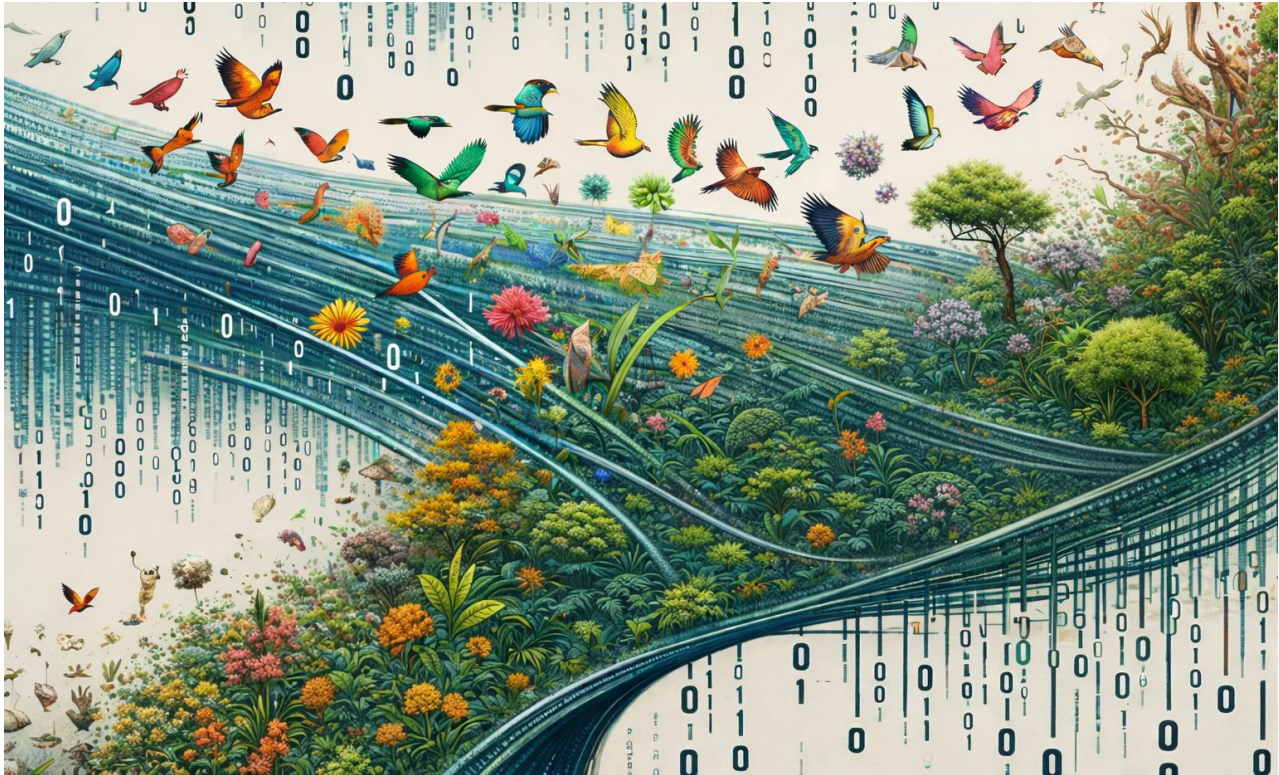




Building & Applying Evidence in the Biodiversity Challenge Funds

Piloting Artificial Intelligence solutions in the BCFs – Deep Dive II

Defra

Date: August 2026

Contents

Glossary	3
Authorship and acknowledgements	3
Disclaimer	4
Executive Summary	5
1. Introduction: study purpose and background	6
2. Use-case pilot findings and insights	7
2.1 Information retrieval AI chatbot: finding insights from BCFs project documents	7
2.2 Grantee Report reviewing: AI guidance and tools	8
2.3 An Evidence synthesis AI tool - drawing from findings and lessons across BCFs funded projects	10
2.4 International Climate Fund Rio Marker screening tool	11
2.5 Summary of key challenges and opportunities for full AI operationalisation	13
2.6 Recommendations	15
Annexes	16
Annex 1: AI guidance to reviewers: Final Report Reviews	16
Annex 2: NIRAS Principles for responsible AI use	19

Glossary

Term	Definition
Artificial intelligence (AI)	The intelligence of machines or software, as opposed to the intelligence of humans or other living beings.
Application Programming Interface (API)	A way for two or more computer programs or components to communicate with each other.
Biodiversity Challenge Funds (BCFs)	The UK Government's flagship bilateral grant schemes for biodiversity, namely the Darwin Initiative, The Illegal Wildlife Trade Challenge Fund, and Darwin Plus
International Climate Finance (ICF)	UK International Climate Finance (ICF) helps the UK deliver on our commitments to developing countries under the Paris Climate Agreement. ICF can include activities which reduce deforestation, protect nature, and increase capacity for climate adaptation and resilience.
Generative pre-trained transformer (GPT)	A type of AI large language model that is good at a wide range of tasks (from classification to generating text). The same type of model used for ChatGPT.
Large Language Model (LLM)	A family of AI models that are trained on large amounts of text data (usually internet webpages) and can then do tasks related to processing, analysing and generating text information. GPTs (above) are type of LLM.
Monitoring, Evaluation and Learning (MEL)	Systematic collection, analysis and presentation of data to assess the progress, effectiveness and impact of a project or programme.
Retrieval-Augmented Generation (RAG)	A technique for enhancing the accuracy and reliability of generative AI models with facts fetched from external sources. It allows the AI models to draw from external sources when generating answers.

Authorship and acknowledgements

This report was authored by Matthew McConnachie. Inputs to the planning and execution of the study, and comments on drafts of the report were provided by Alex Harnett and Victoria Reilly-Pinion.

