itad

The evaluator's guide to generative AI

How we're using generative AI at Itad: what works, what doesn't and practical guidance for evaluators who want rigour in their AI-supported analysis.

Chris Perry, Michael Moses, Max Pinter

March 2026

Generative AI is everywhere, but is it the answer to Life, the Universe and Everything?

In this learning brief, we explore the implications of using AI for evaluation, including lessons on what's worked (and what hasn't) for us so far - with just a few Hitchhiker's Guide to the Galaxy references thrown in along the way.

CHAPTER ONE

The rise of generative AI

(or ‘don’t panic’)

Generative AI increasingly sits alongside project data, interview transcripts and theory-based frameworks to help teams search, extract, summarise and sense-make at speed.

Clients are also asking how AI can help them use the evidence they already have, surface signals in large document sets and keep quality high while time and budgets stay tight.

At Itad, we’ve been exploring AI in real delivery with the simple starting premise: AI should augment expert judgement, not replace it.

To approach AI safely and pragmatically, we:

- Use models where they add value - such as document triage, structured extraction, first-pass synthesis
- Design human-in-the-loop checks for anything interpretive and keep quantification and nuance with with analysts
- Work within client requirements and data security expectations, choosing tools and workflows that respect confidentiality

Below are three case studies of AI use in our recent work, and the lessons we’ve learnt from this.

CHAPTER TWO

How we've been using AI

(or, 'almost, but not quite, entirely unlike a research assistant')

Case study one | **Grant proposal analysis and classification with Copilot**

Wellcome's Climate Impacts Awards back short, high-impact projects that combine evidence generation, stakeholder engagement and policy influence to drive urgent climate policy action at scale. Our evaluation assesses the effectiveness and implementation of the Awards.

We used **Microsoft Copilot 365 within Itad's secure systems** to extract and synthesise evidence from across hundreds of proposals and to classify proposals. This meant the evaluation team could focus on sense-making, triangulation and recommendations.

We used Copilot to:

- Iteratively develop an analytical framework aligned to three evaluation questions
- Extract structured evidence from year one and two proposals
- Synthesise this data and identify themes and patterns
- Classify year three proposals against a range of criteria

This fed into deeper, human-led synthesis, integrating Key Informant Interviews (KIIs) a document review, and a survey.

How successful was CoPilot?

A post hoc verification of around 15% of year one and two proposals informed a strong confidence rating for extracted evidence on completeness, relevance, accuracy and fidelity and absence of hallucinations/fabrications.

The evaluation team also assessed the syntheses generated by Copilot as 'strong' for relevance and 'adequate' for insightfulness.

The syntheses provided comprehensive information; however, it lacked nuanced differentiation across applicant groups and was insufficient on its own for generating deep insights. Copilot was unable to reliably quantify trends observed in the synthesis, and these counts were therefore not used.

When attempting to classify year three proposals, CoPilot required a two-step process: data extraction to excel *then* classification. It performed well on structured data but struggled with nuanced classifications requiring contextual understanding.

We engaged with the Wellcome team throughout the process to transparently build an understanding of Copilots strengths and limitations.

