From social media noise to advocacy-ready evidence

How AI-assisted collection, classification and human review can help civic and human-rights organisations produce credible evidence at scale

A ModAstera technical brochure, drawing on the OGBV Tracker, an online gender-based violence (OGBV) monitoring system that we built and operate as technical partner to RACE Centre, a civic organisation based in Abuja, Nigeria.

ModAstera Inc. · November 2026

Organisations that document online harm need a sustainable way to turn scattered posts into advocacy-ready evidence: records they can review, analyse and communicate. In building the OGBV Tracker, we brought collection, AI-assisted classification, human review and public analytics into one shared workflow. Working with RACE Centre showed what matters in practice. Review capacity has to keep pace with collection. Campaign and awareness content must be kept apart from incidents. Automated decisions need to be traceable. This brochure explains what we built, how it supports evidence production, what a team using it should measure, and how we work with organisations that want a similar workflow.

At a glance

The need	Credible briefings, periodic reports and an annual account of online abuse, produced by a small team from a growing volume of online content.
What we built	A shared workflow that collects public posts from five platforms. An AI model suggests a category and summary for each relevant post, and reviewers decide, apart from the automated decisions each deployment allows. Approved evidence feeds filters, charts, maps and shareable chart images.
What changes in the workflow	Records reach reviewers with suggested categories and summaries. That reduces the discovery and transcription needed within the review workflow. Reports and briefings then start from approved records, filters and exported charts.
Who stays in charge	The organisation. It owns the categories, its reviewers apply local context, and it decides how much automation to allow. Approvals and reversals are recorded with their source, so automated decisions can be inspected and undone.
How we report results	Precisely. We separate what was intended, what the code does, what was observed in operation and what has not yet been measured. Appendix A shows how we grade each claim and how we propose to measure the rest.

Who this brochure is for: programme, monitoring and evaluation (M&E) and digital-safety leads, and their technical partners, at organisations that document online gender-based violence, attacks on journalists and human-rights defenders, election-period abuse, hate speech and similar harms.

1. What organisations like our partner need to produce

Our partner is the Centre for Redefining Alternative Civic Engagement for Africa (RACE Centre), a civic organisation based in Abuja, Nigeria, that works on gender justice, technology accountability and research. It already ran an established monitoring practice when we met. Its team watched X, Instagram, Facebook, TikTok and YouTube. They screenshotted likely cases with their replies and engagement, recorded each one in a spreadsheet, and compiled quarterly reports. The partner also published an existing public dashboard. Its goal was evidence for advocacy: material that could inform policy conversations, public communication and a credible annual account of online abuse.

The constraints were familiar to any small team:

- **Manual effort.** Every case was found, captured and typed in by hand.
- **Uneven platform coverage.** Platforms that were harder to watch by hand were under-represented in the records.
- **Inconsistent context.** How much detail each case captured depended on who recorded it.
- **Dependence on individual analysts.** The pipeline depended on the hours a few people could give it.
- **The work of turning records into reports.** Each quarterly report meant compiling and cleaning spreadsheets first.

Our proposal set five goals:

1. broader collection;
2. faster first identification and classification;
3. preservation of local language and cultural context;
4. richer analysis;
5. a stronger public dashboard.

The AI role was a first pass: find likely relevant content, suggest categories and structure records for review. In the original design, a reviewer made every decision.

As volumes grew, we added automated decisions. The tracker can now approve or reject some records without an individual review, using thresholds each deployment sets, and automated approval can be switched off (Section 2.2). The partner's reviewers were trained on these rules, and administrators can adjust them. We describe them openly, because whether to use them is the decision that matters most for evidence quality.