Disclaimer

NIRAS is the fund administrator for the Biodiversity Challenge Funds and commissioned this work on behalf of the Department for Environment, Food and Rural Affairs (Defra) under Workstream 5 of the Biodiversity Challenge Funds.

NIRAS works with a range of specialists and consultants to carry out studies and reviews on the Biodiversity Challenge Funds. The views expressed in the report are entirely those of the author and do not necessarily represent the views or policies of Defra, NIRAS or the Biodiversity Challenge Funds. Defra and NIRAS, in consultation with wider stakeholders as relevant, are considering all findings and recommendations emerging from this study in how they manage the Biodiversity Challenge Funds.

Your feedback helps us ensure the quality and utility of our knowledge products. Please email BCF-Comms@niras.com and let us know whether or not you have found this material useful, in what ways it has helped build your knowledge base and informed your work, or how it could be improved.

Cover photograph: Image generated with DALL·E 3

Executive Summary

This second phase of the *Piloting Artificial Intelligence solutions in the Biodiversity Challenge Funds* deep dive tested how artificial intelligence can be used safely, responsibly and practically within BCFs grant management. Building on the first phase, which identified areas of opportunity, this study focused on four priority use-cases:

- portfolio information retrieval;
- report review support;
- evidence synthesis; and
- International Climate Fund (ICF) Rio Marker screening.

The pilots show that Artificial Intelligence (AI) can add most value where BCFs processes involve large volumes of qualitative information, repeated review tasks, structured screening against guidance, and the need to draw lessons across a broad portfolio of projects. The strongest immediate opportunities are in the evidence synthesis and ICF screening tools. Both tools are operational, support traceability through references to source material, and can help staff extract, classify, and structure information more efficiently while retaining human oversight and final decision-making.

The evidence synthesis tool was able to replicate findings from a previous manual analysis, indicating strong potential to reduce the level of effort required for future synthesis tasks. The ICF screening tool provides draft classifications, explanations, and page references, which can streamline reviewer quality assurance and may generate clearer time savings during future application rounds. For report reviewing, immediate time savings were more limited because AI was used deliberately as an assistant rather than a substitute for reviewer judgement. However, the VfM case remains strong where AI improves consistency, quality assurance, and risk reduction within existing review effort, particularly as reporting requirements expand.

Overall, the pilots suggest that AI adoption should now move from isolated testing to managed operational rollout. This will require clear BCFs ownership, secure tool access, practical guidance, targeted training, and ongoing tracking of both time savings and quality improvements. AI should be embedded where tasks are repeated, evidence-intensive, and sufficiently bounded, with clear safeguards to ensure outputs are checked, transparent, and aligned with BCFs standards.

1. Introduction: study purpose and background

The purpose of this second BCFs AI-focused deep dive study was to pilot and operationalise the use of Artificial Intelligence (AI) for priority use-cases across BCFs grant management processes. A previous deep study (in late 2023) identified broad areas where AI could have potential to enhance BCFs grant management work processes (Table 1, column 2). This second deep dive focused on piloting and operationalising the use of AI for four priority use-case areas (Table 1, column 3). Use-cases were prioritised that could offer strong improvements in quality and productivity of BCFs grant management processes. For each use-case described in Section 2, we explain the specific reasons for selecting each use-case.

For all use-cases, the focus was on how to get the most out of AI and how to do so in a responsible and ethical way. Like with the use of all technologies, the use of AI comes with risks that need to be carefully managed. Examples of key risks are data leakages, accuracy, language and cultural biases of AI models, and environmental impacts.

The overarching purpose of the deep dive was to draw lessons and insights that would help operationalising AI for grant management. The intended audiences for the study included the BCFs team and Defra, as well as external stakeholders.

Table 1: The use-cases that were piloted for the deep dive (emboldened text, column 3). Column 1 lists the overarching BCFs grant management process areas and column 2 summarises the findings from the initial scoping study.

Overarching process areas of BCFs fund management	Potential for AI (deep dive 1 identified)	Use-case tasks (bold indicates focus of deep dive 2)
Knowledge management of the BCFs portfolio	Moderate to High	BCFs portfolio information retrieval (see Section 2.1)
Application phase (screening of applications)	Moderate to High	Assessments of applications/projects: ICF Eligibility tests , ESA10 (Section 2.4)
		Feedback letter preparation
		Expert reviews into sift tables (data transposition)
Reporting phase (review of progress reporting)	High	Annual report review
		Final report review (Section 2.2)
		Half year report reviews
		Synthesis of reviews
Evaluation and evidence workstream	Moderate to High	MEL (Monitoring, Evaluation and Learning) synthesis work and reviews of project evidence systems. (Section 2.3)
Communications	Moderate	Creating content for comms.

Overarching process areas of BCFs fund management	Potential for AI (deep dive 1 identified)	Use-case tasks (bold indicates focus of deep dive 2)
Administration of fund	Moderate	Document redaction
		Checking financial claim accuracy

2. Use-case pilot findings and insights

Below for each of the selected four use-cases (listed in Table 1), we present findings and insights from piloting the use-cases. We document the progress and challenges made in operationalising the use-cases, along with next steps that could be taken by the BCFs team for each use-case area.

2.1 Information retrieval AI chatbot: finding insights from BCFs project documents

Purpose and background

The purpose of this use-case was to pilot an AI chatbot tool that could search through BCFs project documents (e.g. application forms and reports), and find relevant examples and insights based on questions asked by BCFs staff and Defra. The chatbot tool would be highly secure and integrated with the BCFs SharePoint where documents are currently stored. It would be able to reference the site and documents that were used to answer questions. It would be based on the Microsoft Azure AI tool ecosystem, using Retrieval-Augmented Generation (RAG) technologies. Azure is one of the leading enterprise solutions which developers use for building highly secure AI solutions. An initial prototype chatbot tool was developed and tested during the first scoping deep dive study.