What enabled good results

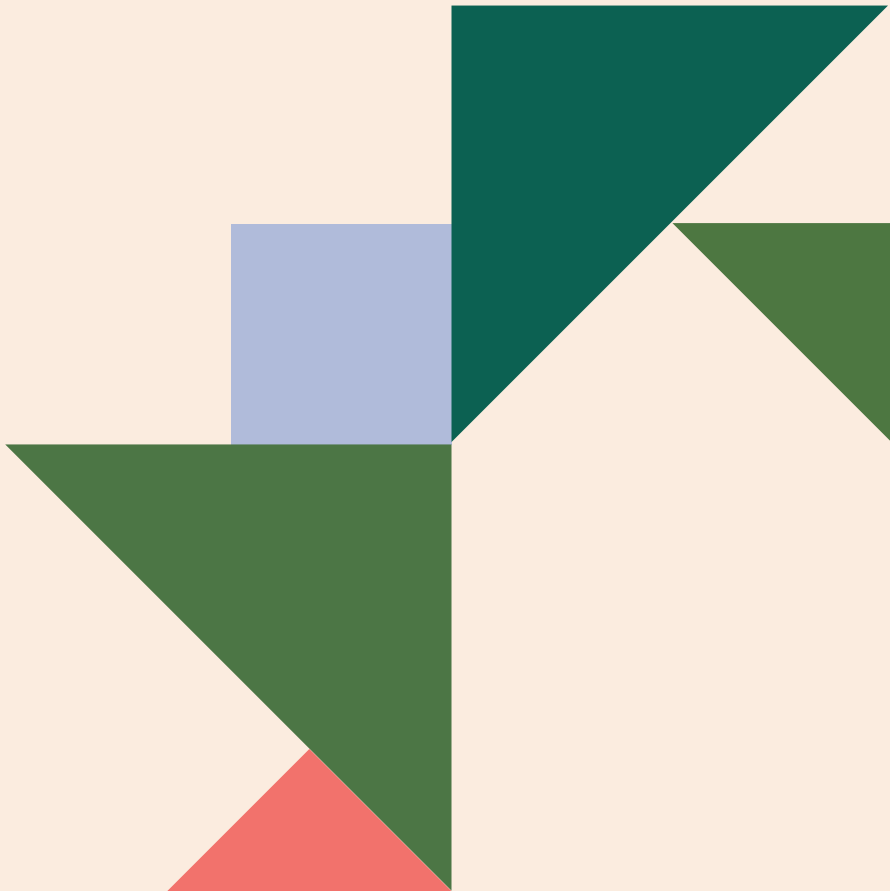
- **Framework-led prompts:** clear components, definitions and exemplar answers helped anchor Copilot, reducing drift and off-topic content and improving fidelity across all proposals.
- **Task decomposition:** separating extraction from classification made Copilot's job more straightforward and facilitated easier verification.
- **Human-in-the-loop:** spot checks and rubric-based scoring built confidence, and syntheses generated by Copilot helped augment, rather than replace, evaluator judgment.

Limitations and mitigations

- **Quantitative fragility:** aggregated counts produced by Copilot were inconsistent. We limited its role in quantification and relied on spreadsheet formulas and human quality assurance (QA) for totals and derived metrics.
- **Nuanced classification challenges:** context-dependent criteria and ambiguous phrasing in proposals led to misclassifications. Instead, we relied on some manual coding and prioritisation of other classifications to meet time constraints.
- **Limited differentiation across applicant groups:** Copilot syntheses sometimes exhibited a "meets criteria" bias, blurring distinctions between longlisted/shortlisted/awarded groups. We anchored insights in KIIs and committee scoring data, using Copilot outputs as corroboration rather than lead analysis.

In summary

In short, Copilot is a strong analysis assistant for data extraction, synthesis and classification where signposting is clear. It is best used within an extract to classify workflow and paired with human verification for nuanced attributes and any quantification. This combination helped us deliver timely, comparable insights while preserving evaluative rigour.



CASE STUDY TWO | **Processing key informant interviews with CoLoop AI**

The MacArthur Foundation's Big Bet On Nigeria invested around $154m to support Nigerian-led efforts to reduce corruption by promoting transparency, participation and accountability. We recently completed the final evaluation of the ten-year programme.

We deployed a mixed methods approach on the final evaluation, including development and analysis of a representative set of case studies. We used CoLoop.AI to:

- Record and transcribe 50+ KIIs as they happened.
- Translate interviews conducted in non-English languages.
- Support analysis of the collected data, with a view towards accelerating and strengthening the evaluation team's traditional coding and analysis process.

We theorised that CoLoop would enable us to collect and use more data than we otherwise would have been able to.

How did CoLoop perform?

Overall, CoLoop performed well at transcription, and was significantly faster than our traditional approach to writing up transcripts manually. Our team thoroughly reviewed and corrected every transcript CoLoop produced, finding only minor mistakes each time.