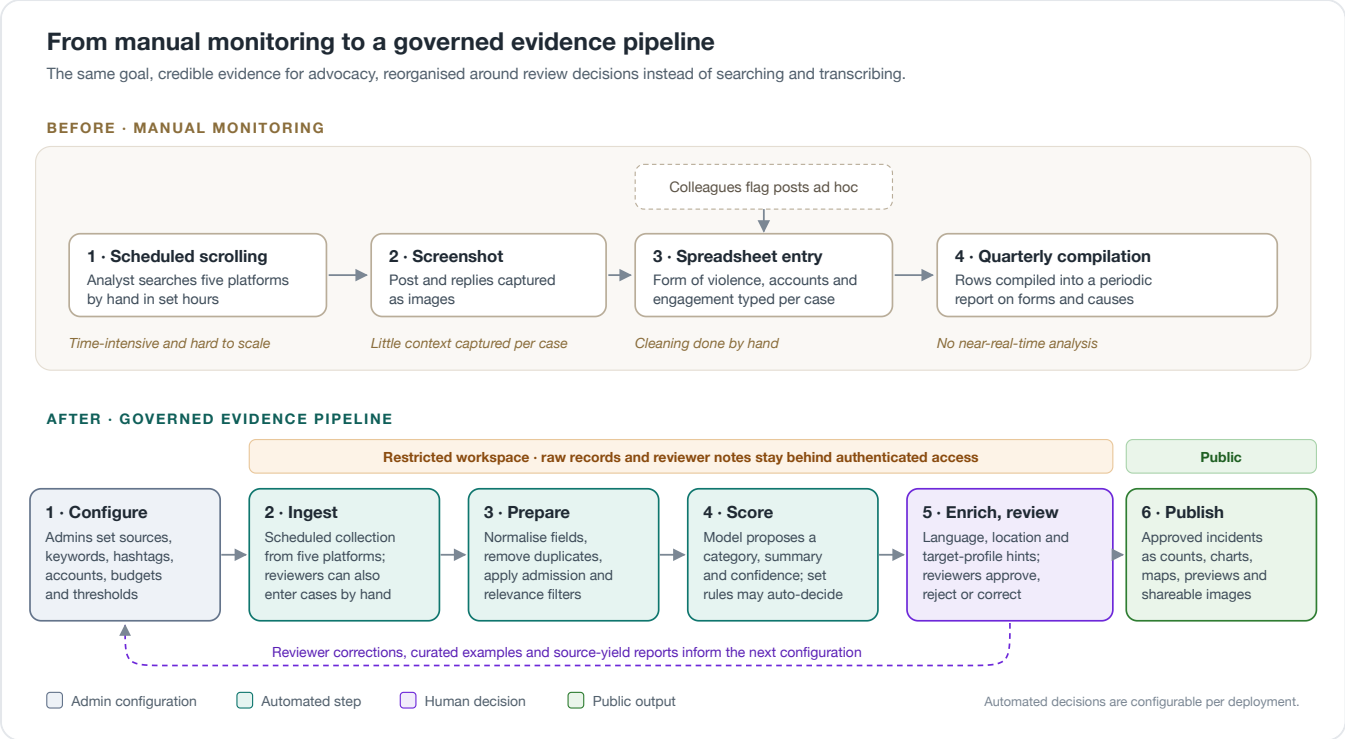

Figure 1. Manual monitoring compared with the shared workflow.

2. How the workflow turns posts into evidence

The tracker works in six stages, shown in Figure 1.

1. **Configure.** Administrators choose the keywords, accounts and pages to follow, and set budgets and thresholds.
2. **Ingest.** The tracker collects public posts on a schedule. Reviewers can also enter cases by hand.
3. **Prepare.** It puts the posts into one common format, removes duplicates and filters out posts that do not pass its admission and relevance checks.
4. **Score.** An AI model suggests a category and a summary for each relevant post, with a confidence score. Set rules may then approve or reject some records automatically (Section 2.2).
5. **Enrich and review.** The tracker adds details, suggesting each record's language, location and target profile. Reviewers check the suggestions and decide the remaining records.
6. **Publish.** Approved records appear as counts, charts, maps and short previews, and charts can be shared as images.

Each of the five goals is matched by something we built: scheduled collection across five platforms, a first classification by the AI model, reviewer control over language and context, filters, charts and measures for analysis, and a public tracker fed by approved records. The table shows what each part does for the team. Its third column, the evidence boundary, marks the limit of what we can claim.

What was built	How it helps the team	Evidence boundary
Scheduled collection from five platforms, using keywords, accounts and pages that administrators choose	Puts all records into one common format and one shared workflow, so the team depends less on searching by hand	We have not shown that every platform is covered fully or equally. The tracker collects only public content that the services supplying its data can reach
Suggested categories and summaries, using the organisation's own categories	Prepares each record for review. A confidence score shows how sure the AI model is, and points reviewers to the uncertain records	Depends on the AI model being switched on and given a budget
Reviewer corrections, and examples chosen to guide the AI model	Gives the organisation's experts a controlled way to guide the suggestions	The AI model does not update itself automatically from these corrections
Separate labels for incidents and for awareness content, such as campaign and support posts	Keeps awareness content out of the counts of abuse	
A record of who or what made each approval, and of how often reviewers changed or rejected the AI model's suggestions	Lets the team check human and automated decisions, and undo them	The record began part-way through the project, and settings differ between deployments
Filters, charts, maps and chart images to share	Helps teams prepare briefings, presentations and public posts from approved data	These tools do not write reports, and they do not show whether advocacy has had an effect
Limits on spending and collection, and a view of how much approved evidence each source produces	Supports decisions about where collection effort and money go	

2.1 Language and local context

One of the five goals was to keep local language and cultural context, so the tracker has to recognise the languages people post in. It works out the language of each post from several clues: the language tag that comes with the post, a language-detection tool, an AI model instructed to recognise Nigerian Pidgin, and common words in Pidgin, Yoruba, Hausa and Igbo.

Reviewers have the final say. They can correct the language of any record, and a sample of the language labels that were accepted automatically is sent back to them for checking. The tracker supports Pidgin, Yoruba, Hausa and Igbo. Appendix A.1 gives the technical detail.

2.2 Automation, and how it is governed

Some records can be decided by the tracker without a person reading them first. This is where evidence quality is most at risk, so the organisation controls how far automation goes.

Each record receives a confidence score between 0 and 1, which shows how sure the AI is of its own suggestion. The organisation sets two thresholds on that score: one for approval and one for rejection. A score at or above the approval threshold counts as high. A score below the rejection threshold counts as low. Anything in between is medium.

Confidence score	What happens to the record
High, country or city known	The record can be approved automatically
High, country or city not known	The record waits for a reviewer
Medium	A reviewer decides. The least certain records are marked for priority review
Low	The tracker rejects the record automatically, and it stays out of the evidence base

These are the standard settings when a deployment starts, with automated approval switched on. Three things keep automation under the organisation's control:

1. **The organisation sets the rules.** Its administrators choose both thresholds. They can also switch automated approval off, so that a reviewer approves every record. Automated rejection is a separate step that follows the rejection threshold.
2. **Automated decisions are labelled and can be undone.** Each automated approval is marked as made by the AI, and each automated rejection is marked on the record (Section 2.3). A reviewer can reverse an approval, or return a rejected record to review.
3. **A second check is available.** The team can add a second, separate AI model to examine each record before it is approved automatically. The check starts as a trial run, called shadow mode, in which it only notes what it would have decided and changes nothing. After the trial, the team can let the check act. A record is then approved automatically only if the check clears it, and a record it does not clear goes to a reviewer. The second check never rejects a record itself.