Reason for selecting the use-case and expected return on investment

The potential return on investment for the chatbot is hard to fully quantify but would help focus NIRAS capability for requests which require specific examples on a wide range of topics (i.e. better and more rapidly draw on the portfolio of projects). This would help support Defra commissions and both internal / external comms on a wide range of topics ensuring these are the best possible examples. These types of requests are hard to anticipate and thus categorise projects in advance, with lots of potential offered by AI.

Status and findings

During implementation of the AI deep dive, updates were still ongoing on with the new BCFs platform, therefore it was decided to pause the full implementation of the chatbot until after development was completed. A key step with building the chatbot required linking it to the correct documents through an established folder structure. With the updates being made in the BCFs folder database, it was decided to pause the implementation of the chatbot. Since then, the decision was made to retain our current folder structure.

Since the first deep dive, the introduction and widespread availability of Microsoft 365 Copilot across NIRAS has altered the business case for developing a separate BCF-specific information retrieval chatbot. Copilot is now directly integrated with SharePoint and, while not as comprehensive as a dedicated AI tool, it can search, retrieve, summarise, and analyse information from documents that users already have permission to access. This means that many of the core functions originally envisaged for the bespoke chatbot can now be achieved using Copilot

without the need to develop and maintain a separate user interface or technology solution. At the same time, improvements in AI model capabilities mean that substantially larger volumes of documents can be processed in a single interaction, while costs have fallen considerably compared to when the original concept was developed in 2023¹.

Recommended next steps:

- As a result of the development and integration of Copilot, the BCFs' focus should shift from building a standalone chatbot towards ensuring that BCFs documentation is well organised, appropriately structured within SharePoint, and readily accessible to AI-enabled search and analysis tools.
- The BCFs team should be trained in how to maximise the use of Copilot to retrieve information on SharePoint and other information sources that Copilot has access to.
- A bespoke retrieval tool may still be justified in future if there is a need for specialist functionality, tailored workflows, or access to document collections that are not easily supported through standard Copilot capabilities.

2.2 Grantee Report reviewing: AI guidance and tools

Purpose and background

The purpose of this use-case was to test AI tools and guidance that can help reviewers use AI in an effective, responsible, and ethical way when reviewing project annual and final reports.

Reason for selecting the use-case and expected return on investment

With additional reporting requirements on BCFs projects come additional requirements of desk-based reviewers to appropriately conduct due diligence of projects through the review process – for example, expanded reporting on Standard Indicators and additional requirements on assessing projects' approaches to integrating Gender Equality, Disability and Social Inclusion (GEDSI). Our standard level of effort for a desk-based review is 1 day (of ~200/50 desk-based annual/final report reviews per year respectively), but with the additional requirements on our reviewers the length of time (and so potential cost) has been creeping upwards. There is a risk that reviewers therefore sacrifice important parts of the process to keep within this day of effort. Therefore, this use-case seeks to provide AI some of the more procedural parts of the review process to ensure reviewers add value in the places where it is needed. We anticipate reviewers will increasingly be using AI, so it is important to establish robust guidelines.

Status and findings

Steps taken to operationalise the use of AI for project report reviewing:

The piloting involved sharing AI guidance with final report reviewers and giving access to a secure AI tool that they could use. The guidance covered three main areas (see Annex 1 for a full copy of the guidance):

1. Overarching AI usage best practices that should be applied across all reviewer tasks,

¹ Since 2023, the cost of using advanced AI tools has fallen by around 90-99%, while their capabilities have improved significantly. Tasks that previously required relatively expensive AI systems can now be carried out using more capable models at a fraction of the cost. This has made routine use of AI for activities such as document review, evidence synthesis, information retrieval and report assessment increasingly cost-effective. For example GPT-4 launched in 2023 at roughly \$30 per million input tokens, whereas models offering similar performance became available for around \$0.15-\$2.50 per million tokens by 2024-2026 (<https://tokencost.app/blog/ai-price-index>)

2. examples and tips for how reviewers could use AI to support each of the specific tasks related to their reporting template sections, and
3. details on the AI tools that we recommended reviewers use.

On the AI tool side, we gave reviewers access to an OpenAI ChatGPT Team licence which allowed us to assign and manage access to a secure version of ChatGPT. Unlike the free and individual-based ChatGPT licences, the tool ensures that information shared with the AI is more protected and not used for training the OpenAI AI models.

Progress and feedback given:

An email was sent to reviewers sharing the AI guidance and the option of using the AI tool. Reviewers were still able to use the guidance if they did not use the shared tool (assuming they had access to a secure AI tool themselves). Only two reviewers signed up for using the AI tool from which we got guidance and tool feedback. This is in part due to the time of year that the guidance was developed and shared with most final report reviews having already been completed. It may also reflect that a lot of our reviewers have been completing reviews for a number of years and are less inclined to change their process. One reviewer was very positive about both the guidance and tool, saying that the guidance aligned with their approach to using AI. The other reviewer was also positive about the guidance but struggled to effectively use the AI tool. This highlighted how more focused training might be needed for reviewers newer to using AI. We did not get feedback from reviewers who did not use the tool and might have used the guidance.

