to restore it; nothing is restored unless you choose to.

Stage 1 • AI

Choose the engine. The default is AI auto-select, which resolves the best available Gemini model automatically. Paste your API key if you are using the AI engine, or continue with the local rule engine for a fully offline, deterministic review.

Stage 2 • Course & QFEmirates Level

Identify the course and its context, then list the assessment elements **as per the course syllabus**. The fields are:

Field	Purpose
Target QFEmirates Level	Level 6 (Bachelor's), 7 (Master's) or 8 (Doctoral). Drives criterion 1c, which judges cognitive demand against the level descriptors.
Course year	Year 1–6. Scoring allows an introductory assessment in year 1 and expects progressively greater demand in later years.
Course name and code	Identify the review. The code appears in the PDF filename.
Previous offerings	Whether the course has run before. Selecting first offering hides the previous-exam slot and awards criterion 2c the maximum score, because recycling is not possible.
Assessment elements	One row per element in the syllabus, with its name. Each element is reviewed and scored separately.

Stage 3 • Evidence Upload

Nine evidence slots, each with an information icon describing exactly what is required and how it is used.

Mandatory slots are listed below; the standard-setting sheet and moderation sheet are optional, and leaving them empty needs no justification.

Evidence	Required	Used for
Blueprint of the exam	Yes	CLO/PLO coverage matrix (1a, 1b), assessment method (2a), mark totals, grading criteria (2b, 3a)
Course syllabus	Yes	Learning outcomes, topic and skill coverage (1b), course nature (2a), published grading structure (3a)
Exam of the current year	Yes	Item segmentation, all 31 defect rules, cognitive demand (1c), similarity baseline (2c)
Exam of the previous year	If repeated	Item reuse and recycling (2c). Appears only when the course has previous offerings.
FOLIO Grades report	Yes	Grade distribution and psychometrics (3b). De-identified in the browser at upload.
Standard setting sheet	Optional	Named method and cut score strengthen 2b and 3b
Moderation sheet	Optional	Second-marker or double-marking evidence strengthens 2b
Moodle evidence — blueprint published	Yes	Proof the blueprint and criteria were published before the assessment (3a)
Moodle evidence — feedback given	Yes	Timeliness, cohort coverage and specificity of feedback (criterion 4)

Evidence is routed, not pooled

Each criterion is scored from its own named evidence only. Evidence is never imported from one criterion to another, so a strong blueprint cannot inflate the feedback score. Missing evidence produces an explicit "insufficient evidence" result with a stated floor, rather than a silent low score.

Stage 4 · Analysis & Report

The pipeline runs in stages, with progress shown: text extraction from every attachment, item segmentation, similarity measurement against the previous paper, defect detection, grade and psychometric analysis, then criterion scoring. Each criterion is then presented with its score, the reasoning, what was found and what was missing.

You can record your own verdict on each criterion — accept, adjust or reject — and the report keeps both the computed score and your decision. Sign-off and PDF export are at the foot of the page.

4. The Rubric and How the Score Is Computed

The rubric is versioned (**v2.1.0**) and follows the CAA OBEF KPI 2.1 structure: four domains, nine criteria, each criterion scored 1–5 by the reviewer or the engine.

Domain	Title	Weight	Criteria
1	Alignment with Learning Outcomes	45%	1a, 1b, 1c
2	Assessment Design	25%	2a, 2b, 2c
3	Grading Clarity	20%	3a, 3b
4	Feedback	10%	4

The nine criteria

Code	Criterion	Weight (of 5)
1a	Does the assessment measure the intended learning outcomes?	1.5
1b	Are specific skills or competencies provided by the curriculum being evaluated?	1.5
1c	Does the assessment meet the relevant level of QFEmirates?	2.0
2a	Is the assessment method (e.g., exam, project) appropriate for its intended purpose?	2.0
2b	Does the course use methods to evaluate assessment accuracy and consistency?	1.5
2c	Was item reuse and assessment recycling avoided in major assessments within the past three years?	1.5
3a	Are the grading criteria clearly defined, published, and aligned with the assessment tasks?	3.0
3b	Is the grade distribution fair, unbiased, and not inflated?	2.0
4	Do students receive timely, actionable, and constructive feedback to support their learning?	5.0

The calculation

Criterion weights sum to exactly 5.0 within every domain, and domain weights sum to 1.00. The composite is therefore always on the same 1–5 scale as the individual criteria.

