INNOVATION & RESEARCH CAUCUS

USING AI IN EVALUATION: EVIDENCE, PRACTICE AND PRINCIPLES FOR RESPONSIBLE USE

IRC Report No. 093

REPORT PREPARED BY

Hamisu Salihu
University of Warwick

Ully Nafizah
University of Warwick

Halima Jibril
University of Warwick

Delivered with ESRC and Innovate UK

Authors

The core members of the research team for this project were as follows:

- Dr Hamisu Salihu - University of Warwick
- Dr Ully Nafizah - University of Warwick
- Dr Halima Jibril - University of Warwick

This document relates to IRC Project IRCP0034: The use of Artificial Intelligence in evaluation: A review of the evidence base

Acknowledgements

This work was supported by Economic and Social Research Council (ESRC) grant ES/X010759/1 to the Innovation and Research Caucus (IRC) and was commissioned by the Engineering and Physical Sciences Research Council (EPSRC). We are very grateful to the project sponsors at UK Research and Innovation (UKRI) for their input into this research. The interpretations and opinions within this report are those of the authors and may not reflect the policy positions of EPSRC.

We would like to express our sincere gratitude to all research participants for their time, openness, and valuable contributions. Their insights made this work possible. All views, interpretations, and conclusions presented in this report are the authors' own and do not necessarily reflect those of the participants or any affiliated organisations.

We would also like to thank the Project Working Group for their input and particularly the project manager Cameron Ross for his very helpful contributions to the study.

We would also like to acknowledge and appreciate the efforts of the IRC Project Administration Team for their support in preparing this report.

About the Innovation and Research Caucus

The Innovation and Research Caucus supports the use of robust evidence and insights in UKRI's strategies and investments, as well as undertaking a co-produced programme of research. Our members are leading academics from across the social sciences, other disciplines and sectors, who are engaged in different aspects of innovation and research systems. We connect academic experts, UKRI, IUK and the ESRC, by providing research insights to inform policy and practice. Professor Tim Vorley and Professor Stephen Roper are Co-Directors. The IRC is funded by UKRI via the ESRC and Innovate UK, grant number ES/X010759/1. The support of the funders is acknowledged. The views expressed in this piece are those of the authors and do not necessarily represent those of the funders.

Contact

You are also welcome to email us if you have any questions about this report or the work of the IRC generally: info@ircaucus.ac.uk

Cite as: Salihu, H., Nafizah, U. & Jibril, H. 2026. *Using AI in Evaluation: Evidence, Practice and Principles for Responsible Use*. Oxford, UK: Innovation and Research Caucus.

CONTENTS

EXECUTIVE SUMMARY 5
Introduction 10
Methods and Evidence Base 12
2.1 Phase 1: rapid evidence review 12
2.1.1 AI-assisted literature discovery exercise 12
2.1.2 Flexible rapid literature review approach 13
2.1.3 Scope of the evidence review 13
2.2 Phase 2: Qualitative Interviews Approach 14
2.3 The funding evaluation cycle as an organising framework 15
2.4 Definitions of AI tools 16
Findings from the Evidence Review: AI Use Across the Funding Evaluation Cycle 18
3.1 Choosing intervention areas (project prioritisation) 18
3.1.2 Implications for UKRI 18
3.2 Assessing project proposals (grant assessment) 19
3.2.1 Implications for UKRI 19
3.3 Project implementation, monitoring and process evaluation (M&E) 19
3.3.1 Implications for UKRI 20
3.4 Impact evaluation and value for money evaluation (M&E) 20
3.4.1 AI use in quantitative impact analysis. 20
3.4.2 AI use in qualitative impact analysis. 20
3.4.3 AI use in summaries, syntheses and evaluation reports 21
3.4.4 Implications for UKRI 22
Findings from qualitative interviews: How evaluators are using AI in practice 24
4.1 AI awareness and patterns of use among evaluators 24
4.1.1. Evaluators' understanding of AI 24
4.1.2 A mixed landscape of adoption 25
4.1.3 AI use across the evaluation process 26
4.2. Findings from thematic analysis 27
4.2.1. The human-in-the-loop consensus 27
4.2.2 Qualitative coding and synthesis as a current strength 28
4.2.3 The scale threshold 29
4.2.4 Perceptions of benefits, limitations and views on output quality 30
Risks, limitations and mitigation conditions 32
5.1 Structural risks 32
5.1.1 Equity and bias 32
5.1.2 Data privacy, confidentiality and security 32

5.1.3 Legal and ethical frameworks 33
5.1.4 Transparency and accountability 33
5.2 Operational risks 34
5.2.1 Scientific rigour, validity and reliability 34
5.2.2 Overreliance and skills erosion 34
5.2.3 Stilted outputs and limited interpretive value 34
6. Using AI responsibly in evaluation: Emerging Best Practices 36
7. Principles to guide AI use in evaluation 39
7.1. The case for a principles-based approach 39
7.2. 10 principles for AI use on evaluation 39
7.3. Key responsibilities of UKRI as evaluation commissioner 41
7.3.1. Governance and oversight: What UKRI should consider 41
7.3.2. Skills and learning- UKRI should consider: 41
8. Conclusion and Implications for UKRI 42
8.1 Triangulation of evidence 42
8.2. Gaps in the evidence base 43
REFERENCES 47
APPENDIX 58
Appendix A: Evaluator Skills and Attitudes for Effective Engagement with AI 58
Appendix B: Additional Guide for the Responsible use of AI in an Evaluation 59
Appendix C: Tool-specific Considerations: Challenges and Good Practice 61
Appendix D: AI Maturity Level Classification 62
Appendix E: Potential uses of AI Across the Funding Evaluation Cycle 63

EXECUTIVE SUMMARY

This report examines the use of Artificial Intelligence (AI) tools in policy evaluation. It brings together two phases of work conducted for UKRI. Phase 1 was a rapid review of the evidence base, examining current and potential applications of AI globally and across different types of funding organisations and sectors. Phase 2 complements this with primary qualitative research, drawing on semi-structured interviews with evaluators to understand the extent and nature of everyday adoption of AI in evaluation. Together, the two phases aim to provide UKRI with evidence-based insights to inform appropriate principles for the responsible and effective integration of AI into evaluation processes.

The report draws on academic and grey literature as well as on interview evidence, and develops a funding evaluation cycle as a framework for understanding AI use in: i) choosing and prioritising intervention areas; ii) assessing project proposals; iii) programme monitoring and process evaluations; and iv) impact evaluation and value for money assessments.

It is worth noting here, however, that throughout this report, "evaluation" is used broadly to cover assessment activity across the whole funding cycle, including the appraisal of grant proposals. This is wider than the sense in which UKRI normally uses the term internally, where evaluation refers to the monitoring and evaluation (M&E) of funded programmes. Where findings relate specifically to grant assessment, this is stated.

Use cases and reported benefits of AI in evaluation (Section 3)

The evidence review found notable examples of AI applications across different phases of the funding evaluation cycle (Section 3). Here we highlight five areas where literature review evidence on AI effectiveness is relatively stronger, as well as areas where its use may be less effective or too risky.

1. International evidence shows that Natural Language Processing tools can effectively support horizon scanning and strategic agenda setting, which rely on analysing large volumes of real-time data. This could help UKRI identify emerging funding priorities.

2. LLMs, ML, and Generative AI can increase efficiency in the administrative stages of proposal assessment, such as pre-screening and classifying applications, thereby reducing administrative burden. However, AI should not be used beyond these stages for peer review or final funding decisions.

3. LLMs can provide efficient and reliable summaries of large documents, supporting evaluators in preparing reports. They are much less reliable for evidence synthesis, which continues to require significant human input.

4. ML tools are effective for real-time monitoring and data collection, though this may be less applicable to UKRI programmes where recipient organisations, such as universities or businesses, operate with a high degree of autonomy in grant use and real-time monitoring may be infeasible.

5. LLMs, ML and GenAI tools have performed well in quantitative and qualitative data analysis, including code standardisation and replication, provided that strong data management, security, and governance measures are in place.

AI use in Practice (Section 4)

The Phase 2 interviews complement these findings and shed a light on how AI is currently used by evaluators and their perceptions of its risks and benefits. Key findings from the interviews are:

- There is a striking variation in the maturity of AI application across the organisations in which respondents work, ranging from beginners to those with comprehensive in-house systems and sophisticated, end-to-end approaches.

- Evaluators with the most advanced and widespread use of AI also perceive higher quality of AI- assisted outputs and have the most transparent disclosure norms.

- There is widespread application of diverse AI tools for various evaluation related tasks. The tools most frequently cited by respondents were Microsoft Copilot, ChatGPT, Claude, Gemini, Fireflies, Google NotebookLM and, in a few cases, bespoke in-house tools.

- Most evaluators interviewed adopt AI for interview transcription, qualitative coding and analysis. This is seen to represent AI's current strength in evaluation, judged to have high efficiency and quality. Adoption is least prevalent for quantitative analysis.

- The synthesis of large volumes of text and project records is also increasingly treated as AI's current strength. This is contrary to what the evidence review suggests. One possible explanation is timing, as the interviews were conducted more recently than experiments conducted in prior studies. Given the rapid pace of development in AI models, it is possible that text synthesis capabilities improved substantially over time. This points to a wider pattern in which evaluator practice is outpacing the published evidence base, particularly for qualitative synthesis and large-scale document analysis. For UKRI, it reinforces that guidance should not be too prescriptive and should be principle-based, reviewed regularly, and informed by UKRI's own evidence of emerging practice...

