

AI-assisted data audits in results-based programmes

Andrew Armstrong

Executive summary

Results-based programmes depend on rigorous verification, yet auditing evidence at scale remains costly and complex. This represents a persistent barrier to national implementation. AI-assisted data auditing offers a practical way forward. This paper examines the cost and practicality trade-offs of applying AI within large-scale results-based funding programmes to analyse diverse, unstructured datasets spanning image files, spreadsheets, PDFs, and handwritten records in multiple languages.

Since 2023, Uptime Global has applied this approach across two functions in audits of rural drinking water services within a multi-country results-based contract portfolio: data extraction and discrepancy confirmation. Using a combination of specialised document intelligence services and general-purpose Vision Language Models (VLMs), Uptime extracts over 200,000 performance records from thousands of pages of source documentation per audit cycle, at a total AI processing cost of under USD 350. Retrieval-Augmented Generation (RAG) offers further efficiency gains, enabling auditors to query source documents directly and receive evidence-grounded responses with traceable citations.

Cost is not a barrier to scale. Across a twentyfold increase in programme size, AI processing costs at current rates rise by less than a few hundred dollars per year and remain well under USD 1,000 dollars in total. Expressed as a share of total recurring programme costs, AI processing represents a fraction of one percent, even for programmes covering hundreds of thousands of sites.

Realising this potential requires clear expectations about what AI can and cannot do. It is one component of a broader audit pipeline, and many efficiency gains also reflect improvements in sample design, data systems, and script-based processing. AI cannot be a substitute for human judgement in interpreting ambiguous findings, nor for the governance structures that uphold sanctions arising from verification results. AI also depends on consistent, credible source data, an area that continues to require sustained investment in many low- and middle-income country contexts. For governments and development banks designing results-based programmes, AI-enhanced auditing is best understood not as an added cost, but as a tool that protects the integrity and credibility of their investments.

The role of AI in verifying results: opportunities and applications

Results-based programmes depend on verifiable evidence. In practice, this evidence is drawn from a wide range of reports, photographs, logs, receipts, and other types of documentation. These records are typically fragmented, multilingual, and unstructured, making large-scale scrutiny demanding. Traditional approaches rely on human review, which introduces interpretive errors and becomes increasingly costly as programmes expand.

The potential for AI to address such data-related challenges in low- and middle-income countries has been recognised for over a decade ([Cohen & Kharas 2018](#), [World Bank 2018](#), [World Bank 2026](#)). More recently, AI has been applied to verification in development contexts in several ways, with varying degrees of success. The most established application has been the use of satellite imagery to detect physical proxies for service delivery. Other approaches remain at earlier stages of maturity. For example, mobile data and call detail records have proven more effective for targeting beneficiaries than for verifying performance changes within relatively homogeneous populations ([Barriga-Cabanillas et al. 2025](#)); natural language processing applied to social media data suffers from geographic bias ([Lythreatis et al. 2026](#)); and the use of IoT sensor data for verification purposes is nascent.

By contrast, AI offers high potential for data audits, which centre on the extraction and cross-referencing of primary evidence. Where documentation exists, even in unstructured or analogue form, AI tools can process heterogeneous inputs such as images, PDFs, and spreadsheets across multiple languages, converting them into structured, comparable data at a scale that human review cannot match. This reduces reliance on standardised data systems and removes practical limits on the volume of documents a single reviewer can assess. In doing so, AI alters the economics of verification and makes national-scale administration of results-based programmes more feasible without compromising integrity.

What do we mean by "AI"?

Artificial intelligence (AI) refers to computer-based systems that support or automate human decision-making by analysing large volumes of information to identify patterns and generate predictions or outputs. Methods range from supervised and unsupervised machine learning to deep learning (neural networks) and generative AI, which produces new content from learned patterns. In the context of data audits, the most relevant applications are natural language processing, which enables the analysis of text, and computer vision, which enables the analysis of images and video.