1. **Criterion contribution** = criterion score $\times$ criterion weight $\div$ 5
2. **Domain score** = sum of its criteria's contributions, giving a value from 1.00 to 5.00
3. **Composite** = sum of (domain score $\times$ domain weight), again 1.00 to 5.00
4. **KPI percentage** = composite $\times$ 20, giving 20% to 100%

So a review scoring 1 on every criterion yields 20%, all 3s yield 60%, and all 5s yield 100%. The application asserts these three mappings, and the two weight sums, on every page load; a failure is reported to the browser console rather than silently mis-scoring.

Band descriptors

Score	Meaning
1	Major misalignment / poor
2	Partial alignment / weak
3	Acceptable / generally appropriate
4	Strong alignment / well designed
5	Excellent / fully validated

The reported KPI is labelled a **minimum** score: it reflects the evidence supplied. Additional evidence can only raise it.

5. Item Reuse and Similarity Analysis (Criterion 2c)

Criterion 2c asks whether item reuse and assessment recycling were avoided. It is measured, not inferred: the AI engine cannot change this score. When the course is a first offering the criterion is awarded 5 automatically and the reason is recorded.

Two bases of comparison

The application chooses the basis the documents allow, and reports which it used.

- **Question by question** — where both papers can be read as question papers (at least three questions each). Every current question is matched against every previous question and the best match reported.
- **Whole document** — always computed in addition, and the sole basis for assignment briefs, projects and other assessments that are not laid out as question papers.

The score is taken from whichever basis reads higher, so reuse cannot hide behind a document format.

How two questions are compared

Question text is first normalised: the shared scaffolding of question writing is stripped ("which of the following", "most appropriate" and similar), units are standardised, and drug synonyms and brand/generic names are mapped to a common form. Several measures are then computed for each pair:

Measure	What it detects
Whole question	The stem and its answer options taken together — the strictest test
Wording	Shared phrasing between the two stems
Phrasing	Overlapping word sequences, which survives light editing
Answer options	Reuse of the option set, independently of the stem
Numbers changed	A question reused with only its figures altered

Measure	What it detects
Containment	Whether one question is substantially contained in the other — this is what catches rewording

Classification

For a multiple-choice question, the answer options are what make it that question. A matching stem with entirely different options is a shared house template, not reuse; a rewritten stem carrying the old options forward is reuse. Each pair is classified accordingly:

Classification	Severity	Meaning
Identical	Critical	The whole question, stem and options, matches word for word
Cosmetic edit	Critical	Trivially edited, or only the numbers changed
Stem reused, options changed	Major	The stem carried over with new options
Options reused, stem changed	Major	The option set carried over under a rewritten stem
Paraphrase	Major	The same question reworded
Same concept, new item	Minor	The same knowledge point, genuinely rewritten
Shared stem template, new question	None	Only the department's house stem wording is shared
Unrelated	None	No meaningful overlap

A pair the classifier judges harmless is never flagged, whatever the reviewer sets the sensitivity threshold to. Every matched pair is listed with its similarity and classification for you to confirm or reject.

Scoring bands

The higher the measured similarity, the lower the score:

Measured similarity	Score	Interpretation
More than 90%	1	The assessment is last year's, with at most minor edits
60% – 90%	2	Substantial repetition
30% – 60%	3	Partial overlap
10% – 30%	4	Limited overlap
Less than 10%	5	Recycling avoided

Interpreting whole-document similarity

Two documents from the same course always share the course template — the title block, weighting, submission route, late-penalty and academic-integrity paragraphs, and criteria headings. On assignment briefs this alone typically measures 40–50% even when the task is entirely new. Where whole-document comparison is the only basis available, read the reported figure alongside the listed question pairings before accepting the band.