- AI use is also prevalent in proposal and report writing stages, mostly assisting with ideation, editing and structuring rather than the automated generation of text.

- Evaluators apply a pragmatic volume threshold when deciding whether to use AI, reflecting considered thinking about AI's comparative advantage in analysing large volumes of data at speed.

- The interviews reveal a clear consensus by evaluators against AI-only approaches, instead favouring having a 'human-in-the loop' approach.

- Respondents consistently cite AI's principal benefits as speed, the ability to work at a scale or depth that manual methods cannot reach, and supporting coding and synthesis.

Risks and challenges of using AI in evaluation (Section 5; mitigations in Section 6)

Using AI tools in policy evaluation carries significant risks and challenges (Jacob, 2024, 2025, OECD, 2025). This may relate to structural challenges such as potential biases, gaps in ethical and legal frameworks, difficulties ensuring data privacy and security, and issues with transparency and accountability. At the operational level, challenges include underperformance of some AI tools in maintaining scientific rigour, validity, and reliability, as well as the tendency of Generative AI to produce stilted outputs with low artistic value. There is also a risk that patterns of tool-user interaction may lead to overreliance on AI and erosion of evaluator skills.

Evaluators also had similar perceptions of AI limitations, specifically highlighting concerns around hallucination and accuracy, constraints arising from the volume and quality of available data, bias against protected groups, the erosion of analytical skills, the risk of over-reliance, and the difficulty of transparent disclosure.

Triangulation of evidence from literature review and evaluator interviews (Section 8)

The literature and interview evidence are broadly aligned on where AI currently adds value in evaluation, but there are also important differences between documented capabilities, current practice and the perception of risks.

- Both evidence strands point to the clearest benefits of AI use in qualitative coding, classification and analysis, with gains concentrated in speed, scale and efficiency rather than consistently higher quality.

- There is also strong agreement that human oversight remains essential, and strong concerns about over-reliance and skills erosion, particularly for junior staff

- Interviewees reported more sophisticated approaches to evidence synthesis than earlier studies would suggest, particularly where AI is embedded in structured, traceable workflows with human verification. This most likely reflects the pace of model improvement between the 2023 experiments and our 2026 interviews.

- Quantitative impact analysis is comparatively well evidenced in the literature yet barely used by the evaluators we interviewed, which points to a potential opportunity that UKRI evaluators are currently not exploiting.

- Specific concerns over bias and equity are prominent in the literature but are not front of mind in practice, which argues for treating them as a commissioner-level requirement rather than leaving them to practitioner discretion.

Evidence-based principles for AI use in evaluation (Section 7.2)

Triangulating the evidence from the literature reviews and interviews with evaluators, we develop 10 principles of responsible use of AI in evaluation (Section 7). These relate to the different roles of human judgement, disclosure, governance and the development and preservation of evaluation skills.

- AI should **support evaluator judgement**, not replace it.
- AI outputs should be **checked and verified**.
- AI-assisted analysis should be **traceable back to the evidence**.
- LLMs should be used through an **iterative prompt-and-validate process**.
- AI should be used where it is **appropriate to the task**
- AI use should be **proportionate to the data and evaluation context**.
- AI use should be **transparent and disclosed**.
- AI use should follow clear **data governance and security rules**.
- AI use should meet relevant **ethical, legal and professional obligations**.
- Organisations should **invest in the skills required for responsible AI use**.

These evidence based principles are also consistent with emerging guidance from the evaluation practitioner community, including that of the UK Evaluation Society (UKES, 2025). In particular, the UKES guidance to evaluators emphasises transparency, accountability, human oversight, proportionate use of AI, active risk management and harm prevention, and quality assurance and verification. This alignment provides reassurance that the principles developed through our evidence review and evaluator interviews reflect wider thinking on the responsible use of AI in evaluation.

How can UKRI create the conditions for safe, transparent and responsible use of AI in evaluation? (Section 7.3 and Section 8)

This report recommends a measured and evidence-based approach to any significant integration of AI into UKRI evaluation activities. As a commissioner of evaluation, UKRI's role primarily relates to governance, oversight, encouraging the preservation of evaluation skills, and upholding key principles. We recommend that UKRI consider:

- Defining a clear but adaptable framework outlining appropriate and inappropriate use cases for AI within UKRI evaluations.

- Adopting an open but experimental approach through carefully designed pilots that test specific important applications, capture challenges and generate lessons.
- Ensuring data security and, where feasible, developing UKRI-specific internal AI systems and platforms so that sensitive data is not exposed to third-party tools.
- Providing clear accountability structures for AI-supported outputs, including for unintended errors or harms.
- Ensuring strong human oversight and ethical safeguards so that AI complements rather than replaces human judgement.
- Establishing clear disclosure norms that incentivise transparency and accurate reporting of AI use.
- Adopting a principles-base approach as outlined above, rather than overly prescriptive guidelines or checklists.
- Investing in AI literacy and skills training for evaluators, including the prompt-and-validate workflow and critical appraisal of outputs.
- Maintaining open communication with evaluators throughout the evaluation process and creating opportunities for learning and feedback loops.
- Periodically updating standards to reflect emerging best practices and new tools.
- Supporting the deliberate maintenance of core analytical skills, particularly for junior staff, to guard against skills erosion.

Gaps in the evidence base

Some gaps remain in the evidence base, which future work could usefully address. These include limited evidence on:

- Using AI to design evaluations
- Development of disclosure norms and how disclosure is received by evaluation commissioners
- Longer-term implications of AI adoption for the evaluation workforce and professional practice.

1. Introduction

Artificial intelligence (AI) refers to “a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments” (Russell et al., 2023). The advent of AI and the recent popularity of Generative AI tools is transforming the way in which tasks are performed and has been argued to hold huge potential for enhancing productivity across sectors (Al Naqbi et al., 2024).

Research and evaluation are domains where AI tools have specific potential as they have the ability to efficiently process and analyse large volumes of qualitative and quantitative data (Koliousis et al., 2024; Djunaedi, 2024). AI tools can also enhance the speed and quality of decision-making through automation, rapid data processing, and real-time analysis (Wirjo et al., 2022). By handling both structured and unstructured data at scale, AI can reduce time and workload while lowering costs of evaluations (Flahavan, 2024).

The UK's Research and Innovation agency (UKRI) seeks to understand how AI is being utilised in evaluation contexts by other funders globally and across sectors, and what this might imply for safe and effective AI use within UKRI's own evaluation processes. While there is growing interest in applying AI tools to routine evaluation tasks such as summarising large bodies of text, supporting data collection and analysis, and assisting with the write up of evaluation reports, concerns are often raised about transparency as well as ethical and legal safeguards when AI is introduced into evaluation workflows.

This report brings together two phases of work. Phase 1 was a rapid evidence review on the use of AI tools in policy evaluation. Phase 2 comprised semi-structured interviews with evaluators across consultancy, government and academic settings who are at different stages of adopting AI in their work. The two phases serve different and complementary purposes. The evidence review examines what the published and grey literature suggests about AI use, benefits, risks and emerging practices in evaluation. The interviews look into how evaluators are actually using AI in practice and what risks and benefits they percieve. The primary objective of the consolidated report is to provide UKRI with evidence-based insights to inform the development of appropriate principles and guidelines for the responsible and effective integration of AI into its evaluation processes.

To that end, the Phase 1 review addressed the following research questions:

1. In what ways is AI currently being applied in the impact evaluation of policies and programmes?
2. How is AI being used across other stages of the funding evaluation cycle, including identifying intervention areas for funding, assessing project proposals, programme monitoring and data collection, and impact analysis?

3. What are the key advantages, challenges, risks and trade-offs in using AI for evaluation, and how do these vary by use case and specific AI tools?
4. What are the emerging best practices in the use of AI for evaluation?

Phase 2 extended this enquiry into practice, exploring additional questions in relation to everyday evaluation practice:

5. Which AI tools are evaluators using, and for which tasks?
6. What are the benefits gained compared with manual methods?
7. What are the perceived challenges and risks encountered and how these are managed?
8. What are evaluator's attitudes towards disclosing the use of AI tools?

The report proposes a funding evaluation cycle, based on the policy-making cycle (Cairney, 2023; 2024), as an organising framework. This enables consideration of AI use at multiple related phases of the evaluation process: choosing and prioritising intervention areas; assessing project proposals; monitoring and process evaluation; and impact evaluation. The latter is the primary focus of the report, but earlier phases are also considered in order to give a fuller picture of the potential benefits and risks of AI use across evaluation-related activities. In other words, we deliberately adopt a broad usage of evaluation, since many of the AI tools, risks and safeguards discussed here cut across both grant assessment and M&E. This may thus be broader than UKRI's internal convention/practices. Hence, in practice, Sections 3.1 and 3.2 concern prioritisation and grant assessment, while Sections 3.3 and 3.4 concern evaluation in the narrower M&E sense that UKRI teams will recognise.