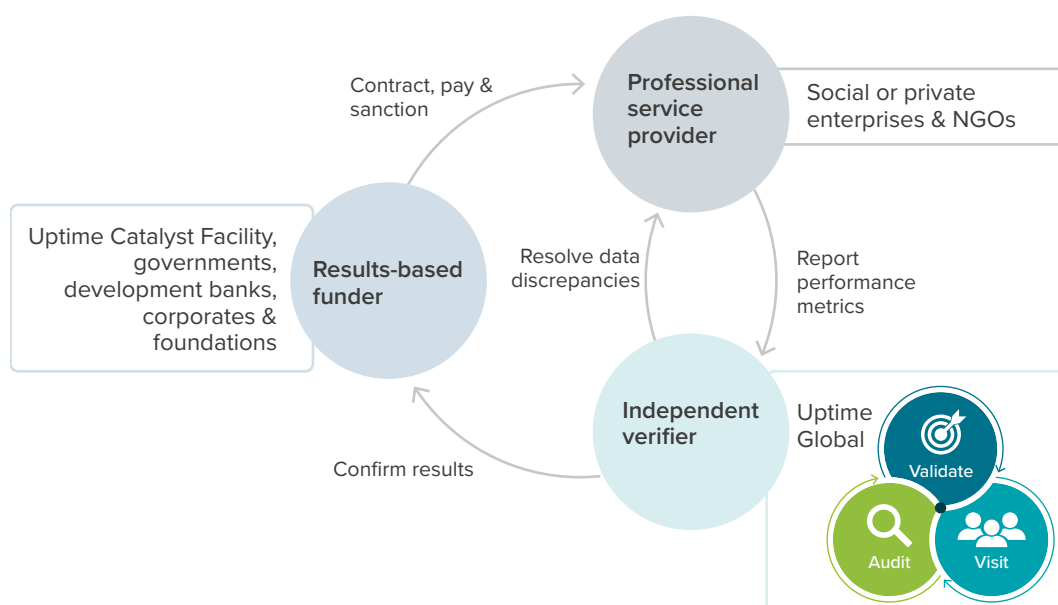
Case study: AI-assisted audit of rural water data

Verifying rural drinking water results

Uptime manages a portfolio of results-based contracts across 16 countries, covering more than 5,500 piped schemes and handpumps, where payments to professional service providers are contingent on verified delivery of reliable drinking water services. Since 2020, more than 5 million rural people have benefited from this model, with over USD 8 million disbursed.

Validation and verification is central to the approach: Uptime employs a triangulated framework combining quarterly data submissions cross-checked against historical records, annual data audits, and site visits. Together, these measures are designed to identify discrepancies, discourage misreporting, inform sanctions, and assure results-based funders that payments reflect actual service delivery.

Figure 1: Uptime's data integrity framework. (Source: Armstrong et al. 2025)¹



¹ Armstrong, A., Naughton, M., Trethewey, F., Nilsson, K., McNicholl, D. and Hope, R. 2025. [The Data Dividend: Applying a data integrity methodology to deliver better value for results-based funding for safe drinking water services](#). Working Paper, Uptime Global.

The data audit is a core component of this framework. Each year, service providers submit supporting documentation for a probability-based sample of handpumps and piped schemes, typically covering several hundred sites in total. The sample is sized to reliably detect material discrepancies across the portfolio, so a meaningful number of sites must be audited regardless of programme size. This means that even for service providers with relatively small operational footprints, the volume of supporting documentation submitted for review each cycle can be substantial.

That documentation spans four performance areas: service reliability, water volumes, local revenues, and water safety. It also reflects wide variation in record-keeping practices across the portfolio, ranging from integrated digital platforms with automated meter readers and cloud-based dashboards to handwritten logbooks and paper receipts, submitted across multiple languages, formats, and levels of completeness².

The scale and diversity of this documentation present a practical challenge in the audit process. In the 2025 audit cycle, approximately 6,500 pages of PDFs, images and spreadsheets were submitted, ranging from handwritten receipts in the Central African Republic to automated meter exports from India. Manually extracting, cross-referencing and assessing this volume of primary evidence is time-consuming, costly and prone to error. As results-based financing programmes scale nationally, the limitations of purely human review are becoming increasingly difficult to sustain.

From human review to machine-assisted auditing

Reducing the cost and complexity of data audits without compromising their integrity is a prerequisite for scaling results-based programmes. Uptime has pursued this by investing in two complementary AI-based approaches to document processing: specialised document intelligence services and general-purpose Vision Language Models (VLMs).

For the first approach, Uptime has utilised Google Cloud's Document AI, a service that combines Optical Character Recognition (OCR) with trained entity extraction processors. Separate processors have been built and trained for each of Uptime's core document types: maintenance and breakdown records, water meter readings, revenue receipts and financial statements, and water safety documentation. Each processor reads a submitted file, identifies the relevant data fields, and returns structured spreadsheet outputs alongside annotated copies of the original documents with bounding boxes drawn around each extracted value (Figure 2). This visual annotation layer is particularly valuable for human review, allowing auditors to quickly verify what the model found and where, without reopening the source files. The approach is consistent, scalable, and largely automated in day-to-day operation once the processors have been trained.