In relation to value for money (VfM) and level of effort (LoE), internal testing by the NIRAS team did not indicate significant immediate time savings from using AI to support final report reviews. This is because the tool was deliberately used as an assistant within the existing review process, rather than as a replacement for reviewer judgement. Reviewers still needed to read the report, make their own assessment, check AI-generated outputs against the source material, and ensure that final comments reflected the fund's guidance and standards. As a result, any time saved on procedural tasks, such as summarising sections, checking whether specific issues were addressed, or drafting initial wording, was partly offset by the time required to verify and refine the AI outputs.

However, the testing did indicate that AI can improve the quality and consistency of reviews when used appropriately. In particular, AI helped reviewers to cross-check whether they had considered relevant evidence, identify points that may have been missed, test the clarity of draft comments, and support more structured responses to sections that require judgement across several criteria, such as GEDSI, standard indicators, lessons learned, scalability and durability, and value for money. These benefits are important because the review process is not only intended to produce a completed template within the allocated time, but also to ensure that reviews are sufficiently robust, evidence-based and useful for fund management decisions.

The VfM case for AI in report reviewing should therefore be understood as a combination of efficiency, quality assurance and risk reduction, rather than time saving alone. In the short term, the main benefit is likely to be improved review quality within the existing one-day LoE, helping reviewers manage expanded reporting requirements without reducing attention to important areas of analysis.

Recommended next steps:

- Share the guidance across all review types. Some slight tailoring would be required for the different review type templates.
- Conduct a focused AI survey to gauge the extent to which reviewers are using AI already, the types of tools they use, and feedback on the guidance. Some anecdotal insights have been gathered from a handful of reviewers but little is known about how reviewers are using AI and what their needs are.

- Drawing from the survey above, the BCFs management team could then gauge the extent to which reviewers can use existing secure tools they might have access to already or whether to make the tools accessible e.g. through ChatGPT Team licence. The ChatGPT licences are transferrable and relatively low cost, so it might be possible to just keep a few licences that reviewers only use when needed. All NIRAS staff already have access to a secure AI tool (Microsoft Copilot) so would only be necessary for independent consultant reviewers.
- In addition to general-purpose tools, such as ChatGPT Team or Microsoft Copilot, the BCFs team should explore making specialist AI tools available to support specific review questions and recurring workflow steps, often described as “agentic” AI tools. These tools can reduce the need for reviewers to write detailed prompts and can lower the level of AI experience needed, because the tool guides the user through a structured task. However, Defra should consider the full cost of adoption, not only the software licence. Costs are likely to include secure tool licences for external reviewers where they do not already have approved access, staff or consultant time to develop and test BCFs-specific prompts and workflows, preparation of guidance and worked examples, training sessions for reviewers and BCFs staff, and ongoing support to answer user questions during review periods. Specialist tools may also require developer time to configure, update and maintain them, particularly where BCFs templates, assessment criteria or data sources change. Investment should therefore be prioritised where the same task is repeated many times, where there is a clear quality assurance benefit, and where the cost of development and training can be spread across multiple review rounds or use-cases.
- For general-purpose AI use, the main investment need is likely to be light-touch and practical: ensuring users have access to secure tools, providing short task-based training, and maintaining clear guidance on what reviewers can and cannot do with AI. This is likely to be appropriate for tasks where reviewers need flexible support, for example summarising report sections, checking whether evidence has been addressed, drafting clearer comments, or testing whether their assessment is sufficiently evidenced. For specialist AI tools, the investment case is stronger where the task is more standardised, high-volume or strategically important, such as ICF screening, portfolio evidence synthesis, or structured quality assurance of completed reviews. These tools may have higher upfront development costs, but can offer better consistency, lower user burden and clearer auditability once established.
- Explore using AI to support the quality assurance of completed reviews. AI could be a useful assistant to existing quality assurance measures.
- Over time, as reviewers become more confident using AI, guidance is tailored to different report types, and more structured or specialist tools are introduced, there may be greater scope for measurable time savings. Future rollout should therefore track both time saved and quality improvements, for example by comparing review completion time, number of QA comments, reviewer confidence, and the completeness of evidence-based assessments before and after AI-supported review processes are adopted.

2.3 An Evidence synthesis AI tool - drawing from findings and lessons across BCFs funded projects

Purpose

We developed an AI tool that could categorise and summarise qualitative information and evidence from across BCFs project reporting documents.

Reason for selecting the use-case and expected return on investment

The BCFs team often need to synthesise qualitative information from across project documents, for example lessons learnt on a specific topic drawing from project annual and final reports. The reviews can cover over one hundred documents making it very time consuming. At the time of this deep dive, off-the-shelf general-purpose AI tools like ChatGPT/Copilot were not well suited for this task because it was not possible to upload more than a few documents at a time. At this time, it was felt that a specialised AI tool was required that could parse through

documents without needing to manually upload them. It was also important to support traceability and replicability so that similar conclusions/lessons could be developed if the same question was asked again.

Status and findings

For the use-case pilot we focused on creating a tool that could summarise lessons learned from across project reporting documents. Figure 1 shows how the tool worked and the main steps that were used to create the evidence summaries. The project reporting templates include a lessons learned section from which the tool was able to select this text and then classify it using a combination of AI algorithms using generative AI and traditional machine learning classification methods. The output results were then viewable in Excel and Word documents. The tool can show overarching lessons and sub-lessons. The number of listed themes can also be pre-specified.