The remainder of the report is organised as follows. Section 2 sets out our methods, including the rapid evidence review approach, the qualitative interview design, the funding evaluation cycle as an organising framework, and definitions of AI tools. Section 3 presents findings from the rapid evidence review of use of AI across different stages of the funding evaluation cycle (highlighting both potential and documented use cases of specific AI tools both potential and documented use cases of specific AI tools. Section 4 reports findings from qualitative interviews (Phase 2) on how evaluators are using AI in practice. Section 5 discusses the key risks of using AI in evaluation, limitations and the conditions under which they can be mitigated. Section 6 outlines emerging responsible practice for AI use in evaluation. Section 7 synthesises the evidence into principles and practical guidance for UKRI and its evaluators. Section 8 concludes and sets out implications for UKRI.

2. Methods and Evidence Base

This report draws on two complementary sources of evidence. The first is a rapid evidence review of academic and grey literature. The second is a qualitative interview phase with professionals engaged in evaluation work. The two sources are used in different ways. The review provides the broader evidence base on documented AI use, while the interviews provide practice-based insight into how evaluators are currently interpreting, adopting and governing AI tools in everyday evaluation contexts.

2.1 Phase 1: rapid evidence review

We adopted a rapid evidence review approach (for example, Ganann et al., 2010; Varker et al., 2015) to provide timely insights into how artificial intelligence is currently being used in evaluation, and to capture the opportunities and challenges emerging in this fast-moving field. Unlike a full systematic review, the rapid review allows for a focused but flexible search and synthesis process, balancing breadth of coverage with timeliness.

2.1.1 AI-assisted literature discovery exercise

To engage with the review's topic area, we began with an AI-assisted literature discovery exercise. In particular, we used Elicit AI (Ought, 2023), a machine-assisted literature review platform that leverages language models to identify, extract and summarise relevant scholarly publications using semantic similarity techniques rather than purely keyword-based search engines. Our aim was to test and report, given the topic of this study, whether and how specialist AI tools can be useful in a research context relevant to many policy evaluation exercises (i.e., evidence reviews).

We asked Elicit AI to conduct a review of academic articles, using its mid-advanced function, based on the question “How is Artificial Intelligence being used to support policy evaluation?”. It claimed to have searched across over 126 million academic papers from the Semantic Scholar corpus and retrieved the 499 papers most relevant to the query, of which it retained 25 of the most relevant articles after screening for an explicit primary focus on a policy evaluation context with a defined policy domain. We conducted a similar search for the question “How is Artificial Intelligence being used to support research evaluation?”, obtaining similar outputs with 25 additional articles retained, making a total of 50 articles across both searches.

Although the 50 studies seemed relevant at first, only a few were ultimately included in this review: comparing our final reference list to that of Elicit AI returned 10 matches, representing about 8 per cent of our reference list. This is because, on further manual reviewing, many of the articles focused on the potential uses of AI and in policy domains outside that related to economic and business programme interventions, predominantly healthcare, electricity regulation and education. The Elicit AI reviews were nevertheless useful in helping us confirm our initial assessment that the academic scholarly literature

has not yet covered this topic, possibly reflecting long timelines to publication for peer-reviewed academic papers. It enabled us to turn more quickly towards grey literature to understand emerging use cases. Thus, by returning only a few relevant articles, Elicit AI's outputs enhanced the efficiency of our manual review process. This exercise is itself a small, reflexive example of the use of AI within an evidence review.

2.1.2 Flexible rapid literature review approach

Following our experimentation with Elicit AI, we implemented a literature search strategy that was deliberately wide-ranging. We started by exploring academic sources, using Google Scholar to identify peer-reviewed journal articles, working papers and conference proceedings related to AI applications across all stages of the evaluation process, using general keywords such as “Artificial Intelligence in Research” and “Artificial Intelligence in Evaluation”. This was complemented by targeted searches of organisational websites, including those of international agencies (for example the World Bank, OECD and UN), evaluation networks and government departments, to capture policy reports, guidance documents and technical notes. Recognising that much of the debate and innovation around AI in evaluation sits outside traditional academic publishing, we also included grey literature, encompassing consultancy firms' outputs, think tank briefs, blogs and practitioner reflections. Finally, we carried out targeted web searches to identify rapidly emerging use cases and discussions that may not yet have been formally published or indexed.

Initial literature screening was based on titles and abstracts or executive summaries, after which potentially relevant documents were read in full. We also used snowballing, following references in included papers to identify further sources. For each document, we extracted details on the type of AI tool or approach described, the stage of the evaluation cycle in which it was applied, and the reported benefits, risks or lessons. These insights were then synthesised thematically, using the evaluation cycle as an organising framework. As a rapid review, this work has some limitations. The search was not exhaustive, and while we made efforts to capture both academic and non-academic perspectives, there is a risk that we missed some relevant studies or case examples (Ganann et al., 2010). The lack of a standardised methodology presents challenges for reproducibility relative to systematic literature reviews (Varker et al., 2015).

2.1.3 Scope of the evidence review

The review initially focused on identifying evidence specific to research and innovation (R&I) evaluation contexts. However, evidence in this area proved to be limited. The scope was therefore broadened to include policy domains with strong economic relevance, such as trade, finance and climate-related policies. References from healthcare policy were included only where they were highly relevant, particularly in the UK context or where they illustrated concrete use cases. Only documents available in the public domain were reviewed. Within this scope, we found few sources that directly

link the application of AI to the impact evaluation phase, so our broader scope includes studies with AI applications in other funding and evaluation related activities, including choosing intervention areas, proposal assessment, and monitoring and data collection.

2.2 Phase 2: Qualitative Interviews Approach

To understand how AI tools are actually being used in evaluation practice, and to address the everyday adoption gap identified in Phase 1, we conducted semi-structured interviews with evaluators. In total we interviewed nine respondents across eight organisations (two respondents were drawn from the same organisation), spanning private consultancy, government and higher education, and at various stages of their careers and AI maturity. To preserve confidentiality, respondents are referred to throughout this report by number only, and organisational names are not used. In Table 1 below, we provide an anonymised profile of the respondents we interviewed.

Interviews were recorded using Microsoft Teams, with Fireflies AI providing transcription support. All transcribed interviews were manually checked against the recorded audio for accuracy before being used for thematic analysis. We note the reflexivity of using AI transcription tools within a study of AI in evaluation, and the manual verification step was retained precisely to guard against the quality limitations discussed later in this report. The interviews were analysed thematically, with themes/codes developed inductively from the transcripts and organised around respondents' understanding of AI, their current adoption and tools, use across the evaluation cycle, perceived benefits and concerns, and views on what guidance they consider useful.

Table 1: Profile of interview respondents

Respondent	Role	Sector	AI maturity level
Respondent 1	Mid-level staff	Private consultancy	Advanced
Respondent 2	Senior staff	Private consultancy	Beginner
Respondent 3	Senior staff	Private consultancy	Intermediate
Respondent 4	Senior staff	Government	Limited
Respondent 5	Senior staff	Private consultancy	Beginner
Respondent 6	Senior staff	Private consultancy	Advanced
Respondent 7	Senior staff	Private consultancy	Very advanced
Respondent 8	Consultant	Private (independent)	Expert
Respondent 9	Lecturer / researcher	Higher education	Intermediate

Source: Authors. Two respondents were drawn from the same organisation. Organisation names are withheld.

Note: AI maturity level was assigned by the interviewer rather than self-reported. Based on the interview evidence, this assignment was done considering five indicators thus: breadth of use across the evaluation lifecycle; tool sophistication (general-purpose, configured, or bespoke in-house); workflow integration and reproducibility; governance, validation and quality-assurance practices; and reflexivity and capability to

adapt, build or teach. The levels are ordinal, ascending from Limited to Expert, and are consistent with the breadth of use reported in Table 4. We provide full AI maturity level definitions in Appendix D.

2.3 The funding evaluation cycle as an organising framework

We use a funding evaluation cycle as a framework for organising evidence on the use of AI in evaluation (see Figure 1). The four stages span two functions that UKRI manages separately: funding decisions and grant assessment at stages one and two, and monitoring and evaluation at stages three and four. Presenting them as one cycle reflects the shared tooling and governance questions, not a claim that the functions are interchangeable. The evidence review revealed that AI is applied not only to impact analysis but also to activities such as programme funding decisions and real-time monitoring. Building on the policy-making cycle (Cairney, 2023; IfG, 2024), the funding evaluation cycle illustrates how AI can be integrated throughout the funding and evaluation process. The stages are as follows.

- **Choosing Intervention Areas.** The first stage involves identifying the key problems to address (IfG, 2024) and determining which policy areas, programmes or projects should be targeted. Funding organisations such as UKRI make similar strategic decisions to prioritise intervention areas, which may include defining key selection criteria based on factors such as strategic importance, resource constraints and anticipated impacts.
- **Assessing Project Proposals.** This stage involves systematically reviewing and analysing funding proposals to determine their relevance, feasibility, potential impact and alignment with strategic objectives (OECD, 2021). It includes administrative steps to verify the eligibility of proposals and to score them against predefined key criteria.
- **Project Implementation, Monitoring and Process Evaluation.** At this stage, the selected projects are put into action while continuous monitoring and process evaluation are conducted in parallel. These activities provide timely information on how a project is being implemented, whether it meets its objectives, and whether changes to delivery are required (HM Treasury, 2022).
- **Impact Evaluation and Value for Money Evaluation.** Impact evaluation is the systematic assessment of the changes that occurred as a result of an intervention, the extent to which these changes can be attributed to it, and how far the intervention achieved its intended objectives (HM Treasury, 2022). Value for money analysis assesses whether an intervention has used public resources efficiently, effectively and economically to achieve its outcomes. The results help policymakers understand which interventions deliver measurable benefits and inform future resource allocation and policy design.