² For more information, see Armstrong, A., Nilsson, K. and Tretheway, F. March 2024. [Data Audit | 2022–2023](#). Uptime summary report.

Figure 2: Structured data extracted from supporting documents. *Institution and community names have been changed to protect privacy. Images are from previous audit cycles and do not reflect current documentation practices. Credit: Whave and International Lifeline Fund*

Date transformed to annual quarter:

2025 Q1

Unique waterpoint ID:

School A

Payment amount in Ugandan Shillings:

270,000

Unique waterpoint ID:

Community B

Date and time of service interruption:

January 21, 2025 at 9:28 am

Service interruption duration calculated from date and time of problem resolution:

5.4 hours

The second approach uses VLMs such as Anthropic's Claude Sonnet to process document types that trained processors handle less reliably, particularly complex spreadsheets with multiple tabs, non-standard layouts, or dense formatting. Rather than applying a fixed extraction model, the VLM receives plain-language instructions describing what to extract and returns structured data alongside natural-language explanations of how each value was found. This conversational, schema-driven approach is more flexible and requires no retraining when document formats change or new data fields need to be captured.

In the 2025 audit cycle, these tools together processed more than 200,000 performance records from approximately 6,500 pages of source documentation at a total AI processing cost of under USD 350, representing more than twelve times the volume extracted in the previous cycle from a comparable number of files.

Performance has improved substantially across audit cycles. Both approaches work best with clearly structured forms, typed reports, and screenshots of digital systems, while spreadsheets with large numbers of rows, columns, or tabs and poorly legible handwritten entries remain more challenging. In the 2024 audit cycle, more than half of AI-extracted records were discarded due to errors or missing information, and the majority of the final dataset required manual input. Since then, accuracy has improved considerably, with each subsequent cycle requiring less manual intervention. Human review remains important, particularly for uncertain extractions and ambiguous or incomplete documentation, but the overall trend toward greater automation is clear.

The next frontier: Retrieval Augmented Generation

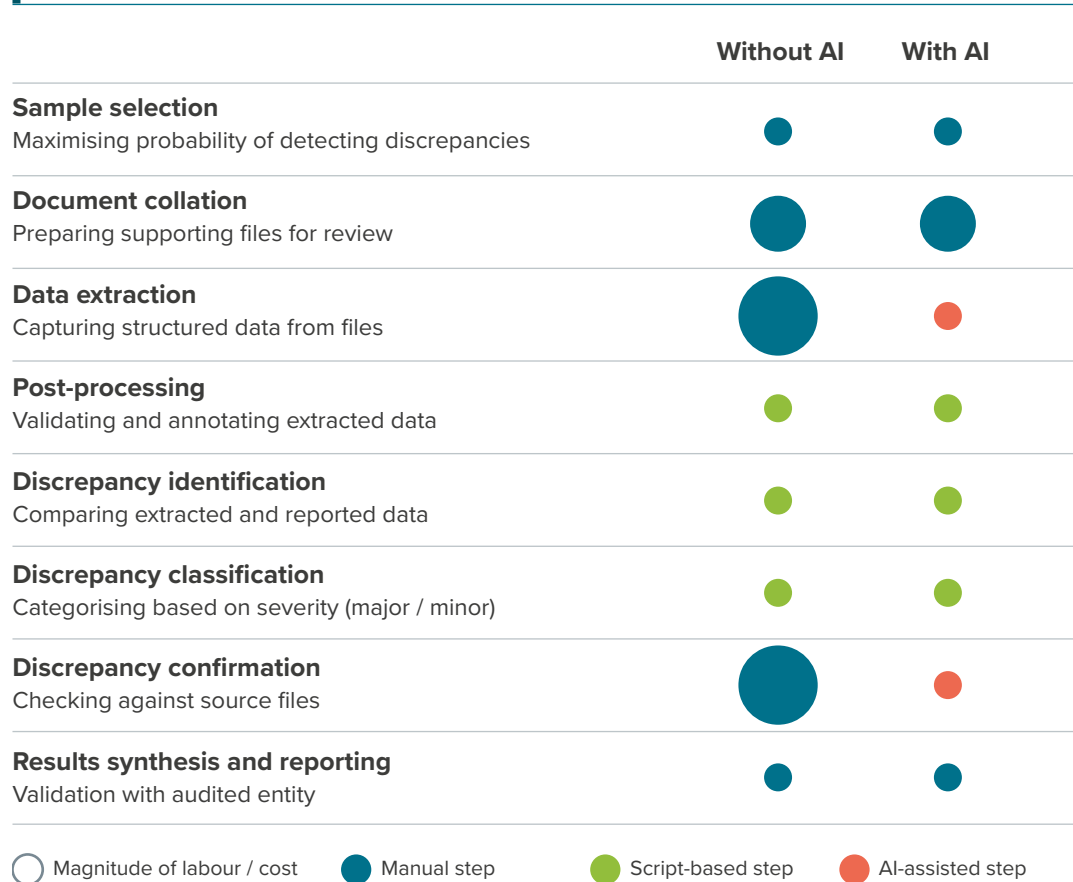
The AI tools Uptime has utilised are primarily designed for data extraction: identifying and pulling specific values from individual documents for comparison against reported performance. A promising next step is to introduce a second stage using Retrieval Augmented Generation (RAG), a technique that automatically pulls relevant source documents to assist analysis, which could further reduce the labour and cost associated with discrepancy investigation.