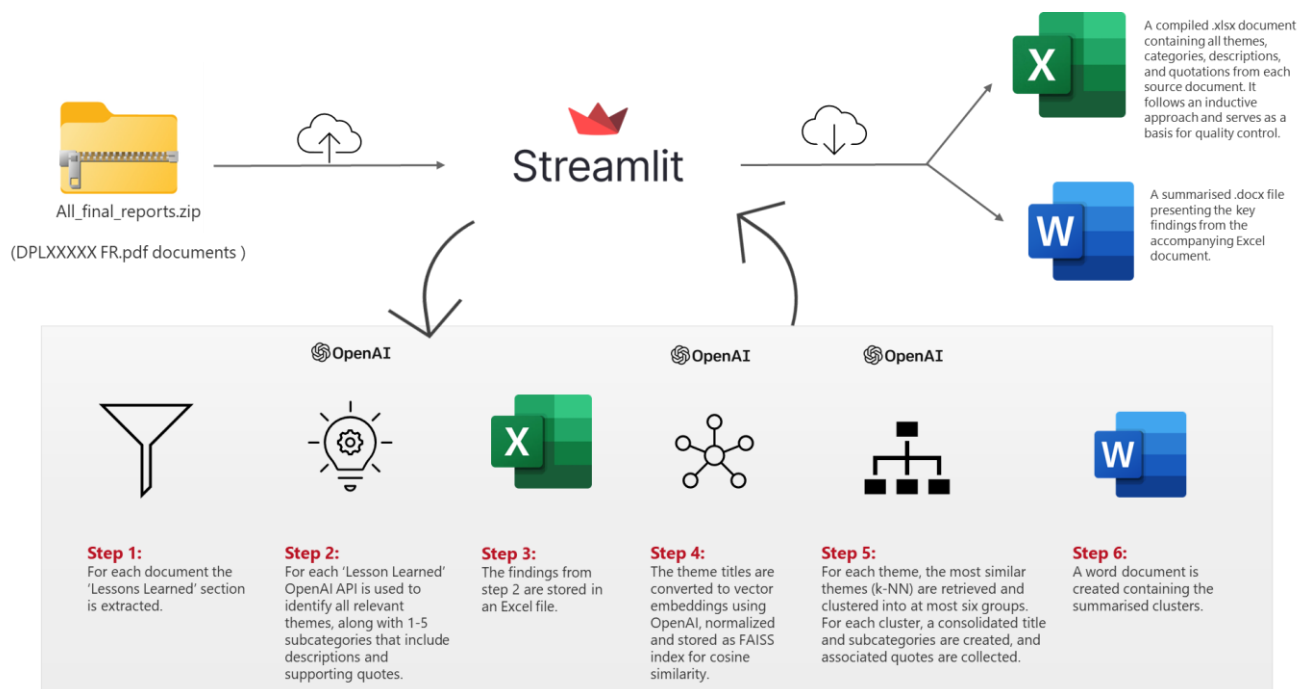


Figure 1: Overview of the key steps used to generate summaries of lessons learned from project documents.

We tested the accuracy of the tool by comparing the lessons it generated with that of a review that had been done manually. We were able to replicate the findings with the tool, signalling that it could provide the accuracy needed for future assessments.

Recommended next steps

The tool is operational and can be used for lesson learning synthesis. Incremental improvements could be made to the tool over time based on staff feedback. With a small amount of developer time, the tool could also be modified for other evidence synthesis tasks across BCFs such as collating lessons learned from Mid-Term Reviews (MTRs).

2.4 International Climate Fund Rio Marker screening tool

Purpose and background

International Climate Fund (ICF) assessments involve scoring projects at application stage against four criteria (called "Rio Markers"), assessing the extent to which they are considered in the project implementation approach

and objectives. For every project, each criteria – Biodiversity, Climate Adaptation, Climate Mitigation, and Desertification – is categorised as either a principal objective, significant objective, or not targeted. The scoring is currently done by BCFs staff and detailed guidance has been written for the staff reviewers, defining the criteria and scoring requirements. A tool was developed during the previous deep dive that could score project applications against the criteria. The updated tool included improved capability to explain the reasoning for the scores given and the specific parts (pages) of the proposal that supported the rating given. This would allow the reviewers to more easily compare and cross-check the scoring they assigned. This tool is already showing value for money.

Reason for selecting the use-case and expected return on investment

Recent adjustments to the guidance for ICF assessments have meant that the BCFs team need to draw on a wide range of potential climate-focused benefits to categorise projects as ICF eligible. To provide Defra with the best possible information to inform their budgetary allocation, speeding up this process would better focus NIRAS efforts on QA, keeping review costs at a minimum and streamlining the process to provide Defra information as soon as possible (carrying out ICF assessments is currently sequenced later in the process to keep costs down).

Status and findings

Figure 2 outlines the steps used by the updated ICF screening tool. The tool's ability to reference specific page numbers greatly improved the transparency and robustness of the tool, as well as traceability and replicability. It is best used as an assistant to augment and quality assure decisions made by reviewers. This added functionality improves the ability of the tool to assist reviewers. The tool could easily be modified to assist with other screening and due diligence tasks.

The tool was tested by two members of the BCFs team, one of whom had done ICF reviews. The tool is fully operational, noting that ongoing accuracy assessments and adjustments will always be required.