We found data analysis was much less straightforward. We initially hoped structured prompts in CoLoop's analysis grid would replace manual coding, but pilot testing showed CoLoop oversimplified and struggled to deal with the nuance and complexity inherent in our data. As a result, the generated themes were inadequate for answering our evaluation questions.

We tried a few different approaches to enrich CoLoop's theme production, including iterating prompts multiple times and using the platform's chat function instead of analysis grids. After several failed attempts, we landed on a new approach: download and code each transcript by hand, then upload back into Coloop, and analyse each set of coded excerpts individually.

This meant that CoLoop didn't have to make sense of lots of divergent data. Instead, we could tell it what the overarching code for a given set of excerpts was and limit the context window. This finally helped us develop within case themes that were both accurate and sufficiently nuanced, while also being easily bolstered by the underlying evidence - ultimately what we needed for our synthesis process.

What enabled good results

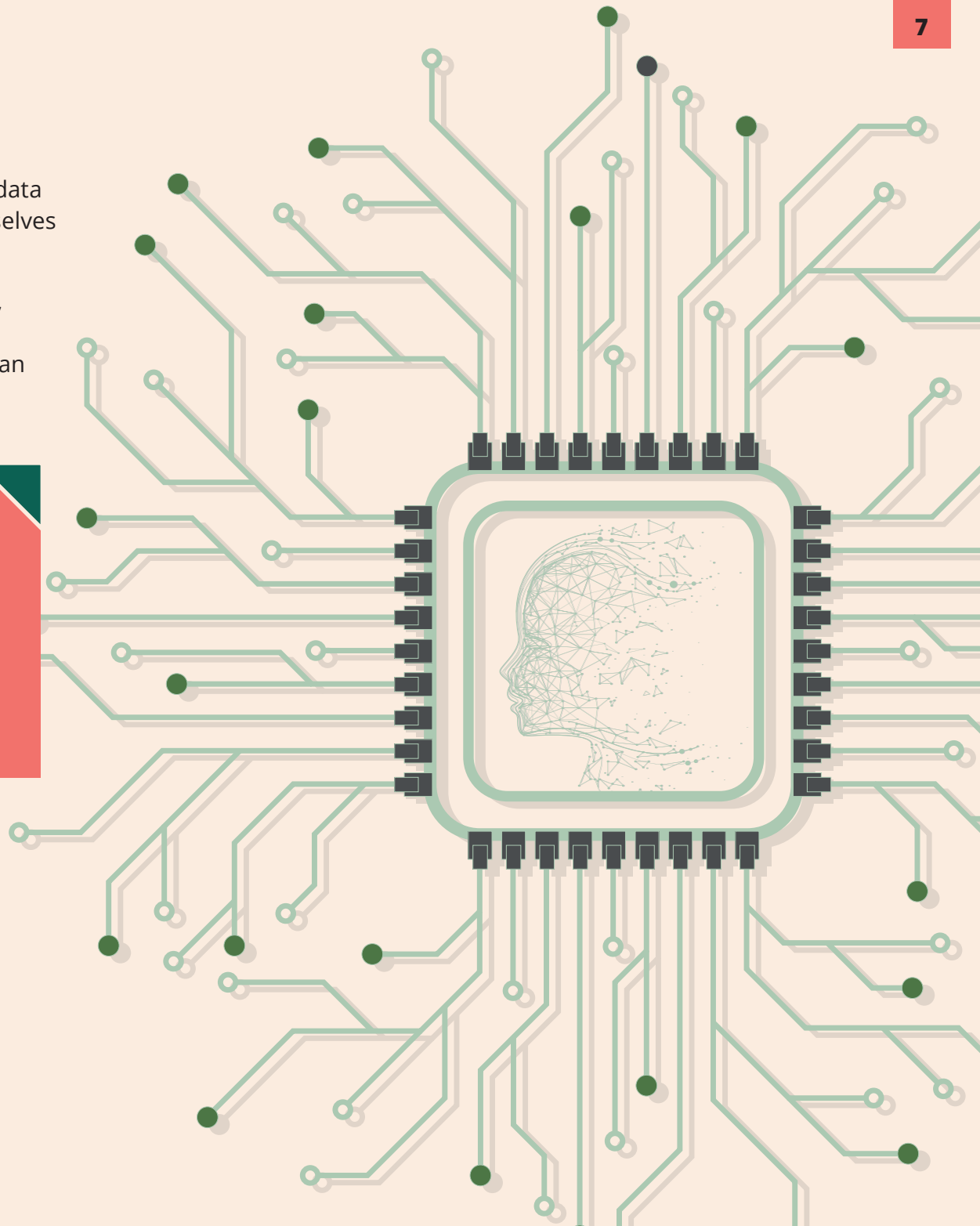
- **QA, over and over again:** we checked everything that CoLoop produced to confirm final transcripts and ensure accuracy (and, eventually, extra efficiency) in our analysis process.
- **Trial and (lots of) error:** team-wide reflection sessions and working together to develop and test new hypotheses were essential steps to find an approach that worked.
- **Make it simple:** asking CoLoop to do tasks related to only one chunk of a workflow at a time helped it better engage with the data and provide richer analysis.