RAG connects large language models to relevant slices of a knowledge base in real time. Rather than relying on whatever a model has learned during training, it retrieves evidence from approved source documents and asks the model to answer with that evidence in view. This enables the system to move beyond extracting individual data points to assessing whether the documentation as a whole is consistent with reported performance, while maintaining a traceable and defensible evidence trail.

In the envisioned two-stage workflow (see Figure 3), document intelligence services and VLMs first extract structured data from submitted files and compare it automatically against data reported to Uptime for results-based payments, flagging potential discrepancies for further review. In the second stage, rather than requiring an auditor to manually reopen and search through source files to investigate each flagged item, a RAG system queries the original documents directly. An auditor could ask, for example, whether a reported service interruption is corroborated by the submitted maintenance records, and receive a response grounded in the specific pages of the relevant files, with citations to the evidence that supports or contradicts the reported value.

A framework for AI-assisted data audits

Figure 3 illustrates how these AI approaches map onto the broader audit workflow and where they deliver the greatest efficiency gains. The framework distinguishes between three modes of processing across the key audit stages: manual steps, script-based steps, and AI-assisted steps.

Figure 3: Areas where AI enhances efficiency in large-scale data audits.

AI offers the greatest gains at the two most labour-intensive stages. At the front end, data extraction is where document intelligence services and VLMs address the core bottleneck, processing thousands of pages of diverse, multilingual documentation automatically and at a fraction of the cost of human review. At the back end, discrepancy confirmation is where RAG adds its most distinctive value: not through volume, but through the quality and precision of its outputs. By querying source documents directly and returning answers grounded in specific passages with traceable citations, RAG enables targeted, evidence-based confirmation of flagged discrepancies without requiring auditors to manually search through large file sets.

Between these two AI-assisted stages, three intermediate steps benefit from script-based automation. Post-processing validates and annotates data that has been extracted by AI, for example by cross-checking waterpoint identifiers against a master roster, standardising date formats, and flagging low-confidence extractions. Discrepancy identification then compares the cleaned data against reported values, applying predefined tolerance thresholds to distinguish meaningful differences from minor rounding or currency conversion variations. Discrepancy classification categorises each flagged item as major or minor based on its potential impact on payment calculations, whether it appears isolated or systematic, and the strength of the supporting documentation. These steps are well suited to scripted, rule-based logic: they involve structured data, defined decision rules, and reproducible outputs that support the consistency and auditability that compliance processes require.

The remaining stages, including sample selection, document collation, and results synthesis and reporting, involve predominantly manual processes. Document collation in particular remains a practical constraint on scalability in contexts where data systems are not digitised or standardised, since service providers must still compile and submit their own records. Results synthesis and reporting requires human interpretation and contextual judgement that automation cannot yet reliably replace.

Taken together, this framework positions AI as a targeted intervention at the stages where the volume and complexity of primary documentation create the greatest bottlenecks. When supported by scripted processing and human oversight, the combined effect is an audit pipeline that is faster, less expensive, and more consistent than purely human review, without sacrificing the rigour and traceability that results-based funding demands.

The cost case for AI-assisted auditing

One of the most common concerns raised by governments and development banks considering results-based programmes at national scale is the cost of verification. If auditing expenses grow in proportion to the number of sites, verification quickly becomes unsustainable. Uptime's experience provides a practical basis for assessing what AI-enhanced auditing actually costs at scale.

Uptime Global's current multi-country portfolio covers approximately 5,500 sites across 20 service providers, generating around 6,000 pages of supporting documentation per audit cycle at an AI processing cost of around USD 450 per year. A national results-based programme of 100,000 sites supported by 50 to 100 service providers would generate an estimated 8,000 to 10,000 pages, at an AI processing cost of USD 500 to 600. Across a twenty-fold increase in programme size, AI processing costs remain well under USD 1,000 per year.