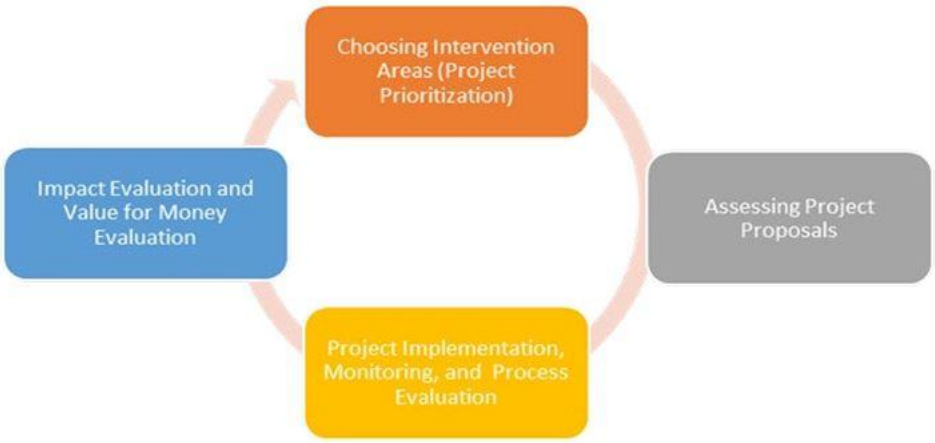


Figure 1: The funding evaluation cycle

2.4 Definitions of AI tools

Here we provide definitions of AI tools commonly used in evaluation along with example applications. These tools are referenced throughout this review, and we specify, wherever possible, the types of AI tools to which evidence relates.

1. Natural Language Processing (NLP). NLP enables computers to process and interpret human language (Hirschberg & Manning, 2015). It underpins tools such as grammar checkers, speech recognition, translation software and chatbots. In evaluation, NLP can support evidence synthesis, sentiment analysis, proposal screening and data extraction from large text sets (Jacob, 2025; Mungalpara, 2023).

2. Large Language Models (LLMs). LLMs are advanced forms of NLP trained on vast datasets and fine-tuned using human feedback. Models such as GPT-4, Google Gemini and Meta's Llama generate fluent, context-aware language and can analyse large volumes of text quickly. Studies show that combining LLMs with human review improves both efficiency and interpretive quality in qualitative analysis (Liu & Sun, 2023; Thomson Reuters, 2025). Evaluation examples include using LLMs to code interview transcripts or summarise large programme reports.

3. Generative AI (GenAI). GenAI refers to tools capable of producing new content, from text to images, video and audio (Feuerriegel et al., 2024). While LLMs are a subset of GenAI, wider applications include image generation, video synthesis and automated design. Evidence suggests that pairing GenAI's speed and structure with human contextual judgement can strengthen project and programme planning (Barcaui & Monat, 2023). In evaluation, GenAI can help simulate potential programme outcomes or generate dashboards.

4. Machine Learning (ML). ML involves algorithms learning patterns from data to make predictions or classifications with limited human intervention (Jasper et al., 2019). It is especially useful for detecting complex relationships in large or unstructured datasets. In evaluation, ML is increasingly applied to automated coding, risk identification and impact assessment (Bravo et al., 2023). Successful adoption typically follows staged implementation, including feasibility checks, pilot testing and long-term integration planning.

5. Big Data Analytics (BDA). BDA draws insights from high-volume, high-variety datasets such as mobile phone records, satellite imagery, electronic transactions and administrative databases (Bamberger, 2024). It enables richer disaggregation and faster feedback than traditional approaches and supports longer-term tracking of programme results (Meier, 2015; Mistry, 2024). Evaluation examples include real-time monitoring of service delivery using telecom data, or mapping programme reach with satellite and GIS tools.

3. Findings from the Evidence Review: AI Use Across the Funding Evaluation Cycle

In its recent report Governing with Artificial Intelligence, the OECD states that AI use within government for policy evaluation is still in its early stages and that its use "has been limited and has progressed slower than in other functions" (OECD, 2025, p.221). Similarly, the report found that "the impact of AI on the practice of policy evaluation is still modest and difficult to measure" (OECD, 2025, p.225). This aligns with the findings of our evidence review, which found that case studies documenting AI use in evaluation, and assessing its effectiveness relative to human effort, were rare.

There are many more studies looking into the potential uses of AI across the funding evaluation cycle, largely from the academic literature. These studies offer UKRI a forward-looking perspective on what types of activities AI could support at each stage. This literature is summarised in the Appendix. Here, we focus on the rarer documented use cases, predominantly from grey literature, which show how AI tools are actually being used at different stages and with what effectiveness, noting in each case the implications for AI use in UKRI evaluations.

3.1 Choosing intervention areas (project prioritisation)

At the first stage of the cycle, the evidence suggests that AI tools support data-driven inputs that help identify key intervention areas. For example, an APEC Policy Brief by Wirjo et al. (2022) highlights the use of NLP technology developed by CitizenLab in Belgium, which helps civil servants process large volumes of data from digital public participation platforms, enabling classification of public input and clustering of similar contributions by theme, demographic profile or geographic location (OPSI, n.d., in Wirjo et al., 2022). In the UK, the Department for Science, Innovation and Technology launched a consultation tool, Consult, which was able to collect and analyse more than 50,000 responses to a government review relating to the Independent Water Commission, reportedly within two hours and while matching human accuracy, with positive implications for efficiency (DSIT, 2025). Other examples include the analysis of crowdsourced data in Bulgaria to identify issues and behaviour trends in the urban environment (Policy Cloud, n.d., in Wirjo et al., 2022), and the Victorian State Government's use of a syndromic surveillance programme combining automated data capture with NLP to monitor reported symptoms in hospitals (Patel et al., 2021).

3.1.2 Implications for UKRI

UKRI strategy teams may be engaged in setting funding agendas and determining the types of support provided to recipient organisations. The evidence suggests that AI tools can support UKRI in agenda setting through NLP techniques to scan, process and synthesise large volumes of existing R&I-related evidence. This could include, for example, mapping the newly defined IS-8 sectors, understanding what works in stimulating R&I within these sectors, and detecting emerging issues. This

approach is comparable to CitizenLab's use of NLP tools by civil servants in Belgium and Australia's use of a syndromic surveillance programme in public health.

3.2 Assessing project proposals (grant assessment)

Although the literature highlights several *potential* uses of AI in the administrative and assessment stages of funding proposals (see Appendix E), we found limited evidence of actual use. A use case from the Independent Evaluation Group (IEG) at the World Bank tested both ChatGPT and the Bank's enterprise version to classify text data related to disaster risk reduction; comparing AI results to manual classifications, ChatGPT achieved over 76 per cent accuracy while the enterprise version reached 57 per cent (Raimondo et al., 2023A). RoRI (2025), in its guidelines for AI use by research funders, identifies current applications in '*automated matching of reviewers and proposals, similarity check between proposals, eligibility check and quality assurance of expert feedback. Uses are mostly limited to the preparation and support of peer review'* (RoRi, 2025, p.43). The Swiss National Science Foundation used ML for reviewer matching which, depending on field, training data and algorithm, achieved accuracy of between 67% to 92% relative to human-selected reviewer choices (RoRI, 2025). 'la Caixa' foundation also experimented with ML assisted classifications of proposals for possible rejection after verification by a human; only one of 86 projects identified by the AI tool for rejection was selected for funding by human experts, showing a high accuracy rate (RoRI, 2025; Cortés et al., 2024).

3.2.1 Implications for UKRI

The evidence suggests opportunities for UKRI to use AI in the administrative stages of assessing proposals for grant funding. Use cases have demonstrated varying levels of success but strong potential for using LLMs, ML and GenAI to assist with pre-screening proposals against eligibility criteria and conducting simple thematic classifications, which, if appropriately implemented, could reduce administrative burden and streamline internal UKRI processes. However, given the ethical, legal and reliability challenges outlined in Section 5, using AI tools for actual assessment and decision-making regarding grant outcomes should be avoided.

3.3 Project implementation, monitoring and process evaluation (M&E)

Some emerging use cases illustrate AI application in programme monitoring. An example from the Tony Blair Institute for Global Change (2024) shows how the UK NHS created an Early Warning System during the Covid-19 pandemic to track real-time and predicted patient demand and resource capacity, down to the availability of specific beds, enabling real-time monitoring and proactive decision-making. Other use cases, although not concerned with policy monitoring, are relevant to broader real-time monitoring efforts: AI tools using ML and NLP allow automation of compliance monitoring in finance by detecting risks, automating audits and enforcing regulatory policies (Atlan, 2025); and HCLTech's automation of gaming reviews for a global

technology company used GenAI to automate data collection and conduct sentiment analysis, resulting in a reported 70 per cent reduction in manual effort alongside improvements in accuracy and turnaround time (HCLTech, n.d.).

3.3.1 Implications for UKRI

Applying AI to real-time monitoring of ongoing support may present challenges for UKRI, given that major funding recipients such as universities and businesses typically operate with high levels of autonomy, so it may be infeasible to track in real time how grant funding is being used. However, in cases of short-term, intensive programmes, such as accelerator programmes, regular digital data collection could be set up and AI tools used to automate the collection and analysis of that data, providing real-time insights into which aspects of the programme are performing well and whether broader adjustments are needed. This is consistent with findings from our interviews (Section 4) where evaluators report using AI for post-intervention data collection and survey analysis rather than for continuous tracking of how programme beneficiaries use support..