6. Grade Distribution and Psychometrics (Criteria 3a and 3b)

Finding the right marks

The grades report usually contains several mark columns. The application matches the assessment element name against the column headers rather than assuming a position, and reads a maximum from the header — a column headed "End Semester (50)" is scaled from 50, not treated as a percentage. This prevents the common error of reporting raw marks as percentages.

Distribution statistics

Computed for every element: number of candidates, mean, median, standard deviation, quartiles, minimum and maximum, pass rate, and ceiling and floor effects. Distribution flags are raised for patterns that indicate a problem, for example a mean at or above 88% with a standard deviation of 8 or less — a compressed high distribution consistent with grade inflation or an assessment that does not discriminate — or 25% or more of candidates scoring above 95%.

Letter-grade distribution

Criterion 3b is scored on the **letter-grade distribution only**, not on individual marks. Grades are ordered A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F, and the concentration in the top bands is assessed against the cohort size.

Item analysis

Where the grades export also contains the per-item response matrix — which option each candidate chose for each item — the application additionally computes:

Statistic	Interpretation
Item difficulty (p-value)	Proportion answering correctly. Very high or very low values indicate an item that does not discriminate.
Point-biserial discrimination	Correlation between getting the item right and total score. Low or negative values indicate a defective item.
Distractor efficiency	How many distractors attracted a meaningful share of responses. Non-functional distractors reduce the effective number of options.
KR-20 reliability	Internal consistency of the whole paper.
Standard error of measurement	Derived from KR-20 and the total-score standard deviation.

Strongest available evidence

A grades export containing the response matrix produces the most defensible criterion 3b score, because reliability and item discrimination are measured rather than inferred. Where only totals are available, the score rests on the distribution alone and the report says so.

7. The Report and PDF Export

On-screen report

Stage 4 presents a domain compliance matrix, the minimum overall KPI, and a card for each assessment element containing every criterion with its score, reasoning, findings and gaps. Strengths and recommendations are summarised per element.

Recording your decisions

Each criterion carries a reviewer verdict — accept, adjust or reject — recorded with your name and a timestamp. Both the computed score and your verdict are retained, so the dossier shows what the engine measured and what the reviewer concluded.

Sign-off and export

The **Reviewer Sign-off** card records the reviewer's name and role. **Save** stores the review in the browser; it does not download a file.

Export report as PDF is always available. The exported PDF is an exact copy of the on-screen report: it is captured from the rendered report itself, sliced to A4 pages, with a running footer carrying the application name, course code, rubric version, sign-off state and page numbers.

State	Export behaviour
Not signed off	Exports with a diagonal PROVISIONAL watermark on every page and _PROVISIONAL appended to the filename
Signed off	Exports clean, with the sign-off name and timestamp in the footer

Filename pattern: `AQR_KPI2.1_<CourseCode>_QF<Level>_<YYYY-MM-DD>.pdf`

Export is PDF only

Word and Excel export are deliberately not offered. The dossier is an evidence record, and PDF preserves the report exactly as reviewed and signed. If an editable version is required, produce it from the source evidence rather than from the dossier.

Download inside an iframe

When the application is embedded in a Wix widget or an iframe, the browser may open the PDF in a new tab instead of downloading it. Use the browser's Save As option from that tab.

8. Save, Restore and Reset

The application opens from the beginning every time. Nothing from a previous session is loaded unless you explicitly restore it.

- **Save progress** stores a snapshot in the browser. It does not download anything.
- **Restore** is offered by a banner at launch when a saved snapshot exists on the device. Declining leaves the session empty.
- **Save** on the sidebar downloads a JSON file, for moving a review between devices or archiving it outside the browser.
- **Reset all data** clears the session, the stored snapshot and all attachments. It asks for confirmation in an in-app dialog, so it works correctly inside sandboxed embeds where browser dialogs are blocked.