Appendix A gives the exact scores, schedules and settings. It also describes the test we built to check automated approval against past reviewer decisions.

2.3 Every approval leaves a trace

Each approval or reversal is stored as an event. The event says whether a reviewer, the AI, an import or an unknown source made the decision. For automated approvals it also keeps the model, confidence and threshold. Reversing an approval adds a linked reversal event rather than erasing history. Category corrections are recorded separately and feed the correction measures in Section 4.1. This lets a team answer "how much did automation approve, and how much of it was overturned?" Automated rejections are marked on the record itself rather than logged as events.

2.4 Public and restricted data

Public incident counts, charts, maps and previews use approved records only. Raw post text, handles, post URLs, account identifiers and reviewer notes are not public fields.

Each incident preview shows:

- the date, platform, category and place;
- the target profile, severity and confidence;
- a short summary.

Trend annotations are separate from reviewer notes. Reviewers enter them to mark events on the trend charts, and they are public by default. Figure 5 in Section 6 sets out the boundary.

3. One record, end to end

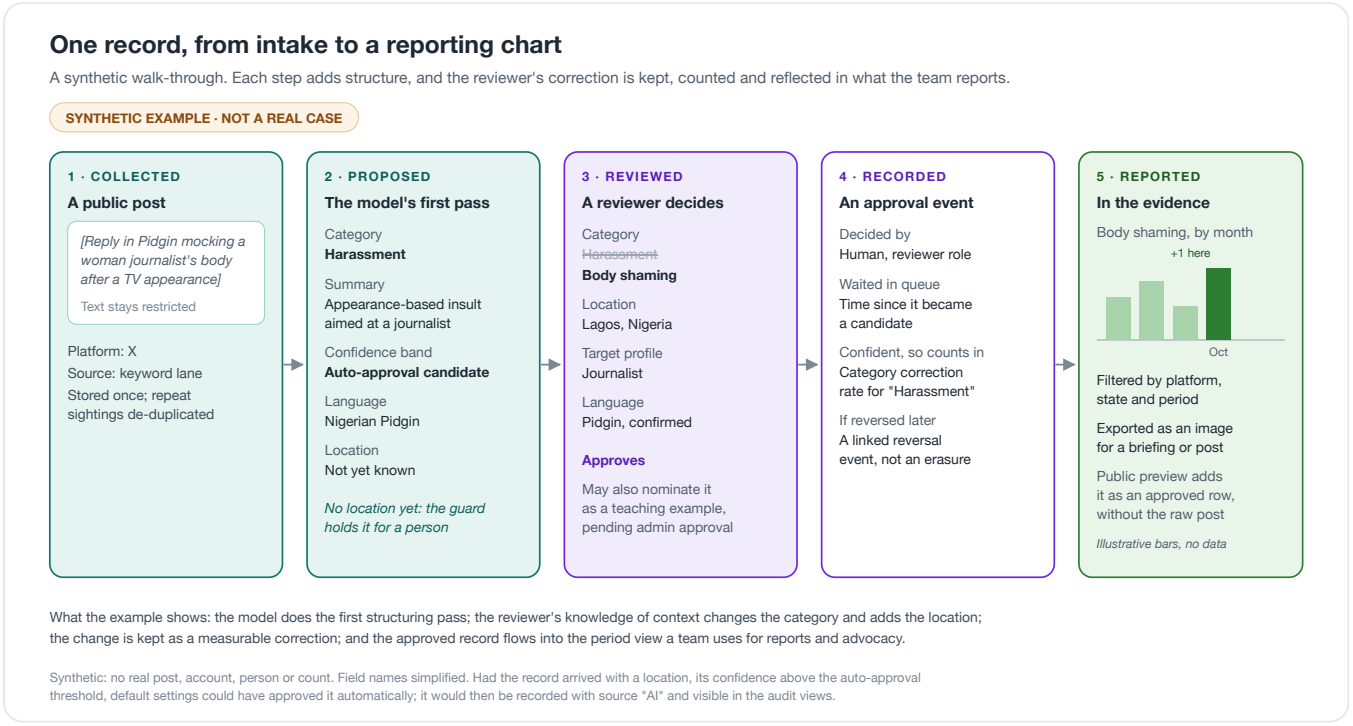

Figure 2. A synthetic example, not a real case. The confidence value used in the text is illustrative.

Figure 2 follows one invented post:

1. **Collected.** A keyword search collects a reply, in Pidgin, that mocks a woman journalist's body after a television appearance.
2. **Proposed.** The model suggests "harassment" with a confidence score of 0.84 and writes a one-line summary. No location is known yet, so the record waits for a reviewer instead of being approved automatically.
3. **Reviewed.** A reviewer, reading the context, changes the category to "body shaming", adds the location and target profile, confirms the language and approves.
4. **Recorded.** The approval is stored with its source. The model's score was above the cut-off the correction measures use (0.80 by default), so the change counts towards the category correction rate for "harassment" (Section 4.1). That tells the team where the model needs clearer examples.
5. **Reported.** The approved record appears in the October count for body shaming. A team member filters by platform, state and month, then exports the chart as an image for a briefing.

The example shows the division of labour. The model structures the record, the reviewer supplies context, the correction is kept and measured, and the result flows into the period view the team reports from.

4. Measures that support decisions

4.1 Start from the decision

A measure is only useful if it helps the team decide something. Each measure below starts from a question a monitoring team asks.