Limitations and mitigations

- **Nuanced classification challenges:** CoLoop struggled to 'understand' the context of cases and tended to flatten the data when generating themes, requiring us to code the data ourselves pre-analysis by CoLoop.
- **Conceptual fuzziness:** CoLoop struggled to relate nuanced, general concepts to the specifics of the program we were assessing. Providing a set of definitions in our prompts was an important part of the analysis process.

In summary

CoLoop is a good translator, and - if set up to succeed - can help to accelerate production of themes across a big set of KIIs. However, human-led quality assurance is essential, as is manual coding and breaking down analysis tasks into their components pre-AI intervention.



CASE STUDY THREE | **Grantee report analysis and synthesis using multiple AI tools**

One of our partner's supports organisations throughout Africa that focus on generating evidence and data. The aim is to enhance local evidence production and strengthen the ability and motivation of policymakers to utilise evidence in their decision-making processes.

We assessed the value contributed by the broader sector working towards these objectives, evaluated our partner's role within it, and considered potential responses to developments in the field.

We used several AI tools during the evaluation:

- **Copilot** was used to iteratively develop and refine our analytical framework, using trial and error to develop prompts that produced structured and consistent responses. We also created a 'Copilot Notebook' to store contextual information and extract evidence from around 450 grantee documents. Copilot was used to synthesise results by grantee and country, identifying patterns and gaps.
- **ChatGPT** was used to develop high-level contextual analysis for three country case studies
- **MaxQDA's** built-in AI tool, **Tailwind**, supported the synthesis of our coding of KIIs and FGDs.

What enabled good results

- **Different tools for different jobs:** we tested different AI tools throughout the evaluation to understand where they added value and where their limitations lay.
  ChatGPT was most effective for contextual analysis; Copilot's security features made it appropriate for sensitive grantee documentation; and MaxQDA integrated directly with our manual coding.
- **Human supervision:** AI worked best as a supporting tool for human analysts rather than as drivers of analysis, complementing rather than replacing human judgment.

Limitations and mitigations

- **Weaknesses in judgement:** AI outputs were stronger when synthesising information from material that had already undergone a level of analysis. When asked to conduct a standalone analysis, they lacked the nuance required for evaluative work.
- **Eager to please:** the AI tools often identified too many patterns in the data. For this reason, we chose not to use MaxQDA's automated AI coding, as it significantly over-coded transcripts.
- **Formatting weaknesses:** despite improved prompting, Copilot remained inconsistent in producing outputs in the precise format required for subsequent Excel coding, requiring analyst intervention and increasing the level of effort.

In summary

AI tools proved valuable throughout the evaluation, particularly for discrete, well-defined tasks. They were most effective when embedded within analysts' workflows as one tool among many - supporting, but not defining, the analysis.