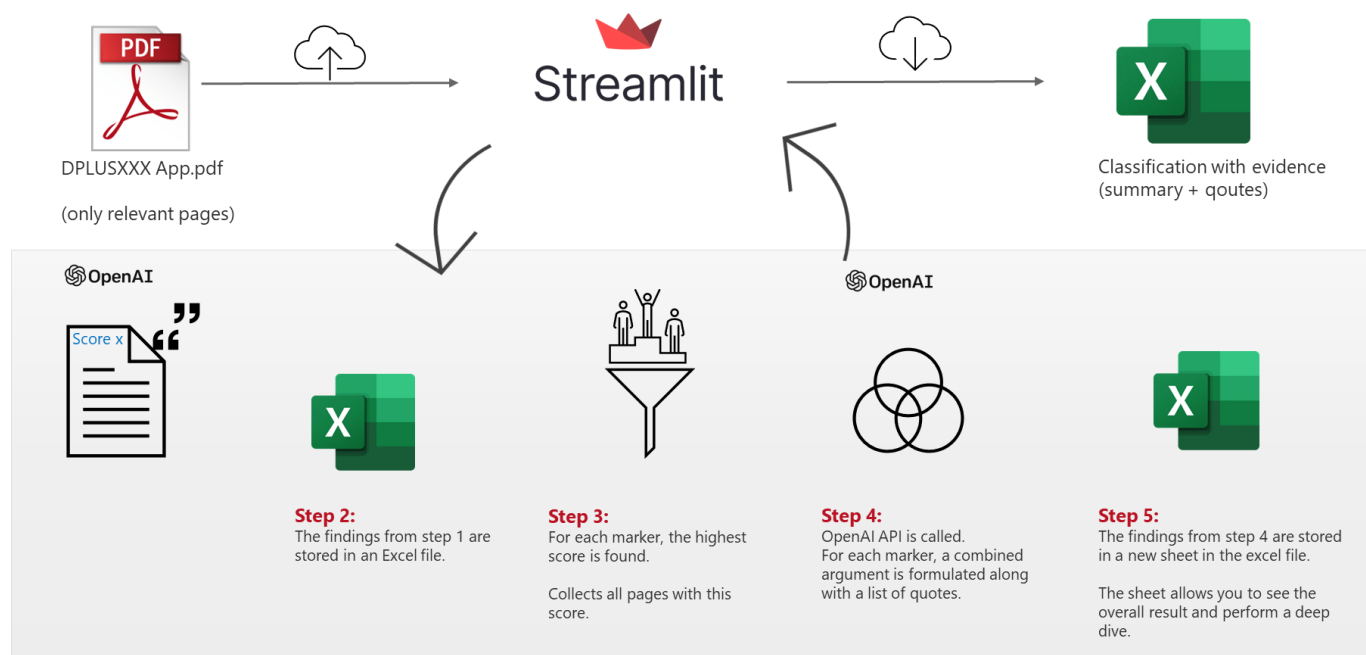


Figure 2: Key steps used by the tool to classify application documents against the ICF Rio markers.

Recommended next steps

The tool is fully operational and ready to be used by the NIRAS team. As with all AI tools, we would only recommend using the tool as an assistant cross-checking and supporting the ICF screening process. The time saving of this will be tested in the next round of applications.

2.5 Summary of key challenges and opportunities for full AI operationalisation

Despite identifying strong opportunities and successful piloting of AI use-cases (in deep dive one and two respectively), take-up and full-integration of AI into BCFs grant management processes has been slow. Below we list the main challenges that have caused this:

- **Staff time and headspace can be a major barrier for uptake of AI.** As successful grant management relies on highly optimised workflows, it can be difficult to find the time to learn new ways of doing work. We found that this applied to both types of tools that we used for the deep dive: the general-purpose AI tools (e.g. ChatGPT Team to support Final Report reviewers) and specialised (agentic) tools that can help with specific tasks (e.g. evidence synthesis and ICF tools). Specialised tools take time to develop and tailor before becoming useful but offer benefits when operational in terms of the skill required to operate the tools (e.g. not needing to write good prompts), unlike general-purpose tools which require more training and skill to utilise properly. The BCFs team will need to navigate this, deciding when to make specialised tools versus use general purpose tools that require more training.
- **AI risks make adoption hard in grant management contexts where there is no margin for errors and fairness implications.** Using AI comes with significant risks (e.g. accuracy, fairness, environmental impacts) which slows how quickly the team can adopt AI. Our approach in the deep dive was to only use AI as an assistant, providing supplementary support. The challenge with this is that the marginal gains from using AI can be small and combined with the time required to learn and adopt the technology, it can make it difficult for staff to justify fully adopting the technology.

- **Getting adoption and uptake from external team members is challenging.** The BCFs team relies heavily on a team of external reviewers and experts. It can be challenging to give external staff access to internal IT tools (like Copilot 365) and support training and uptake. There are also additional cost implications for giving access to secure tools.
- **Gains from AI are sometimes only marginal and not time-savings related.** We have found that AI can save time but most of the gains were improvement in the quality of grant management work. This was partly because we only used AI as an assistant, augmenting existing workflow practices. Marginal gains can reduce enthusiasm for using the technology, especially when the expectations are high.
- **Linking AI to internal documents and data offers huge opportunities but requires having an established database (folder library)** and only linking relevant information (avoiding 'garbage in, garbage out' related issues).
- **The grant management team, leadership, and staff need to have ownership.** Our deep dive involved an external team of AI experts leading on the development and uptake of the tools, working closely with the BCFs team as the users. With hindsight this might have reduced ownership of the outputs of the deep dive affecting adoption of the tools by the team. It is key that the BCFs leadership determine the strategy and have sufficient time and budget to drive uptake, aligned with their vision and values.