Storage uses `localStorage` for the session snapshot and IndexedDB for attachment text, each with an in-memory fallback. Where a browser blocks storage — private windows, or a sandboxed iframe — the application runs normally for the session and simply cannot offer a restore afterwards.

9. Technical Specifications and Deployment

Technology stack

Component	Technology
AI engine	Google Gemini API (v1beta) — live model catalogue, ranked automatically, newest stable model chosen, step-down through up to 8 candidates
PDF reading	PDF.js 3.11.174 (Mozilla), with the worker configured for sandboxed embedding
Word reading	Mammoth 1.6.0
Spreadsheet reading	SheetJS 0.18.5
PDF export	html2canvas 1.4.1 + jsPDF 2.5.1, entirely client-side
Interface	Bootstrap 5.3.2, Inter and JetBrains Mono, and an inline SVG icon set — no icon font
Deployment unit	A single self-contained HTML file (~430 KB). No installation, no build step, no backend.
Storage	Browser <code>localStorage</code> and IndexedDB only — no server-side storage

All logic is contained in one classic script with no ES modules, so the file works when pasted into an HTML widget that does not serve it from its own origin.

Cloudflare Pages deployment

- Log in to dash.cloudflare.com, then **Workers & Pages → Create → Pages → Upload assets**.
- Rename the application file to `index.html` and upload it.
- The tool is immediately available at the generated `*.pages.dev` URL.

Wix HTML widget embedding

- In the Wix Editor choose **Add → Embed Code → Embed HTML**.
- Select **Code** rather than a website address, then paste the entire contents of the HTML file.
- Set the component height to at least 900 px, width 100%, with scrolling enabled.

Built for embedding

Browser dialogs, clipboard access and downloads are all restricted inside a sandboxed iframe. The application uses in-app dialogs throughout, degrades gracefully when storage is unavailable, and keeps every icon inline, so no feature depends on the page being served from its own domain.

Responsive behaviour

- Evidence cards and statistics reflow to a single column on narrow screens
- Wide tables scroll inside their own container — the page body never scrolls sideways
- The report is laid out at a fixed width off-screen so that the PDF export is identical on every device

10. Privacy, Security and Data Handling

De-identification at upload

Grade reports are de-identified in the browser the moment they are read, before any analysis and before any network request. Columns whose headers indicate personal data — student name, email, student ID, username, registration number, mobile, phone, national ID, passport, Emirates ID — are redacted cell by cell with a mask. The header row and the number of cells are preserved, so the table keeps its shape and the marks column is still found correctly.

What leaves the browser

- Under the **local rule engine**, nothing. No network request is made after the page loads.
- Under the **AI engine**, de-identified evidence text is sent to the Gemini API for the criteria that require interpretation. Raw grade data is never sent — only computed, de-identified statistics.
- The API key is sent only in the `x-goog-api-key` header to generativelanguage.googleapis.com, and never appears in a URL.
- No data is transmitted to any Gulf Medical University server, to the application author, or to any third party other than the Gemini endpoint.

Prompt-injection defence

All evidence sent to the model is wrapped in explicit data markers and the model is instructed that everything between them is data submitted for review and never an instruction. This matters because evidence files are supplied by third parties and may contain arbitrary text.

Reviewer responsibility

Treat all exam content as confidential. Ensure that your use of a personal Google account and API key complies with institutional data governance and examination security policy. Where policy does not permit exam content to be sent to an external service, use the local rule engine.

11. Development and Attribution

Field	Detail
Application name	AQR KPI2.1 AI
Version	2.1.0 — 4 domains, 9 criteria, composite 1.00–5.00, KPI 20–100%
Institution	College of Pharmacy - Gulf Medical University
Assessment Lead	Dr. Mohammed Abdullatif · dr.abdullatif@gmu.ac.ae
Platform	Universal HTML5 / JavaScript — Client-Side · Gemini AI-Powered
Data storage	Browser localStorage only — no server-side storage