3.4 Impact evaluation and value for money evaluation (M&E)

Some emerging use cases illustrate AI application in impact evaluation, including in quantitative and qualitative impact analysis, evidence synthesis and production of evaluation reports.

3.4.1 AI use in quantitative impact analysis.

Wirjo et al. (2022) highlight how the World Bank developed a Machine Learning algorithm to quantify and evaluate the impact of trade agreements on trade flows (Breinlich et al., 2021), enabling data-driven identification of which interventions most strongly influence trade flows and quantification of the marginal impact of each. The same report highlights the use of ML methods, specifically causal forests, to estimate heterogeneous treatment effects when evaluating climate-related policies in the UK, revealing how policy effectiveness varies across regions and economic context (Abrell et al., 2022). A further use case from World Bank IEG shows how ChatGPT can be used to conduct econometric analysis of the association between interventions and desired outcomes in the context of the Bank's economic response to the pandemic; AI was effective at generating code, which made it easier to replicate the study's results (Raimondo et al., 2023A).

3.4.2 AI use in qualitative impact analysis.

Experiments from the IEG team also explored how GPT-4 can be used to conduct sentiment analysis, by classifying whether factors are positively or negatively associated with desired outcomes (Raimondo et al., 2023A). They highlight that GPT-4's accuracy in sentiment analysis can be high (i.e., 94.5% performance level).

Another use case highlights how Large Language Models (LLMs) can be used to conduct qualitative interviews to human subjects in the case of underlying factors influencing non-participation in the stock markets, resulting in rich, high-quality data at significantly lower costs compared to traditional human-led interviews (Chopra and Haaland, 2023). Interestingly, the study highlights how the interview data can better predict economic behaviour. Similarly, the use case from the Behavioural Insight Team (BIT) highlights how Large Language Models (LLMs) can be used to categorize qualitative interviews responses in developing gambling-related interventions (BIT, 2023)

The UK has developed AI technology, 'Consult', in trial with Scottish Government to accelerate public consultation responses on government policies (UK's Government Digital Service, 2025). These artificial intelligence (AI) tools can assist in automatically identifying themes, public sentiments and emerging impacts of a policy in the form of a dashboard. This allows evaluators to better understand how policies affect different groups and incorporate diverse perspectives into the assessment of outcomes.

This documented strength in qualitative coding and analysis is strongly corroborated in practice, as highlighted by our interview findings (Section 4).

3.4.3 AI use in summaries, syntheses and evaluation reports.

Beyond the analysis of impact, AI can help in developing impact evaluation reports. For instance, British International Investment (BII) commissioned an AI-assisted report to assess how well their investment aligned with the priorities, challenges, and development strategies of African and South Asian governments (Wagstaff et al., 2025).

Another use case from the Independent Evaluation Group (IEG) World Bank experiments highlights how accurate OpenAI GPT-4o can be used to produce high-level summaries of evaluation documents using Morocco Country Program Evaluation (Raimondo et al., 2023A). In particular, the World Bank's team found that the OpenAI GPT-4o generative models they used in evaluation "performed well on tasks such as text summarization and synthesis, achieving high scores on metrics related to relevance , coherence , and faithfulness of the generated text". Consequently, given that their AI evaluation experiments yielded "satisfactory" results, the team identified a set of "good practices" that can help in the successful application of AI in evaluations (see later sections).

Beyond summaries, however, World Bank IEG use cases requiring *synthesis* highlight the limitations of using ChatGPT. The IEG World Bank team tested the capacity of ChatGPT to synthesize information from a set of reports by feeding the text from six project evaluations to produce an evaluative synthesis report. The IEG points out that the writing included some high-level insights but the model fabricated examples and evidence (Raimondo et al., 2023B). While the use of LLMs facilitates faster evidence synthesis that draws upon a larger volume of documents and data than humans can

use, the result of the synthesis is often of lower quality: agreement between AI and human judgement can vary significantly, such that in evidence synthesis '*AI judgement cannot yet replace human assessment*' (OECD, 2025 p.225).

3.4.4 Implications for UKRI

In impact evaluation, the evidence suggests UKRI could effectively use ML and GenAI tools to support quantitative impact analysis, particularly the standardisation of software code to enable replication of econometric impact analysis, which is well evidenced by the World Bank Group. UKRI evaluators may also effectively employ GenAI and other LLM-based tools to conduct qualitative impact assessments, such as collecting and analysing interview data and producing summaries of large documents.

The evidence from Raimondo et al., (2023B) suggests AI tools were ineffective for producing evidence syntheses or evaluation reports, since persistent issues such as hallucination and the generation of inaccurate text mean these outputs require substantial human oversight and revision. However, findings from our qualitative interviews, conducted in 2026, show that evaluators see value in using AI tools for synthesis of large amounts of text. This divergence may reflect advances in AI capability, and emphasises how rapid technological change in this area may render previous concerns obsolete.

Although not yet evidenced in the literature, based on IRC experience, we believe there is potential for UKRI to leverage recent advances in industrial classification using ML techniques to enhance quantitative impact evaluation. These approaches can complement commonly used quasi-experimental methods such as Propensity Score Matching. For certain intervention areas, particularly those involving innovative firms in advanced technology sectors, recent developments in web scraping and ML can help identify firms closely resembling supported ones. For example, The DataCity and Beauhurst use taxonomies and keywords to create alternative industrial classifications that extend beyond standard SIC codes by grouping companies according to their technologies and activities as described on their websites (Garcia and Chibelushi, 2023). For R&D interventions targeting highly innovative companies, selecting a control group based on these classifications may be more robust than using firms drawn from broadly defined SIC sectors. These techniques should be combined with more traditional approaches to constructing counterfactuals, since AI-based techniques remain imperfect and continue to be refined.

Overall, the evidence review shows AI's effectiveness in horizon scanning, the administrative stages of proposal assessment, real-time monitoring and data collection, qualitative and quantitative impact analysis, and producing summaries of large documents. Wider decision-making functions are less effective areas in which to incorporate AI. This list is not exhaustive, and experimentation will be required to identify other areas of responsible, safe and effective use. An experiment among Boston Consulting Group employees found that, for realistic, knowledge-intensive tasks,

consultants using AI within its known capabilities completed 12% more tasks, 25%faster, and with 40% higher quality, whereas in tasks outside AI's known capabilities, consultants who did not use AI made significantly fewer mistakes (Patel et al., 2021).

Tables 2 summarises the documented use cases associated with specific AI tools.

Table 2: Summary of evidence on the use of specific AI tools and relevant use cases

Evaluation cycle phase	AI tool	Use case	Source
Choosing intervention areas	NLP	Helps civil servants process large volumes of data from digital participation platforms (Belgium)	OPSI, n.d., in Wirjo et al. (2022)
Choosing intervention areas	Big Data Analytics	Identifies issues and behaviour trends in the urban environment (Bulgaria)	Policy Cloud, n.d., in Wirjo et al. (2022)
Assessing project proposals	Gen-AI / GPT	Classifies text data (limited on more complex tasks)	Raimondo et al. (2023A)
Implementation and monitoring	ML and Big Data	Tracks real-time demand and resource capacity during Covid-19 (UK)	Tony Blair Institute (2024)
Implementation and monitoring	ML and NLP	Automation for compliance monitoring in finance	Atlan (2025)
Impact evaluation	Machine Learning	Quantifies and evaluates the impact of trade agreements on trade flows	Breinlich et al. (2021)
Impact evaluation	Machine Learning	Estimates heterogeneous treatment effects to evaluate carbon pricing policy	Abrell et al. (2022)
Impact evaluation	LLMs	Conducts qualitative interviews	Chopra & Haaland (2023)
Impact evaluation	LLMs	Categorises qualitative interview responses	BIT (2023)
Impact evaluation	Gen-AI / GPT	Analyses associations between interventions and outcomes using econometric models; sentiment analysis	Raimondo et al. (2023A)
Impact evaluation	Gen-AI / GPT	Conducts evidence synthesis (limited) and produces high-level summaries of documents	Raimondo et al. (2023A, 2023B)

Notes: Not all use cases are directly relevant to R&I funding evaluation. Table based on reports that identified the AI tool used.

4. Findings from qualitative interviews: How evaluators are using AI in practice

This section draws on the Phase 2 interviews to set out how evaluators are using AI in practice. As evaluators are mainly involved in data collection and impact analysis, this section relates primarily to the impact evaluation stage of the funding evaluation cycle (Figure 1). It therefore complements the evidence review by delving deeper into changes in impact evaluation practices through AI, and by providing more up to date perspectives than in previous studies. We examine what practitioners involved in R&I evaluation in the UK actually do in terms of AI use, the maturity of their adoption, the tasks where AI adds most value, and the working norms that have emerged.

Taken together, the interviews point to evaluation as a profession in transition, where diverse AI tools are being integrated at different stages of the evaluation process in a measured and pragmatic manner rather than driving wholesale transformation of evaluation practice. However, here it is important to point out that interview findings are based on a relatively small number of respondents with some of them being advanced AI users.