This favourable scaling reflects an important characteristic of the underlying technology. The per-page cost of AI processing tends to decline as volumes grow and models improve, so the cost curve flattens as programmes scale. Expressed as a proportion of total recurring programme costs, AI processing represents approximately 0.01% for a 10,000-site programme and as little as 0.004% to 0.005% for a 100,000-site programme (based on cost estimates from the SafePani³ and SafeManzi⁴ programmes).

A further driver of cost efficiency is the sampling logic itself. Sample sizes are determined primarily by the assumed discrepancy prevalence and the target detection probability rather than the total number of sites, so the required sample does not grow proportionally with programme size. A national programme of 100,000 sites requires an audit sample of only 400 to 500 waterpoints per year, compared to 300 for Uptime's current portfolio of 5,500.

3 REACH, 2023. [Cost estimates for safe drinking water in schools and healthcare centres in Khulna District, Bangladesh](#). University of Oxford.

4 Hoque, S.F., Lishimpi, K., Phiri, F., Tambatamba, B., Mulundika, M.M., Kalapa, C., Simwanza, T., Siajunza, M., Mugode, M., Hope, R., Charles, K., Nowicki, S., Nshenda, E., McNicholl, D., Nyirenda, U., Muchelenje, J. and Haankuku, C. 2025. [Results-based funding to provide safe drinking water services for public schools and healthcare facilities in Zambia](#). Working Paper. Oxford, UK: University of Oxford and Uptime Global.

AI processing is, however, only one element of total verification costs. Site visits, human review of flagged discrepancies, results reporting, and programme management all contribute, and these do not scale as favourably. Across its current portfolio, Uptime's combined cost of validation, verification, and payment administration averages less than 9% of total results-based contract value. As programmes scale and verification workflows become more automated, there is a credible path to reducing this proportion further, particularly as the AI-assisted and script-based elements of the audit pipeline mature and reduce the labour burden at the discrepancy identification and confirmation stages.

Implications

The case study and cost analysis presented above point to a broader conclusion: AI-assisted data auditing is not a specialist capability confined to technically sophisticated programmes. Results-based contracts for rural drinking water services illustrate what is possible, but the underlying logic applies across development sectors wherever the audit challenge is defined by the volume and diversity of primary documentation.

Realising this potential depends on a clear understanding of what AI can and cannot do in this context. First, AI is one component of a broader audit pipeline. Intentional sample design, progressively improving data systems, and script-based automation all contribute to more efficient and effective audits in their own right. The gains from AI are real, but they are best understood as complementing these foundations rather than substituting for them.

Second, AI does not replace all manual processes. Even when contract terms are clearly articulated and evidence is well-structured, the interpretation of ambiguous findings, the enforcement of sanctions, and the progressive improvement of service delivery require human judgement and institutional oversight. Verification systems will continue to rely on human input and feedback loops, accountability structures, and decentralised technical assistance. AI can reduce the cost of generating and processing evidence; it cannot substitute for the governance arrangements that give that evidence its consequences.

Third, and most fundamentally, AI cannot substitute for reliable primary data. Its effectiveness depends on consistent and credible source data generation. In sectors such as rural water, most data is still produced through error-prone manual processes. These data systems require sustained investment, not only in digital tools but also in skilled administrators, routine maintenance, and ongoing user training. Even when digital platforms are free, substantial recurring investment is required to maintain data quality and usability.

Finally, organisations adopting AI should anticipate a learning period. Early experimentation will involve failures, rework, and tempered expectations. Tools, workflows, and performance thresholds all require iterative refinement before they can be relied upon at scale.

For governments and development banks designing results-based programmes, the practical implication is straightforward. Robust, AI-enhanced data auditing at national scale is not a luxury reserved for well-resourced programmes. As the cost figures above demonstrate, the technology component of verification is well within reach. The more substantial investments lie in building the data systems, training the service providers, and designing the contractual architecture that makes results-based funding work in the first place. AI-enhanced auditing is best understood not as an added cost, but as a tool that protects the integrity and credibility of those investments.

Acknowledgements

Technical expertise on the AI approaches described in this briefing note was provided by Olivier Mills and Ahmed Nasser Mostafa, who contributed to early-stage conceptualisation and were responsible for code development, tool building, and maintaining computational infrastructure. The note was reviewed by Duncan McNicholl, Rob Hope, Meleesa Naughton, Fin Trethewey, and Kristina Nilsson.

This paper should be referenced as:

Armstrong, A. (2026). AI-assisted data audits in results-based programmes. Briefing Note, Uptime Global.