- **What evidence can we report for this period?**

  Look at **approved incidents**, by period, category, platform and place. One approved incident is one post. Several posts can be about the same event or person, so 100 approved incidents do not mean that 100 different people were targeted. The count shows what was documented, not how common the abuse is.

- **Is review keeping pace with collection?**

  Look at the **pending, approved and rejected counts**. Pending records are still waiting for review. The counts show where things stand today. Note them each week to see whether the number of pending records is growing.

- **Which sources are worth keeping?**

  A source is a keyword, account or page that the tracker collects from. Look at **source yield**: of the records from a source that have been approved or rejected, the share that were approved. Records still waiting for review are left out, and automated approvals count as approved.

- **How long do records wait?**

  Look at the **median approval age**: the typical time between a record entering the review queue as a candidate and its approval. Half of the approvals took longer and half took less. It is shown separately for approvals made by reviewers and by automation. The figure includes waiting time, so it does not show how long a reviewer worked on a record.

- **Where does the model disagree with our reviewers?**

  Look at the **high-confidence rejection rate**: of the records scoring 0.80 or more that reviewers decided, the share they rejected. This cut-off is the default for these measures. Appendix A calls it the confidence filter, and it is separate from the thresholds in Section 2.2. A high rate shows where reviewers and the model disagree, so it tells the team where to look more closely. It does not by itself show how often the model is wrong.

- **Which categories need clearer guidance or better examples?**

  Look at the **category correction rate**: of the reviewed records scoring 0.80 or more, the share whose category the reviewer changed. Check how many records each rate is based on, because a rate based on only a few records can change a lot.

- **How much is automation deciding, and how much is overturned?**

  Look at the **AI share**: of the approvals made by the AI or by a reviewer, the share made by the AI. Approvals of imported records (for example, records brought in from an earlier spreadsheet) and approvals from an unknown source are left out. Then look at **reversals**: approvals, whether made by the AI or by a reviewer, that a reviewer later undid. Recent approvals have had less time to be reversed.

The tracker calculates each of these measures. This brochure describes them without reporting live values. Appendix A gives the exact definitions of the rates and times, and Section 4.2 explains the counts.

4.2 Five important numbers

Every report from the tracker draws on five numbers. Each one counts records at a different stage of the workflow, so a report should say which number it uses and for which period.

1. **Collected records.** Every unique post the collectors stored, including posts later judged off-topic. It shows how much public content the tracker looked at.
2. **Candidates.** Collected records that the classifier judged relevant enough for review. It shows what was put in front of reviewers.
3. **Decided records.** Candidates that have been approved or rejected, whether by a reviewer, by an automated rule or through an import. It shows how much has been judged. Candidates not yet decided are still waiting in the review queue.
4. **Approved records.** The evidence base the organisation stands behind. Approved incidents and approved awareness posts are counted separately, and approvals can be split by who or what made the decision.
5. **Public outputs.** Approved incidents as they appear on the public tracker: counts, charts, maps and short previews. It shows what the public can see.

Two relationships help when reading the numbers together. Source yield (Section 4.1) is the approved share of decided records. Approvals can outnumber approved records, because a record that is approved, reversed and approved again counts once as a record and as two approvals.

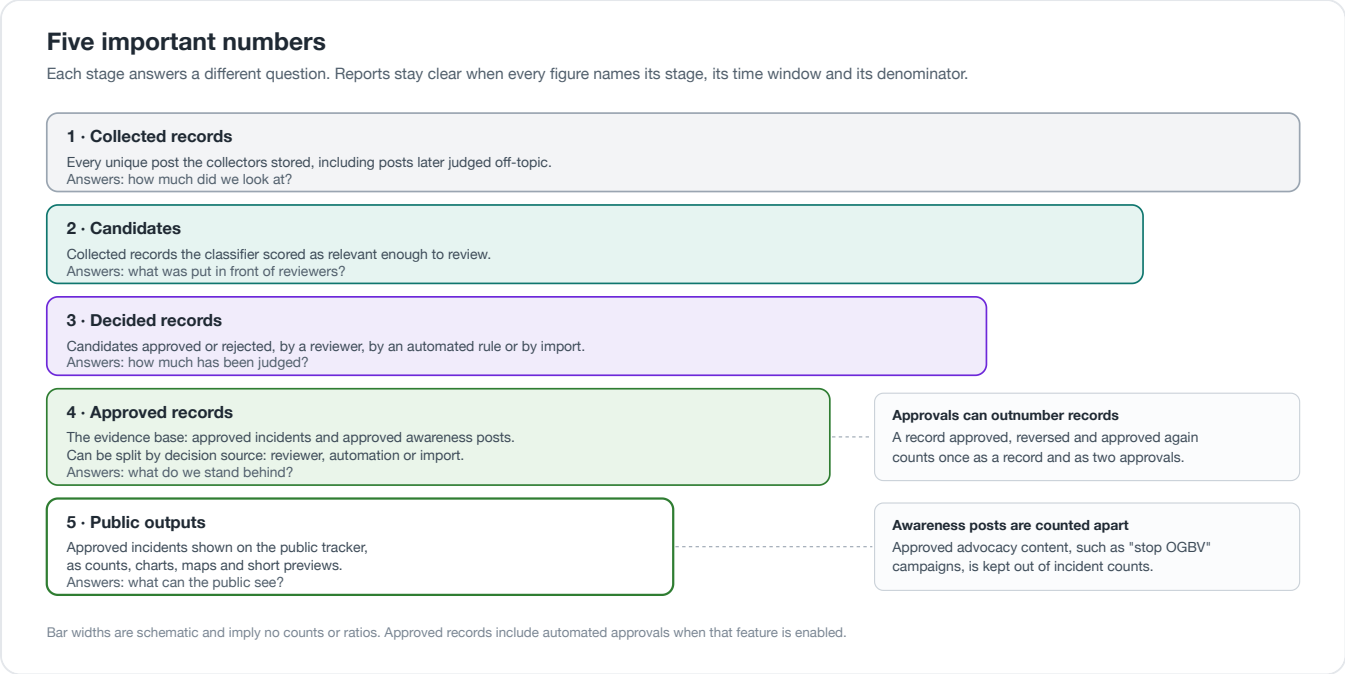

Figure 3. The five numbers and the question each one answers. Bar widths are schematic.