Over the last two years since the first phase of the deep dive, the landscape for operationalising AI within BCFs grant management has improved considerably. The capabilities of modern AI systems, coupled with more accessible tools, create strong opportunities to embed AI safely and effectively across BCFs workflows. These developments support the transition from pilot testing to widespread operational use. Below we list some of the key opportunities:

- **Marked improvements in model performance, reasoning, and reliability.** AI systems have improved in two key areas that directly address earlier constraints identified in the deep dive studies.
 1. Models can now process significantly larger volumes of information in a single interaction, enabling them to review entire application files, annual reports, and evidence sets without needing extensive prompt management by staff or significant costs/resources associated with searching large number of documents.
 2. New reasoning-focused models show stronger performance on tasks involving structured judgement, multi-step logic and synthesising diverse data sources.
- **Expanded ability to customise and deploy specialised AI tools.** Modern platforms allow development of tailored interfaces that guide staff through structured workflows, including application screening, progress review tasks, and portfolio information retrieval. These tools can integrate BCFs rules, templates, and domain knowledge directly into the system, which helps ensure alignment with fund requirements and reduces the burden on reviewers. Importantly, richer monitoring and control features now exist, allowing administrators to track how AI tools are used, conduct quality assurance, and ensure that outputs remain consistent with ethical and responsible use principles.
- **Stronger alignment with operational needs and quality standards.** Models are now better at following instructions precisely, producing consistent outputs that match required formats and applying BCFs review criteria without deviating from guidance. This reduces the risk of errors creeping into outputs and enhances the potential for efficiency gains, as reviewers can spend more time on expert judgement and less on extracting, formatting, or compiling content.
- **Potential to integrate AI across end-to-end workflows.** Because current AI systems can manage more complex, multi-step tasks, there is now realistic potential to operationalise AI across an entire workflow rather than at isolated points. For example, a single system could support reviewers to extract key information from reports, cross-reference this with evidence from previous reporting cycles, highlight areas requiring scrutiny and draft initial comments that reviewers refine. Similarly, in application screening, AI can assist with eligibility checks, summarise key strengths and risks, and pre-populate scoring matrices. This end-to-end capability was not feasible two years ago due to model limitations and the lack of stable agent-like tools.

2.6 Recommendations

Drawing from the above, we recommend the following:

- The BCFs team and leadership should continue to lead the uptake of AI, when and how they want to use AI. They can draw on AI and IT experts where needed (e.g. building advanced customised tools) but the bulk of the work would be led by them. The uptake work should be run like a project e.g. with designated coordinator, goals and tracking of progress and designated budget. Surveys of current use and uptake, especially by external reviewers, would be helpful in informing strategy.
- Linked to the above point, develop principles for AI usage that reflect the values and approach of the BCFs team and leadership. The principles could be made public and apply to the team and grantees. It could be tailored from the NIRAS development consulting principles (see Annex 2).
- Ensure all staff have suitable access to secure general-purpose AI tools. In the case of internal staff, this would include Copilot 365. For external staff, a list of approved (secure) AI tools could be shared.
- Continue with the use and uptake of the specialised agentic AI tools that were developed for this deep dive (for evidence synthesis, ICF reporting), including their ongoing refinement and maintenance. For new ideas for specialised tools, encourage staff to identify new use-cases (building on deep dive 1 work) and decide which areas to prioritise.
- Develop guidance and training for staff and external reviewers, plus spaces for staff to share and document use-case examples.

Annexes

Annex 1: AI guidance to reviewers: Final Report Reviews

Thank you for taking part in piloting the use of this guidance for your Final Report Reviews. After completing your review, we will share a short survey to gauge how useful the guidance was and the AI tool and use-cases that we recommended.

1.1 Purpose of this guidance

AI has high potential to improve the depth and quality of our BCFs final report reviews, but its use also comes with significant risks if misused. The purpose of this guidance is to help you use AI effectively and responsibly.

Because AI is a general-purpose technology the guidance is not prescriptive about where and how you use AI, you will need to develop your own bespoke way of using AI.

The guidance covers three main topics: (1) best practices and guardrails that should be applied across tasks, (2) the AI tool you should use (ChatGPT Team), and (3) tips for how you could use AI to support each of the specific sections in your Final Report review.

1.2 AI best practices and guardrails that apply across tasks

Here are some cross-cutting best practices that should be applied when using AI. These best practices will help you get the most out of AI and avoid potential pitfalls.

- Use AI as an assistant, augmenting not replacing your review tasks: you need to always be in control and have oversight over the review process and what the AI does. Delegating too much to the AI can detach you from the review process resulting in significant risks. For example, if you want to use AI to help you summarise something, first read the material yourself and make your own high-level summary before asking the AI to generate a summary that can augment the summary you created. We have seen reviewers overly relying on AI in some cases, resulting in sub-standard reviews.
- Always check the outputs AI generates linked to the above always fact-check AI outputs, AI cannot be trusted, it can make-up (hallucinate) outputs.
- Give clear and manageable instructions to the AI: giving good instructions (also called prompts) to the AI is very similar to how you would instruct a human. Top tips include (a) being specific about what you ask for, (b) explain the context to the AI, (c) break complex requests into smaller manageable requests and (d) treat prompting as a conversation, refine your requests based on initial answers, don't worry about getting the perfect prompt. Also see [quick guidance here](#) from OpenAI. The guidance given in the Final Report review guidance (blue text) can be a useful starting point for giving instructions to the AI, however we recommend being even more granular with your instructions to get better quality outputs.
- Always read the Final Report before using AI: there is a risk with AI that you don't read the Final Report, instead relying on the AI's interpretation of the Final Report through the summaries and information it shares with you.
- Ask the AI to support its claims with references to specific information written in the Final Report This can cross-check what the AI shares with you by reading the original Final Report text.

- Be specific about what you want the AI to look at when answering instructions about the Final Report you are reviewing: For some requests you might ask the AI to draw on the full Final Report (which you can attach to your conversation). For most other tasks however, you will get better results if you copy and paste the information that is relevant to the instruction that you give the AI. It is also important to tell the AI to only draw on the information you share with it about the project. This reduces the risk of the AI drawing on its own knowledge that it might have (or hallucinate) about the project.
- Use secure AI tools: please only use the AI tools that we recommend (currently ChatGPT team see below). Some tools (especially free ones) are not secure, the data you enter can be leaked.