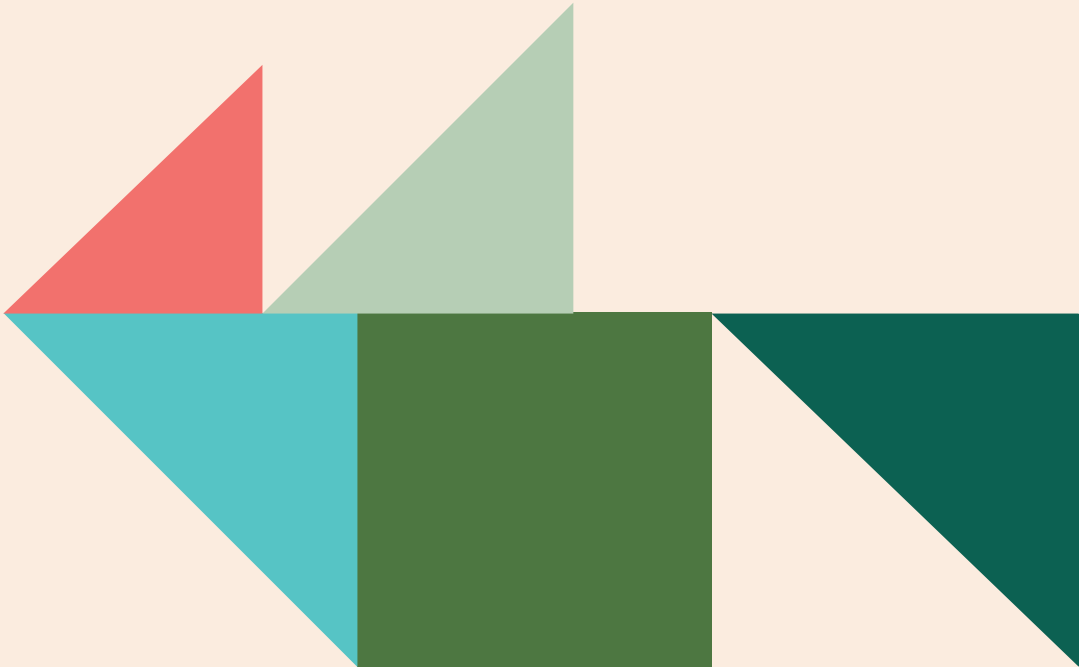
CHAPTER THREE

Lessons for rigorous AI use

(or ‘so long and thanks for all the synthesis’)

Across the cases considered, some constants emerged:

- **AI is great at extraction; humans are great at sensemaking:** treat AI tools as tireless research assistants that structure large volumes and surface candidate patterns. Leave the interpretation, trade-offs and conclusions to evaluators.
- **Break the work into smaller steps:** break complex analysis into stable units to reduce drift and improve reproducibility. For example: *define the framework → extract evidence → classify (where appropriate) → synthesise by component → integrate across sources.*
- **Design prompts around your framework:** provide precise component definitions, examples and formatting instructions. Where possible, ground the model with a lightweight “notebook” or reference pack and reuse it consistently.
- **Always verify the results:** build sampling and scoring into your plan (completeness, relevance, fidelity and hallucination checks). Use spreadsheets/databases for counts and derived metrics.
- **Limit the context for nuanced analysis:** for nuanced concepts, analyse within a limited number of codes, cases, or components rather than ‘everything at once’.
- **Choose the right tool for the job:** consider data sensitivity, model strengths and integration with the team’s existing technology stack (for example, secure environments for confidential docs or dedicated Computer-Assisted Qualitative Data Analysis Software (CAQDAS) for coded data).
- **Document your process:** keep prompts, versions, verification notes and decisions. This supports auditability, client confidence and internal capability building.
- **Mind the ethics and governance:** agree on data handling, client approvals for tool use and explain where AI sits in your method. Be transparent about limits and when human adjudication applies.