5. Engineering contributions and lessons

These are engineering contributions and lessons developed through this project. Human review, confidence thresholds, structured model output and shadow testing are established practices, and we do not claim to have invented them. What is specific here is how they were combined for civic evidence work:

- **Awareness kept apart from incidents.** Prompted by campaign posts being mistaken for abuse.
- **Controlled prompt examples.** Examples nominated by reviewers stay inactive until an administrator approves them for the model's prompt, the instructions it is given. The fallback classifier described in Appendix A.1, by contrast, learns from all approved records.
- **Decision provenance.** Approvals keep their source and automation context, and reversals are linked records.
- **Context-sensitive processing.** Layered language signals with reviewer correction address local-language content.
- **A second check before automated approval.** An optional second model can run in shadow mode beside reviewers before it is allowed to act, which adds another control before automation is expanded.
- **Source-yield management.** Collection is judged by how much useful reviewed evidence it produces, not by volume.

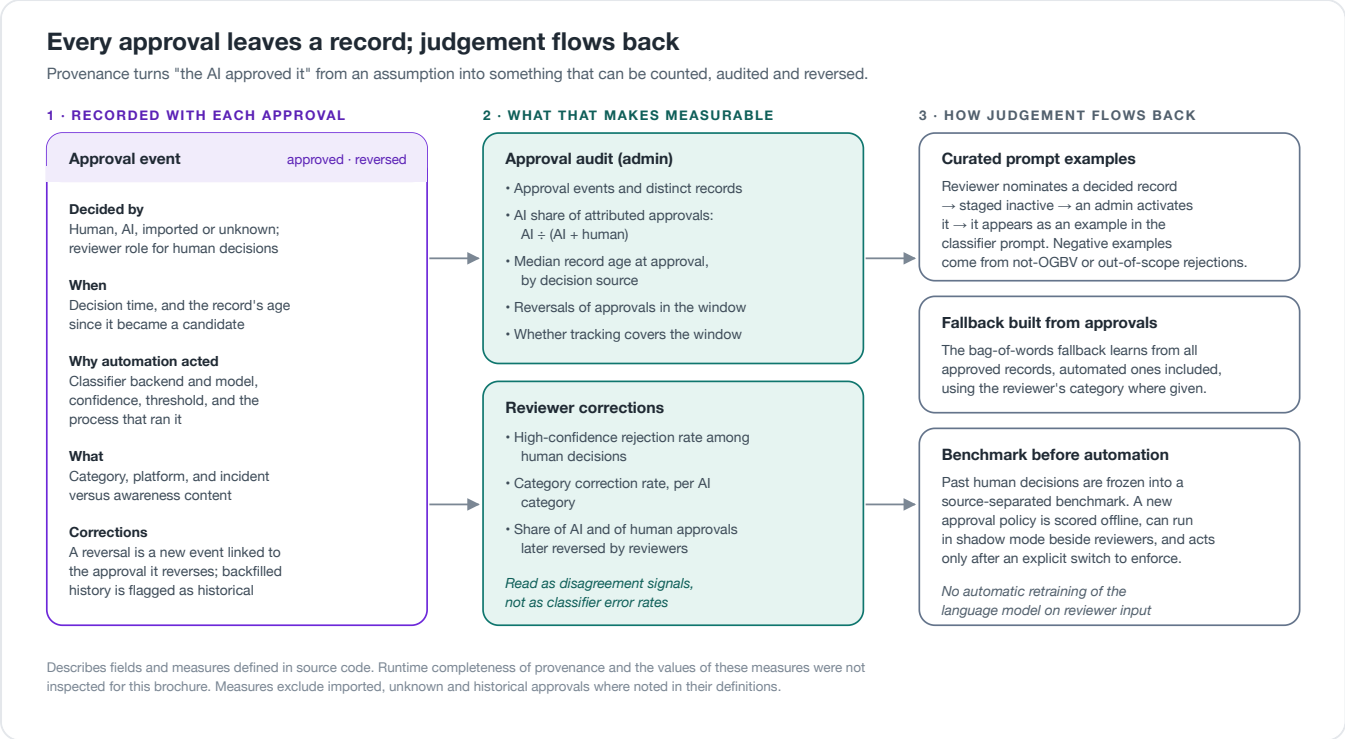

Figure 4. What each approval records, what that makes measurable, and how reviewer judgement flows back. The figure describes the design, not measured values.

6. Working with us: engagement, ownership and data

What you get.

- A public tracker with filters, charts, maps and a methodology page.
- A restricted review workspace with routing, enrichment, corrections, notes and exports.
- Administrator controls for sources, budgets, thresholds, automation and access.
- Audit views for approvals, reversals, corrections and source yield.
- Training for your administrators and reviewers.

What you bring.

- Your categories and decision rules.
- Reviewers with domain and language knowledge, and the time to calibrate.
- Local language partners.
- A safeguarding and ethics position.
- Realistic expectations of platform access.