4.1 AI awareness and patterns of use among evaluators

4.1.1. Evaluators' understanding of AI

Respondents' definitions of AI reveal their practical orientation. One respondent offered a notably precise definition:

> *AI is a series of models that intend to simulate or emulate human reasoning that includes synthesising information, summarising information, but also classifying information or data ... making inferences and making connections. So they are ultimately statistical models and ... they are trying to emulate human reasoning and understanding.* (Respondent 6)

Others focused on practical utility. One independent consultant explained:

> *I tend to generalise and just call them AI, generative AI, because it is the one that most people are actually using ... for me they are basically any tools that help us think through things faster, that help us do things faster, things that we might not have been able to do otherwise because we do not have the expertise.* (Respondent 8)

A respondent in a government setting took a more cautious view, framing AI in the evaluation context primarily as NLP:

> *Typically in this context, people tend to be talking about natural language processing. That is the main thing that people are talking about and finding*

> *opportunities to use NLP in various different contexts for evaluation.* (Respondent 4)

In the final analysis, these accounts illustrate a spectrum of conceptualisations that tracks loosely with maturity of use. It runs from the tool-bounded, where AI is effectively equated with the single sanctioned tool (Respondent 2), through a discipline-anchored framing that falls back on natural language processing and the OECD definition (Respondent 4) and a definitional framing that treats the meaning of AI as dependent on one's chosen definition, having developed a bespoke measure of interdisciplinarity (Respondent 9), and statistical framings that stress probabilistic processing and the absence of genuinely new thought (Respondents 1 and 6), to the most technical accounts of AI as a spectrum of computational methods (Respondents 7 and 8). The more advanced adopters tend to hold the more differentiated conceptions, while the most cautious respondent frames AI largely as natural language processing.

4.1.2 A mixed landscape of adoption

The interviews reveal striking variation in the maturity of AI application across the organisations in which respondents work. At one end, some respondents are only beginning to experiment. One candidly admitted that they are "only just starting to look at how it can be used ... just over the last couple of years", with basic experimentation: "we use Copilot as the main one ... we are just sort of trying it on different things and experimenting. So there is no sort of fixed ways of using it yet" (Respondent 5).

At the other extreme, one organisation has developed comprehensive in-house systems. The respondent described a sophisticated, end-to-end approach:

> *We have built a survey, consultation and analysis tool that will help the full end to end process. So it does the quantitative analysis which you get automated in a lot of off-the-shelf tools. But we have that, it does all the calculation, the cleaning. We also do all of the open-ended analysis ... we have a full workflow end to end all the way from defining themes, the human verification process, through to identifying pertinent quotes where there are divergences between stakeholders.* (Respondent 7)

This disparity may reflect broader organisational readiness and willingness to adapt to a fast-changing field. As one respondent from a private consultancy described their position, prior analytical work on AI regulation meant that "*when GPT was released we already had quite a lot of understanding of what that technology can be doing and how helpful it can be actually for our work*" (Respondent 1).

A thematic analysis of the interviews reveals widespread application of diverse AI tools across the different stages of the evaluation process (summarised in Table 4). The tools most frequently cited by respondents were Microsoft Copilot, ChatGPT, Claude, Gemini, Fireflies, Google NotebookLM and, in a few cases, bespoke in-house tools.

4.1.3 AI use across the evaluation process

Table 3 summarises the stages at which respondents reported using, or not using, AI. Most respondents used AI for tasks related to literature reviews, interview transcription, qualitative coding, data synthesis, and report writing. AI is least used for quantitative analysis and for tasks closest to evaluative judgement. Two extreme cases stand out, consistent with the maturity ratings of evaluators in Appendix D: Respondents 7 and 8 report using AI at every stage of the evaluation process, whereas Respondent 4, in a government setting, reports the least use, largely confined to a few peripheral tasks such as summarising and drafting text and searching existing evidence.

Two stages, proposal writing and report writing, merit particular comment, because the headline figures can overstate the role of AI. In both, the human remains the author. For proposals, respondents used AI as a starting point or a critic rather than a ghost-writer. One uploaded the tender and asked the model how it might approach the questions, treating the output as a beginning and never the end (Respondent 1). Another had begun to consider, at the proposal-development stage, whether AI would add scope or efficiency and to write that intention into the research design (Respondent 2). The most advanced respondent used AI to review a draft proposal and flag omissions (Respondent 8).

For report writing, the pattern is similar but more pronounced. Most respondents used AI only to edit, polish, structure or quality-assure text they had written, or to organise a dictated brain-dump into a first structure that they then revised (Respondents 1, 2, 3 and 4), and one was explicit that reporting is always human-led because the tools are not yet good enough to produce first drafts unaided (Respondent 3). Only the two most advanced used AI to support drafting, and even then, they insisted that the reports are written by humans and verified at each stage (Respondents 7 and 8).

One respondent excluded AI from report writing altogether, on the grounds that a report requires the expert triangulation of different kinds of evidence and a transparent, traceable line of argument that a black-box model cannot provide (Respondent 6).

The prevalence of AI at the writing stages therefore reflects assistance with editing and structuring far more than the automated generation of text.

Table 3: Current AI usage across stages of the evaluation process

Stage of evaluation process	Reported use of AI (by respondent)	Reported no use of AI	Number of respondents reporting use (of 9)
Proposal writing	1, 2, 7, 8	3, 4, 5, 6, 9	4 of 9
Literature review	2, 3, 5, 6, 7, 8	All others	6 of 9
Data collection	7 (surveys), 8 (automated), 9	All others	3 of 9
Interview transcription	1, 2, 3, 5, 6, 7, 8	4	7 of 9
Qualitative coding	1, 2, 3, 5, 6, 7, 8	4	7 of 9
Quantitative analysis	7, 8	Others	2 of 9
Data synthesis	1, 2, 3, 5, 6, 7, 8	4	7 of 9
Report writing	1, 2, 3, 4 (editing, polishing, QA); 7, 8 (drafting support)	6 (kept fully human)	6 of 9
Other uses (summarisation, classification, bibliometric analysis, academic publication analysis, project categorisation)	1, 3, 4, 6, 7, 8	All others	6 of 9

Source: Authors, based on interviews. Respondents 2 and 5 are from the same organisation, and where the organisation rather than the individual reported a practice, both respondents are listed. The counts in the final column are indicative rather than exact: they rest on a qualitative mapping, not every respondent addressed every stage, and reported use spans different intensities. At the report-writing stage, for example, reported use ranges from drafting support (Respondents 7 and 8) through editing, polishing and quality assurance (Respondents 1, 2, 3 and 4), while one respondent deliberately keeps report writing fully human (Respondent 6).

4.2. Findings from thematic analysis

Four main themes emerged from the interviews relating to how AI is used within evaluation practice: the human-in-the-loop consensus; qualitative coding and synthesis as a current strength; the scale threshold governing when AI is used; and perceptions of benefits, limitations and output quality. These are themes are discussed below. Two further themes emerge relating to professional concerns that evaluators have regarding AI use (discussed in Section 5) and their views on guidance for AI use (discussed in Section 7). Throughout, we highlight where interview

findings corroborate or contradict findings from the literature review.

4.2.1. The human-in-the-loop consensus

The interviews reveal a clear consensus against AI-only approaches. One respondent noted that AI is “not a thinking thing, because it is only essentially a statistical, probabilistic thing ... I am not confident of its ability to do analysis properly because it does not think about what it is doing” (Respondent 2). This concern is reinforced by

professional norms and integrity, as another respondent emphasised: “so ethically and professionally, for our integrity, it would just not be appropriate to let AI do the whole thing at this point” (Respondent 3). The consensus among all interviewees is that AI tools are, and should be, used with human oversight. As one respondent put it, “we do not do anything exclusively by AI ... it is human verified ... at different various stages” (Respondent 7). Issues of data privacy also constrain complete AI-only usage.

This practitioner consensus closely mirrors the literature review’s finding that human judgement remains essential in all decision-making functions (Section 3; OECD, 2025).

4.2.2 Qualitative coding and synthesis as a current strength

Interview analysis and qualitative coding represent AI's current strength in evaluation. One respondent demonstrated this through a workflow that synthesises interviews and reports into a unified narrative:

> *Just last week there was this interview campaign with over 30 interviews ... I developed a workflow to get all the information ... and ended up with a final strategic narrative ... extracting claims systematically and then doing thematic analysis.* (Respondent 8)

The reported efficiency gains were significant. The respondent estimated that developing the pipeline took two days, after which the process itself took around half a day, whereas manually working through 31 interview transcripts “*that is like three weeks of work ... now ... you can do in one day*”. The respondent also stressed traceability: “*the most important thing is that everything is traceable because I extracted every claim with metadata ... one claim has this code ... this is stored and then every subsequent step can read this code and then reference it*” (Respondent 8).

Similarly, another respondent recounted using AI to categorise a large set of government-funded projects, describing a long list of around a thousand projects that needed to be read line by line, and validating the approach afterwards: “*we did a calculation ... the amount of resources ... manually ... would probably be pretty much the same ... everyone agreed that it is a good idea*” (Respondent 1).

These accounts illustrate why coding, classification and synthesis at scale are reported as the clearest practical gains. This is consistent with the documented use cases in Section 3.4, except in evidence synthesis where the more positive interview findings likely reflect improvements in AI capabilities for synthesis.

The use of AI for transcription, coding and analysis of interview data requires careful consideration of issues around data privacy and data permissions for uploading and analysing interview data.

