KEEPING CONTROL

Practical Workflows for Using Generative AI in Evaluation

AI can accelerate evaluation, but only when evaluators remain firmly in control.

At Itad, we use human-led workflows that combine generative AI with evaluator expertise to classify, extract, code and synthesise evidence at scale. Drawing on real evaluation projects, this poster shows where AI adds value, the safeguards needed to maintain quality and trust, and the lessons we are learning about traceability, verification and responsible use.

1
Systematic evidence review at scale

Extract evidence from large collections of documents using a predefined analytical framework.

Typical workflow

1. Develop an extraction framework.
2. Provide context documents and guidance.
3. Review against the framework and context using AI.
4. Verify and systematically capture findings for analysis.

Best used when
Conducting literature reviews, portfolio reviews, proposal assessments or large-scale qualitative analysis where consistent evidence extraction is needed.

2
AI-assisted classification for evaluation

Classify documents, proposals, results or other evidence against predefined taxonomies.

Typical workflow

1. Create taxonomies with detailed descriptions.
2. Develop and test AI prompts that apply those taxonomies consistently.
3. Prepare and process data in manageable batches.
4. Review outputs using evidence tables and data visualisations to support analysis.

Best used when
Working with large datasets where evidence needs to be categorised consistently. For example, analysing the characteristics of applicants in a research funding portfolio.

3
Codebook-guided evidence analysis

Apply an agreed codebook to large volumes of qualitative data and generate themes for manual review.

Typical workflow

1. Prepare and batch documents; check readability.
2. Define and test AI coding rules and evidence thresholds.
3. Test outputs against manual coding and run batches.
4. Review, refine and compare emerging themes.

Best used when
Analysing large numbers of reports, transcripts or multilingual documents while retaining human judgement in interpreting findings.

Where AI adds value

Spotlight on Itad's work

AI-assisted multilingual coding

5-10 times faster than manual approaches

For the Stiftung Mercator Foundation, we undertook a Data Analysis, Integration and Synthesis (DAIS) process. We used AI to analyse 80 to 120 grantee reports in German, French and English against a shared codebook.

Value added: Coding was estimated to be five to 10 times faster than manual approaches. Human benchmarking and spot-checking maintained quality, while reusable prompts enabled our client to replicate the approach.

Spotlight on Itad's work

Wellcome Climate Impacts Awards: Process evaluation

Over 600 grant proposals assessed

As part of a process evaluation, we used AI to classify over 600 funding applications - against more than 15 assessment criteria in the latest round. This did not inform the selection or award of applications.

Value added: Between the two rounds we improved how we processed the proposals. This, combined with AI model improvements, has allowed us to review more assessment criteria with greater accuracy and removed the need for manual assessment of some criteria.

4
Zero-draft evidence synthesis

Generate an initial synthesis from structured evidence to support sensemaking and expert review.

Typical workflow

1. Compile relevant evidence.
2. Define the synthesis questions.
3. Prepare a structured prompt and generate an initial draft using AI.
4. Refine through manual review and iteration.

Best used when
Exploring emerging findings, preparing for key informant interviews or sensemaking workshops, and supporting thematic analysis.

5
Portfolio outcome harvesting

Identify, extract and classify outcomes from reports and other documentary evidence to create a consistent outcome harvest, including descriptions of outcomes, contribution, significance, sustainability, and supporting evidence.

Typical workflow

1. Define what counts as an outcome and the information to capture (including relevant taxonomies).
2. Design a structured AI prompt and outcome harvesting framework.
3. Process documents systematically, in manageable batches, using AI.
4. Verify and consolidate results, removing duplicates, and undertaking quality assurance manually.

Best used when
Portfolio evaluations, learning reviews and retrospective analyses where outcomes must be identified systematically across multiple sources.

Spotlight on Itad's work

Document-to-Insight (D2I) system for portfolio learning

1000+ fully searchable, structured insight products

For a global climate Foundation, we developed a system that transformed thousands of grant documents into 1000+ fully searchable, structured insight products.

Value added: Thousands of documents can be analysed in days rather than months. The system supports learning and reporting through fully traceable evidence and transparent source links.

6
Evidence-based learning narratives

Transform evidence into concise narratives that answer learning questions and support strategic decision-making.

Typical workflow

1. Use AI to extract findings and lessons (Box 1).
2. Use AI to classify evidence using common frameworks (Box 2).
3. Use AI to identify patterns across programmes, portfolios or themes (Box 4).
4. Generate insight products using AI to synthesise one to three and review manually.

Best used when
Supporting board reporting, portfolio learning, strategy development and organisational reflection.

Five lessons from real-world use

1 Verification is part of the method, not a separate quality assurance step

Assessing AI for completeness, relevance, accuracy and fabrication helps us understand where it performs well and where additional review is needed.

2 Traceability builds trust

Stakeholders are much more willing to engage with AI-generated findings when they can trace outputs back to original documents, quotations and text passages. This ability to interrogate the evidence matters as much as the quality of the synthesis itself.

3 Asking AI to explain itself improves review

Explicitly asking the tool to explain why it applied a particular code or classification, can often reveal weaknesses in frameworks, prompts and coding structures. This helps enable human review where it is needed most.

4 People care about completeness, as much as accuracy

The first question asked by people using AI outputs is often not 'is this right', but 'did it miss anything important?'. Including with each output a short statement of what the system reviewed and how it selected content can help address these concerns.

5 AI works best when the task is clearly defined

AI performed most consistently when we gave it a clear framework, explicit decision criteria and well-defined outputs. Allowing AI to make open-ended judgements produces much more variable results.

AI and evaluation: Where do you stand?

Author: Chris Perry | chris.perry@itad.com **www.itad.com** Itad

itad