Questions we settle with you before go-live:

- **Hosting and data ownership.** Your data remains yours. We host and operate the system in a cloud environment agreed with you, and you decide what is collected, reviewed, kept and published. At the end of an engagement we hand over an export of your data and documentation.
- **Retention.** You decide how long raw posts, review records and exports are kept, and we configure the system to match. The database is backed up automatically.

- **Provider use.** Which language-model and data providers process post text, on what terms. Post text is sent to the configured model provider for classification, and location text may be sent to a geocoding service.
- **Reviewer responsibilities and wellbeing.** Who reviews, how decisions are calibrated, and how reviewers are supported when reading abuse every day.
- **Training and support.** We train your administrators and reviewers at launch and provide documentation. Ongoing operations, maintenance and small improvements are covered by a support plan with a regular service review.
- **Recurring cost drivers.** Model-provider usage, collection-provider usage, hosting, and your own reviewer time. Budgets, quotas and a paused-by-default model spend control help keep these predictable.

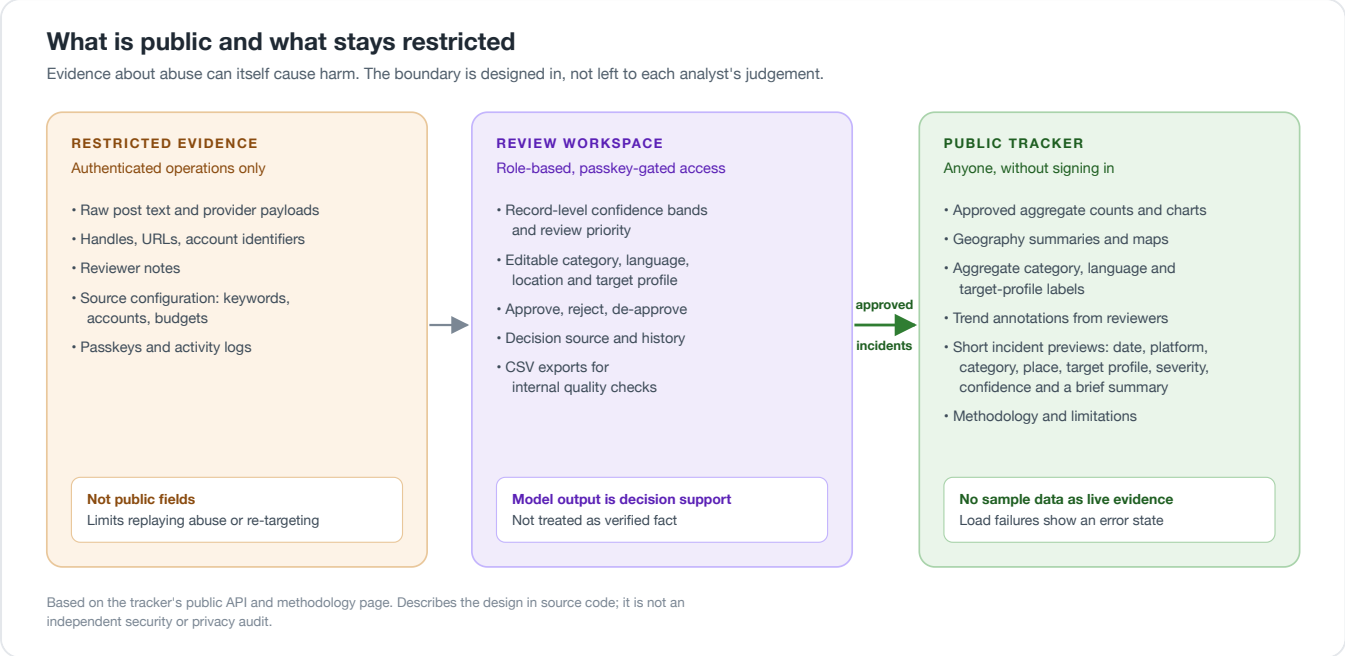

Figure 5. What the public sees and what stays restricted, as implemented in the public API and described from source code. This is not an independent privacy audit.

Talk to us

If your organisation documents online harm and is reaching the limits of manual monitoring, we would be glad to discuss what advocacy-ready evidence could look like for your mission, your languages and your safeguards.

You can explore the public tracker at www.ogbvtracker.africa and read about RACE Centre at www.racecentre.org.

ModAstera Inc.

- Annex Shinsen 901, Shinsencho 10-15, Shibuya, Tokyo 150-0045, Japan
- Email: info@modastera.com
- Web: www.modastera.com

Joshua Owoyemi, PhD, Co-founder and CEO

- Email: joshua.owoyemi@modastera.com
- Phone: +81 90 8929 8658

Appendix A. Technical detail

A.1 Classification and routing

The classifier is LLM-first. It reads each post with the organisation's active categories and administrator-curated examples, and returns strict JSON containing:

- a relevance score;
- a category;
- a confidence score;
- a short summary.

Scores are clamped to 0–1, and categories outside the active taxonomy are discarded. If the model returns no usable category, the tracker tries a bag-of-words classifier, then, if enabled, a rule-based classifier. The bag-of-words classifier learns from approved records, including automated approvals. The rule-based classifier returns fixed confidences.

Paid model calls ship paused and need an administrator-set budget and prices. Records below the relevance threshold (default 0.30) never become candidates. The rest fall into four bands:

- **discard**, below 0.50 by default;
- **priority review**, from the discard threshold up to a fixed 0.60 cutoff;
- **standard review**;
- **auto-approval candidate**, 0.70 or above by default.