CHAPTER FOUR

What comes next?

(Or ‘the restaurant at the end of the universe)

As organisations seek to harness AI for rigorous analysis, the next challenge is scaling up its use to deliver faster, more consistent insights while preserving nuance and ensuring traceability. The priority is to empower teams to focus less on manual tasks and more on judgment, insight, and actionable recommendations.

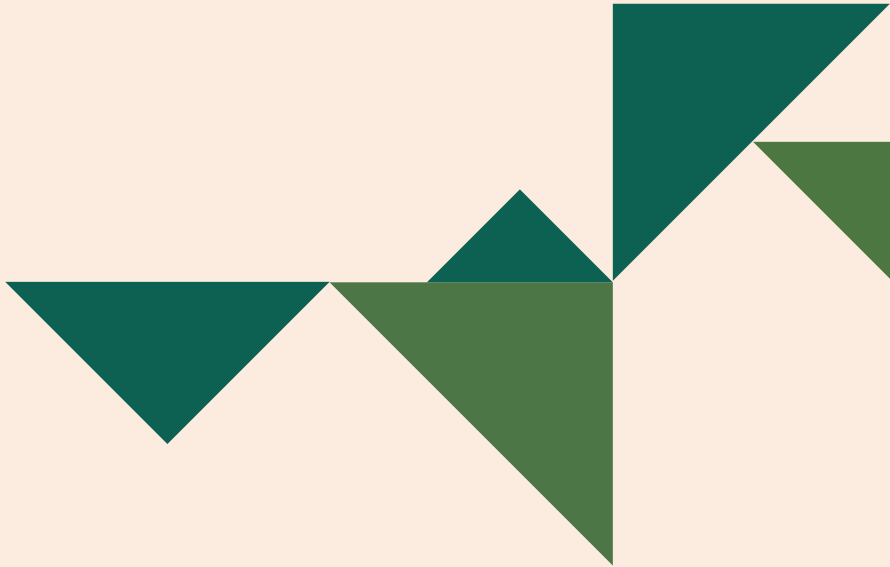
To achieve this, consider the following forward-looking strategies:

- Make use of AI models where they add clear value such as automating document triage, extracting structured data, and synthesizing information from large sources. This frees human experts to interpret findings, make nuanced decisions, and provide context-specific recommendations.
- Implement governance that keeps analysts firmly in the loop. AI can manage routine tasks, but quantification and nuanced analysis must remain under the purview of skilled analysts to ensure accuracy and relevance.
- Adopt tools and workflows that prioritise data security and confidentiality, aligning technology choices with organisational requirements and stakeholder expectations.

At Itad we are building bespoke platforms that fully embed AI models within evaluation workflows – providing even greater efficiency gains through selective automation. For example, we are currently developing a platform for a Foundation partner that will allow them to surface learning and insights from a largely unstructured corpus of grantee documentation amassed over the last 5+ years. This project builds on a tool we developed for an evaluation in 2025 – that allows us to streamline the extraction of insights from large volumes of grant documentation.

In parallel, we are developing our own internal platform – that will raise the baseline across all projects and makes our best AI enabled practice available to every team. The platform will serve as a foundation for supporting emerging service delivery areas such as foresight consulting. For example, we’ll pilot workflows for horizon scanning, signal detection, scenario analysis and options appraisal that combine model-led exploration with expert facilitation and client co-creation.

As we scale these new frontiers in AI-powered analysis, we’ll remember (just like in the Hitchhiker’s Guide) that sometimes the answer is simple, but asking the right questions makes all the difference.



Solving complex challenges with evidence and learning

Itad UK
International House
Queens Road
Brighton, BN1 3XE
United Kingdom
Tel: +44 (0)1273 765250

Itad US
c/o Open Gov Hub
1100 13th St NW, Suite 800
Washington, DC, 20005
United States

Itad Kenya
1870/610 The Westwood Building
Vale Close
Westlands, Nairobi
Kenya

itad.com
@ItadLtd
Itad
mail@itad.com