1.3 The tool we would like you to use: ChatGPT Team

We will share a ChatGPT Team licence with you to use for your review. You will need to login using the email address and password that we share with you.

ChatGPT is a conversational AI tool that can answer questions based on instructions (prompts) you give. ChatGPT Team works the same as ChatGPT free version but has added security and access to better AI models.

1.4 Tips for using AI for specific sections of your final report review

Below are some tips for using AI for the specific sections of your final report review. The tips draw from the above guidance. The sections below align with the Final Report for Darwin Initiative Main and Extra templates, if you are doing a review for another Fund then use the relevant sections.

1. **Project Summary:** We recommend first creating your own summary in bullet form then ask the AI cross-check your summary points and make suggestions for things you might have missed. Then ask the AI to draft your points into full sentences and paragraphs. Make sure to review the AI drafted text for factual errors. We would recommend attaching the full final report to your conversation (instead of copying and pasting) and tell the AI to only draw on information from the attached document.
2. **Project Outputs:** First work through answering the guidance sub-questions on your own then use AI to cross-check and augment your responses. Ask the AI to check your responses at the sub-question level to get more specific responses. Where appropriate be specific about which parts of the Final Report (and Annexes) you want the AI to draw on.
3. **Project Outcome:** For this section the reviewer needs to assess whether outcomes have been met and the quality of supporting evidence. We recommend not asking the AI to give the overall score (e.g. met, exceeded) but rather asking AI to give reasons that might support and non-support giving a met/not-met rating. This is also best done after you have made your own assessment. Make sure you clearly articulate what the outcome(s) were when instructing the AI. For the quality of evidence assessment, you could use the table information in your instructions to help guide the AI.
4. **Project achievements and Impact** As above (2 and 3), use AI to cross-check and augment your assessment, don't use AI to make the overall assessment. You can ask the AI to list reasons and supporting evidence that the project is meeting / not meeting its impact and then combined with your own reasons make an overall assessment. AI can also be useful in drafting your points into full sentences and editing your text.
5. **Contribution to Darwin Initiative Standard Indicators** We would recommend making your own assessment first then asking AI if there are additional indicators that could be used to report against the standard indicators. Ask the AI to give reasons for why an indicator is relevant for a standard indicator so that you can assess how

credible the suggestion is. Into your conversation you will attach the standard indicator list and supporting information.

6. **Project support for multidimensional poverty reduction** Do your own assessment first then cross-check your assessment with what the AI produces. The assessment involves several sub-questions / criteria that you need to consider, use the AI at this level (i.e. don't ask a general question like does the project support poverty reduction).
7. **Contribution to Gender Equality and Social Inclusion:** First make your own assessment for the overall score and then sub-questions that are given in the guidance notes. You can then ask the AI to give its assessment to cross-check what you wrote. Ask the AI to give reasons for its assessment. Give the AI access to the supporting BCFs GESI information guides.
8. **Scalability and Durability** This section includes several sub-questions that you need to answer as a reviewer. Use AI at this level to cross-check and add to your assessment for each sub-question. Ask the AI to give evidence and examples from the Final Report.
9. **Early indicators of transformational change** Share the criteria given in the guidance and ask the AI to find evidence that the project addresses each of the criteria. Do this after you have made your own initial assessment.
10. **What lessons learned/failures/challenges from this project could be used to improve future projects or wider initiatives** First read the relevant sections of the Final Report and make bulleted notes of your assessment then cross-check this with the AI assessment. Ask the AI to reference examples from the Final Report. Work with the AI through each sub-question to make the assessment more granular.
11. **Risk management:** AI can be used to highlight where all risks are explicitly mentioned in the reports and help the reviewer to identify potential risks that could be flagged which were not explicitly listed. As above, work with the AI through each sub-question to make the assessment more granular.
12. **Safeguarding** AI not relevant here.
13. **Project expenditure** AI not relevant but could be used to cross-check your answer for the last sub-question (*"Does the project appear to be good value for money? If not, what could have improved this?"*)
14. **Key Facts for Defra Publicity** The AI could be useful in checking the newsworthy key facts that could be useful to share with Defra. Make your own initial assessment first and then cross-check with the AI. AI can also help with drafting the text after you have decided on the key facts you want to flag.

Annex 2: NIRAS Principles for responsible AI use

1. **AI is a support tool not a substitute** – We lead and take full responsibility for all outputs; AI is used only to assist and improve quality, not as a substitute for the work you would have done. Avoid producing AI slop: generic and low-quality outputs.
2. **Validation is non-negotiable** – All AI-generated content is thoroughly reviewed, fact-checked, and evaluated by relevant human experts. All project and bid deliverables should be quality assured with four eyes, as per our existing quality assurance policy within NIRAS
3. **Transparency is key** – We clearly document and communicate how and where AI was used during each assignment with our relevant team members.
4. **Stakeholder and client consent** – Before each assignment, we propose our intended use of AI, the AI tools we will use and seek approval from relevant NIRAS team members and the client².
5. **Tool selection is rigorous** – We choose only vetted, secure, and fit-for-purpose tools based in consultation with relevant technical experts at NIRAS.
6. **No sensitive data shared/uploaded to non-secure AI tools** – Under no circumstances will NIRAS or client information be input into AI platforms that are not secure and compliant. Tools like ChatGPT Free are not compliant.
7. **Minimise environmental impacts** – We try to minimise the negative environmental impacts of AI tools, taking steps to reduce environmental impacts through the tools we select and using the tools efficiently.
8. **Prevent risks of biased AI content** – AI tools should not generate unfair bias or lead to exclusion of different social groups, especially those that are already marginalised. Maintain principles of non-discrimination and human rights.

² For the BCFs the client is Defra