Thresholds other than the 0.60 priority cutoff are configurable. Section 2.2 describes these bands in plain words. Its rejection threshold is the discard threshold, and its approval threshold is the auto-approval threshold. The discard band is low confidence, the two review bands are medium, and the auto-approval band is high.

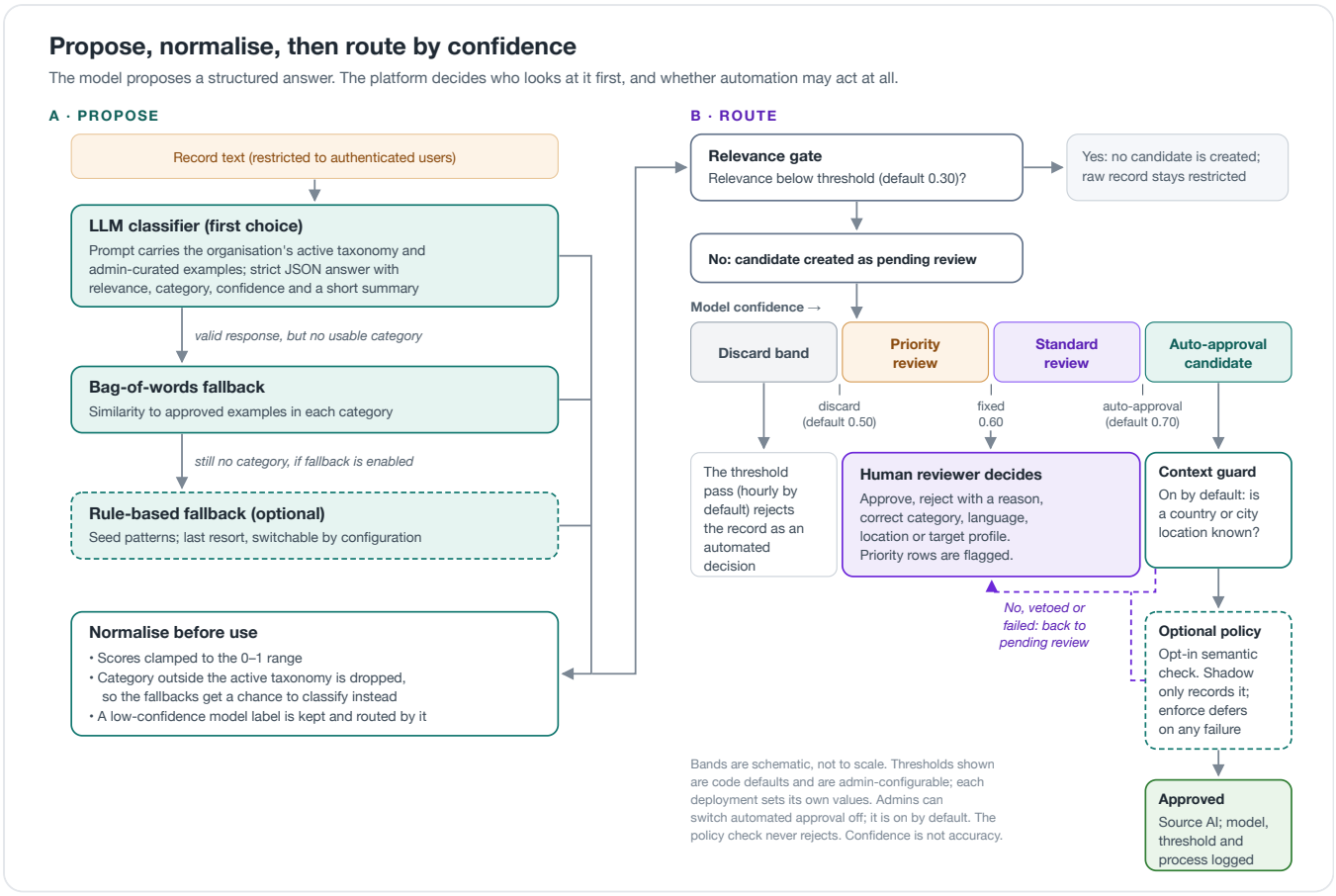

Figure A1. Propose, normalise and route. Thresholds shown are software defaults.

Language identification. Language is identified through a cascade of signals: post metadata, a language-identification model, a language model instructed to recognise Nigerian Pidgin, and marker words for Pidgin, Yoruba, Hausa and Igbo. Reviewers can correct the result, and a fixed sample of automatically accepted labels is sent back for review.

A.2 Automation controls

- **Auto-approval.** Background auto-approval is on by default, and administrators can switch it off so that reviewers make every approval. It applies to records with a confidence of 0.70 or above by default, and it requires a country or city location by default. The confidence usually comes from the language model. When a fallback classifier labels the record, the confidence is that classifier's fixed or heuristic score. An automated approval records source "AI", plus the model, confidence and threshold.
- **Threshold pass.** It runs hourly by default and on demand. It rejects records below the discard threshold (0.50 by default) as automated decisions. These rejections are kept on the record, not as separate events. A reviewer can return a rejected record to review.
- **Optional second check.** This is the second check in Section 2.2, also called the semantic policy. It scores auto-approval candidates with an independent semantic model, and it is off by default. When enabled it starts in shadow mode, where it records what it would have done so that the team can compare it with reviewers. In enforce mode it defers any failed, stale or negative result to a reviewer. It never upgrades a record below the confidence gate and never rejects.