4.2.3 The scale threshold

Respondents apply a pragmatic volume threshold when deciding whether to use AI. One described the decision rule explicitly:

> *The decision of whether to use AI or not depends on the number, the volume of the data sources. So we very much still say if you have a volume that is 50 to 100, do it manually, just go through the records, read them, read your transcripts ... but if it is interviews, reports, also monitoring data ... and the numbers are beyond 100, then we start considering whether we should automate or not.* (Respondent 6)

This reflects considered thinking about AI's comparative advantage. As the same respondent elaborated, where datasets were previously too large to process manually, evaluators would have drawn a random sample; AI now offers an additional option of processing records in a more automated way, rather than replacing sampling altogether (Respondent 6).

The reasoning behind the threshold is less that manual analysis is inherently better and more a pragmatic judgement about whether automation is worth it. The respondent was explicit that AI is reserved for work that would not otherwise be feasible to do by hand: below the threshold, a researcher can read the whole set, so the effort of setting up the tools and, importantly, of validating their output is not justified by the time saved. Automated outputs have to be checked, for instance by building truth tables on a random sample to estimate false positives and negatives and to judge whether the margin of error is acceptable. Where a client requires enough such checks, the respondent observed, the effort can equate to doing the work manually, so manual analysis becomes the more economical option (Respondent 6). The distinctive advantage of AI is in any case specific to scale, namely that it removes the need to compromise on a random and possibly skewed sample and allows an entire large population to be analysed, an advantage that does not arise below the threshold.

On quality and trust, the picture is more nuanced than a simple preference for manual work. This respondent regarded a having a single model, applied consistently, as potentially more consistent than several human coders following the same instructions. However, they considered AI weaker on transparency, given the black-box nature of the models, and noted the risk of staff taking automated outputs at face value rather than scrutinising them critically. The sensitivity of the data is a further factor, with private or personal data analysed manually or only within ring-fenced internal tools. Other respondents do, however, retain manual analysis precisely where precision and accountability are paramount, turning to AI only to check that nothing significant has been missed (Respondent 3).

Taken together, the threshold reflects a measured, task-specific integration of AI, in which automation is adopted when it is necessary (because the volume makes manual work infeasible) and when it is worth the cost of assuring its quality.

4.2.4 Perceptions of benefits, limitations and views on output quality

here we discuss evaluators' perceived benefits and limitations, and their views on the quality of AI-produced relative to human-produced outputs.

Respondents consistently cite the principal benefits in speed, in the ability to work at a scale or depth that manual methods cannot reach, and in support for coding and synthesis. The limitations they raise are similarly shared: hallucination and accuracy, constraints arising from the volume and quality of available data, bias against protected groups, the erosion of analytical skills, the risk of over-reliance, and the difficulty of transparent disclosure. A few respondents add more specific concerns, including inconsistency or drift in model behaviour over time and a broader fatigue with AI (Respondent 8).

Views on the quality of AI-produced relative to human-produced outputs are less consistent among respondents. At one end, respondents judge current tools inferior to a skilled researcher, describing AI as doing only half the job and as not yet good enough to lead reporting (Respondents 2 and 3). At the other end, respondents report AI matching or exceeding human work on suitable tasks, citing a blind classification test with a high accuracy rate and output that is both faster and deeper (Respondents 7 and 8).

It is clear however that AI's advantage lies at scale and breadth, while humans retain the advantage on nuance, interpretation and judgement. The quality an evaluator can extract depends on their own expertise, since an expert can detect errors that, to a non-expert, feels sensible (Respondent 6).

Table 4 brings these dimensions together for each respondent, including their conceptualisation of AI for completeness. Because the quality of outputs also bears directly on methodological rigour, it is considered further as a risk in Section 5.2.1.

Table 4: Respondents' conceptualisations of AI, perceived benefits and limitations, and views on output quality

Respondent	Conceptualisation of AI	Main perceived benefit	Main perceived limitation	View on AI versus human output quality
1	Statistical processing; regurgitation rather than new thought	Categorisation and analysis at a previously impossible scale	Hallucinated figures; over-reliance and skill loss	Improving, but cannot fully match a human
2	Bounded by the single approved tool	Potential efficiency in coding, not yet realised	Does only half the job; weak on qualitative analysis	Not as good as a human researcher; no replacement yet
3	Large language models; a tireless partner	Speed in thematic coding; checks for missed themes and bias	Hallucination; junior staff cannot judge correctness	Reporting still human-led; analysis more comprehensive than human-only
4	NLP in context; OECD definition; AI-based versus AI-enabled	AI-based methods unlock otherwise-impossible insights	Data quality and volume; QA burden; deskilling	Depends on task; AI-based output is harder to verify
5	Pragmatic, largely undefined	Efficiency for suitable tasks such as literature review	Reliability still uncertain	No different from human output if properly checked
6	Statistical models emulating reasoning	Removes the sample-versus-census compromise at scale	Errors and hallucination; expertise needed to catch them	Expert spots errors quickly; deep search markedly higher quality
7	A spectrum of computational methods	More methodologies and robustness; pattern detection	Needs verification; risk of over- and under-trust	Blind test: high accuracy versus humans, bar nuanced calls
8	Generative AI; tools to think and do faster	Faster and deeper; coding; auditable synthesis	Model drift and inconsistency; over-reliance; AI fatigue	Faster and deeper than manual; hallucination a dated concern
9	Depends on definition; built an in-house algorithm	Insights from large unstructured text beyond human reach	Interviewing loses humanness; EDI bias; accountability	AI strong at scale; humans essential for nuance and connection

Source: Authors, based on interviews. Respondent numbers correspond to Table 1; entries summarise each respondent's most salient points and are perceptions rather than measured assessments.

5. Risks, limitations and mitigation conditions

Using AI tools in evaluation carries significant risks and challenges. Drawing together the literature and the interview evidence, we group these into structural risks, which arise from the wider design, data and governance context in which AI operates, and operational risks, which arise in the conduct of evaluation tasks themselves. We then set out the conditions under which these risks can be mitigated.

The interview evidence is generally consistent with the literature: respondents most frequently raised concerns about analytical quality and reliability, skills erosion, and transparency, each of which maps onto the risks identified in the literature review.

5.1 Structural risks

5.1.1 Equity and bias

AI systems can reproduce and amplify biases present in their training data, leading to outputs that disadvantage particular groups (Liu, 2024; Aninze & Bhogal, 2024). Documented biases include those relating to gender and race (Nadeem et al., 2020; Shrestha & Das, 2022; Marinucci et al., 2023; Gebru, 2020). Because large models learn statistical regularities from vast text corpora, they may encode stereotypes and present them fluently and confidently (Bender et al., 2021). In an evaluation context, biased classification, scoring or sentiment analysis could distort which interventions or groups are judged successful, with consequences for resource allocation. Mitigating this requires careful attention to training data, testing for disparate outcomes across groups, and retaining human judgement over consequential decisions.

5.1.2 Data privacy, confidentiality and security

Evaluation frequently involves sensitive personal and organisational data. Generative AI raises specific privacy and security concerns, including the risk that data submitted to third-party tools is retained or re-used, and exposure to data breaches (Golda et al., 2024; Paul, 2024). Responsibility for privacy and ethical incidents is often unclear (Hadan et al., 2025).

Interview findings show that evaluators recognise this risk. A respondent in a government setting underlined how data conditions constrain appropriate AI use, warning that where data "*is not sufficiently large enough ... is inconsistent ... or it is personal*", making policy decisions "*based on insights gained from an AI ... that is risky in itself*" (Respondent 4). This points to the importance of robust data governance, secure or in-house tools, and clear rules on what data may be processed by which systems.

5.1.3 Legal and ethical frameworks

The legal and ethical frameworks governing AI use in research and evaluation are still developing, creating uncertainty around consent, intellectual property, research integrity and accountability (Chen et al., 2024; Novelli et al., 2024).

Evaluators interviewed framed ethical issues partly as a matter of professional integrity. As one put it, "s*o ethically and professionally, for our integrity, it would just not be appropriate to let AI do the whole thing at this point*" (Respondent 3). Until frameworks mature, responsible practice depends on professional norms, human oversight and conservative use in consequential or sensitive contexts.

5.1.4 Transparency and accountability

Many AI systems function as "black boxes" whose internal reasoning is difficult to interpret, which complicates trust and accountability (Hassija et al., 2024; Pieters, 2011; Ferrario & Loi, 2022). Explainable AI methods can partly address this by making outputs more interpretable and aligned with regulatory frameworks (De Carvalho & Da Silva, 2021), but accountability for AI-supported decisions remains contested (Novelli et al., 2024; Hadan et al., 2025).

The interviews surfaced a related and pressing transparency challenge concerning disclosure. One respondent anticipated AI becoming so routine as to be invisible: "*it may become almost like using Google ... do I need to report that? ... something still yet to be defined*" (Respondent 6). Another noted that use is often not "*immediately transparent ... not transparent to the wider teams ... or to us as the client*", and attributed non-disclosure to both habituation and reluctance: "*some people use it so regularly now, they almost do not think of themselves as using it ... maybe they are a bit embarrassed to say they cannot just draft a decent piece of work without using AI*" (Respondent 2). These accounts show that transparency is also a cultural challenge.