A.3 Metric definitions

Measure	Definition
Source yield	Approved ÷ (approved + rejected) per source in a rolling 7-, 30- or 90-day window; null when nothing is decided
Approvals and distinct records	Approvals recorded in a window, and the distinct records they cover. Reversals are counted separately
AI share	AI ÷ (AI + human) attributed approvals; imported and unknown excluded
Median approval age	Median of approval time minus candidate creation, by decision source
Reversals by source	AI or human approvals in a window later reversed through a linked reviewer reversal. The audit summary also shows an overall reversal share across all approvals
High-confidence rejection rate	Human rejections ÷ human decisions among records at or above the confidence filter (default 0.80) in the window, using each record's latest decision
Category correction rate	Changed category ÷ human-reviewed records with an AI category at or above the confidence filter, grouped by AI category

A.4 Offline benchmark for automated approval

The benchmark freezes past human decisions into a versioned, content-hashed dataset. It has training, development and test splits with no shared source connections. Development data selects prompts and cutoffs, and a policy is registered before its test predictions are produced.

The objective is to recover human approvals while minimising automated approvals of posts that reviewers rejected. The test records were inspected in earlier research, so the benchmark is a historical regression check, not a prospective evaluation.

In retrospective testing, the second check did not match reviewers closely enough to replace them. For that reason it is opt-in and off by default. When enabled it starts in shadow mode, and in enforce mode it passes doubtful records to reviewers.

A.5 Evaluation protocol we propose

Question	Method
How accurate is classification?	A stratified, adjudicated sample across platforms, categories and languages, double-labelled with disagreements resolved
Can automated approvals be trusted?	A regular blind audit sample of auto-approved records
Do reviewers agree?	Inter-reviewer agreement on a shared sample at calibration sessions
Did it save effort?	Active reviewer time, measured separately from queue wait, on fixed cohorts, against a documented manual baseline

Question	Method
Did coverage grow?	Unique-record definitions and fixed source cohorts against a defined baseline, with missing data stated
Is local language handled well?	Language-specific samples reviewed with local language partners

A.6 How we grade evidence

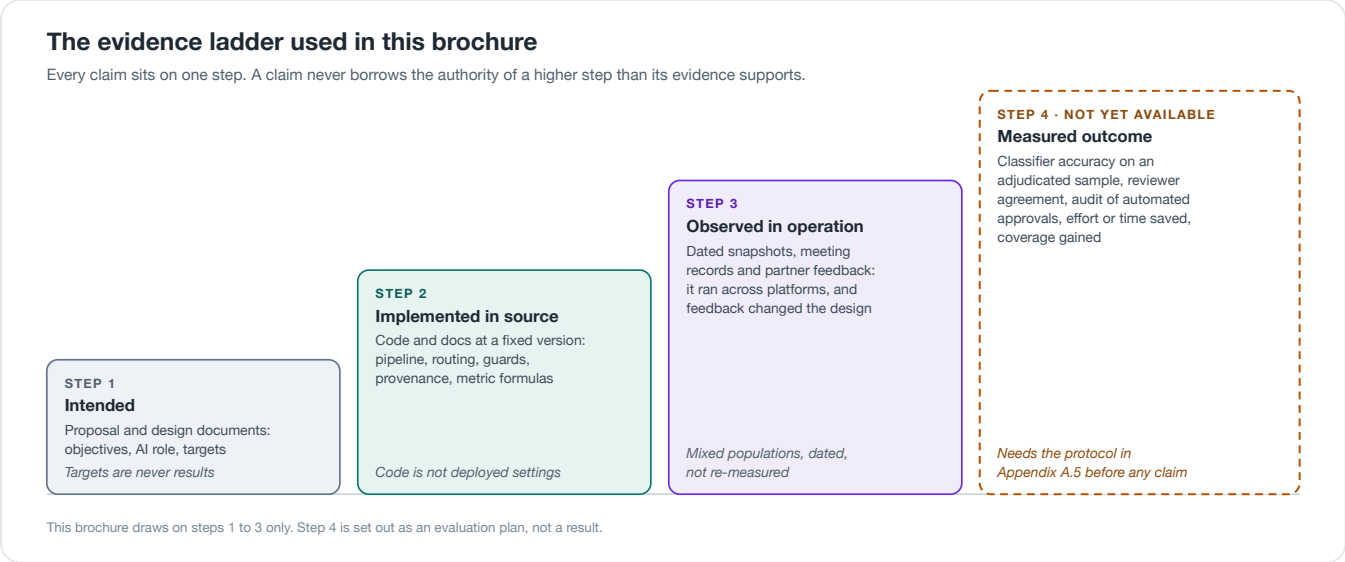

Figure A2. Every claim in this brochure sits on one of the first three steps: intended, implemented in source, or observed in operation. Measured outcomes in live operation are the evaluation plan above.

A.7 Version

This brochure describes the software as it stood on 30 October 2026. Each deployment's settings decide which features are switched on.

Appendix B. Glossary

Term	Meaning
Candidate	A collected record that the classifier judged relevant enough for review
Taxonomy	The organisation's list of abuse categories the classifier may use
Threshold	A score level, set by administrators, at which the tracker changes what it does with a record
Confidence band	A routing label: discard, priority review, standard review or auto-approval candidate
Confidence filter	The cut-off (0.80 by default) used by the high-confidence rejection rate and the category correction rate. It does not change how records are routed
Threshold pass	A scheduled (hourly by default) and on-demand pass that applies thresholds to the review backlog
Shadow mode	Running a new automated policy alongside reviewers, recording what it would have done without acting
Approval event	A stored record of one approval or reversal, with its source, time and automation details
Awareness and advocacy	Campaign, education and support content, counted separately from incidents