Disclosure practices nonetheless varied considerably across the organisations represented. Of the eight organisations in the sample, six reported having a process or policy for disclosing AI use, typically disclosure to the client together with a statement in the methodology, and in some cases naming the specific tool or model used. the academic respondent relied on standard research-ethics approval and methods reporting rather than a dedicated AI-disclosure process, while the government respondent reported no disclosure process at all, with AI use neither routinely asked about nor declared (Respondent 4).

The disclosure mechanisms described included agreeing use with the client at the outset and recording it in a methodological annex (Respondent 6), requiring staff to declare any use of Copilot as part of quality assurance (Respondent 7; see Section 6.3), naming the specific model applied (Respondent 3), and documenting models, dates and prompts in a full written methodology (Respondent 8). This pattern shows for formal disclosure practices by respondents with more advanced usage. A recurring

qualification is that these processes apply mainly to substantive analytical use: embedded or low-level uses, such as drafting assistance or AI built into everyday software, frequently go undisclosed, partly because staff do not always recognise them as AI (Respondents 1, 2 and 4).

5.2 Operational risks

5.2.1 Scientific rigour, validity and reliability

AI tools can underperform on tasks requiring analytical depth, and may generate fluent but inaccurate or fabricated content (Olteanu et al., 2025; Bender et al., 2021). In documented evaluation experiments, models asked to synthesise evidence produced high-level insights but also fabricated examples and evidence (Raimondo et al., 2023B).

The interviews provide direct corroboration on the level of rigour. One respondent found that "*the AI was not being very thorough ... it would just basically do kind of half the job ... it got some things right, but it seemed to have missed quite a lot of stuff*", and that the lack of depth necessitated repeated iteration that undermined the anticipated efficiency gains, to the point of having to "*scale back to just one question*" (Respondent 2). This indicates that current tools are not reliable substitutes for human analysis in rigorous qualitative work and require validation against human-coded outputs.

5.2.2 Overreliance and skills erosion

Sustained reliance on AI assistance can accelerate skill decay and hinder skill development, sometimes without practitioners' awareness (Macnamara et al., 2024), and excessive dependence on AI dialogue systems has been associated with cognitive offloading and weaker critical thinking (Gerlich, 2025; Zhai et al., 2024; Abuzar, 2025). At the same time, complementary AI skills are increasingly valued in the labour market (Stephany & Teutloff, 2024), creating pressure to adopt tools quickly.

The interviews articulate this as a "*skills paradox*". A respondent observed that junior staff "*are using AI by default for everything and they are not actually developing those critical analytical skills*", producing work that "*feels like it should make sense*" yet reads "*like something that is being written by somebody that does not actually understand what they are writing*" (Respondent 4). Another framed it as the loss of formative struggle: "*there is a risk that people are missing a stepping stone in their formation ... if you do not struggle, how do you learn*?" (Respondent 6). This underscores the need for training, supervised use and the deliberate maintenance of core analytical skills.

5.2.3 Stilted outputs and limited interpretive value

Generative tools can produce text that is grammatically polished but stylistically flat, formulaic or lacking in interpretive nuance (Sharma et al., 2025; Lee, 2024; Nilep, 2024), and analyses of AI-generated creative work point to a homogenisation of style (Leitch & Chen, 2025). For evaluation reporting, where the persuasive and interpretive quality of

narrative matters, this implies that AI drafts require substantive human editing to convey meaning, judgement and context. This is in line with practical use for report writing, where respondents report actively editing any AI generated text (Section 4.2.2).

6. Using AI responsibly in evaluation: Emerging Best Practices

The evidence review and interviews together suggest that there are various emerging good practices which may help mitigate the risks identified above. These practices operate at the human, technology and organisational levels, and are outlined below:

1. Human oversight, verification and validation

Our findings underscore the importance of AI outputs being verified by qualified evaluators at multiple stages, and of using AI to augment rather than replace human judgement, particularly for consequential or evaluative decisions (Respondents 2, 3, 7; OECD, 2025). A comparative UK study of AI-assisted versus human-only evidence review illustrates both the potential and the need for careful design and human checking (BIT, 2025).

Experimental evidence from the World Bank IEG provides an exemplary approach for human oversight in the context of LLMs, here suggesting an iterative, validation-led approach to for analytical tasks (Raimondo et al., 2023A; Raimondo et al., 2025a). Synthesising this evidence, a responsible workflow for applying an LLM to a coding or classification task involves the following loop (Figure 2).

- Representative sampling: select a small, representative development sample and a separate validation sample from the data.
- Prompt development: draft an initial prompt with clear instructions, relevant context and worked examples of the desired output.
- Model evaluation: apply the prompt to the development sample and compare the AI outputs against human-coded results to identify discrepancies.
- Prompt refinement: revise the prompt to address errors, ambiguities and omissions, and repeat until performance is acceptable.
- Validation: apply the refined prompt to the held-out validation sample to confirm performance before scaling up.
- Application and human review: apply the validated prompt to the full dataset, with human verification of outputs and clear documentation of the process.

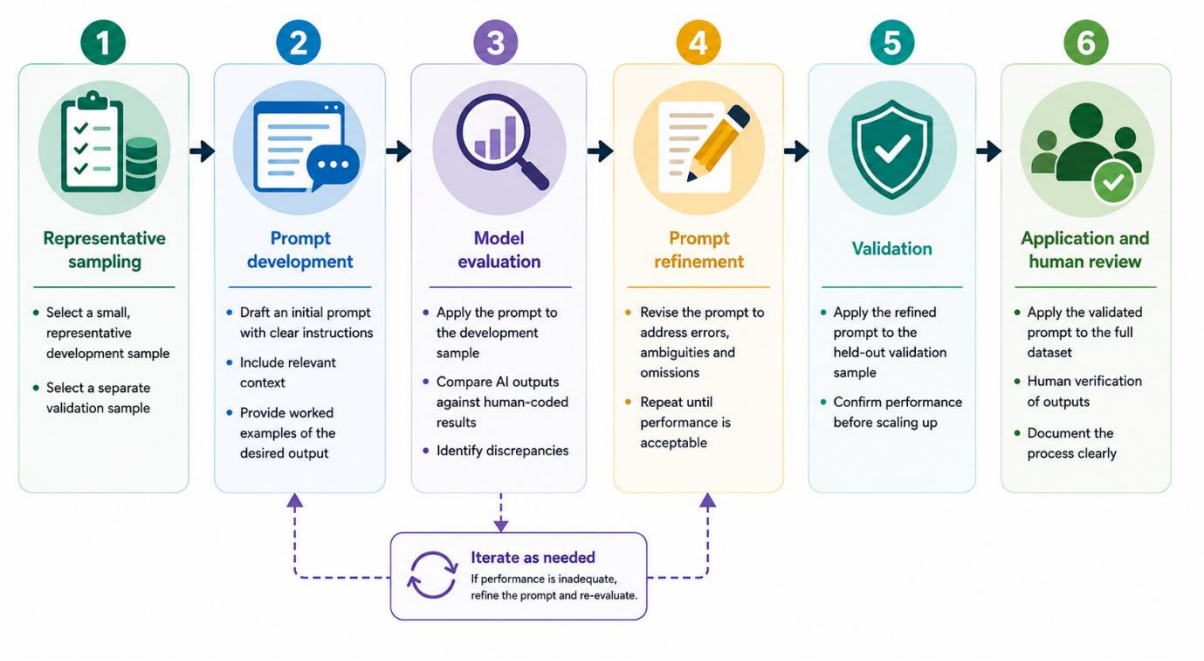


Figure 2: An iterative prompt-and-validate loop for applying LLMs to evaluation tasks
Source: Authors' Synthesis

2. **Using AI for tasks that are within its known capabilities:** Our findings also suggest that responsible AI use required application to tasks where it performs reliably, such as transcription, coding, classification and summarisation, and avoided where evidence shows weaker performance, such as in final decision-making (Section 3; Patel et al., 2021).

3. **Having clear volume thresholds for AI use:** Automation of evaluation tasks appear beneficial only where data volume justifies it and where data quality permits, with manual methods retained for smaller or sensitive datasets (Respondents 4, 6). This recognises AI's competitive advantage of volume and speed.

4. **Traceability and reproducibility:** It is important that, in AI assisted analysis, every claim is extracted with metadata codes so that "*every subsequent step can read this code and then reference it*" (Respondent 8), allowing outputs to be traced back to source. The prompt-and-validate approach (Raimondo et al., 2023A; Raimondo et al., 2025a) also reflects this traceability practice.

5. **Evaluator skills and attitudes:** Effective and responsible use of AI depends on evaluator capability. The literature on evaluator competencies in an AI-enabled environment emphasises a combination of technical fluency, critical appraisal and ethical judgement (Cekova et al., 2025; Mason, 2023). Investing in invest in training and quality assurance could mitigate the risk that AI adoption leads to skills erosion (Respondents 6, 7; Macnamara et al., 2024).

Appendix A sets out a fuller list of skills and attitudes which evaluators should develop, but in summary they should be able to:

- Frame clear, well-contextualised prompts and instructions for AI tools, and refine them iteratively.
- Critically appraise and verify AI outputs against source data and domain knowledge, rather than accepting them at face value.
- Understand the limitations and failure modes of specific tools, including bias, hallucination and inconsistency.
- Safeguard data privacy, confidentiality and security when selecting and using tools.
- Apply ethical and professional judgement, including about when not to use AI.
- Maintain and continue developing core analytical skills, so that AI complements rather than erodes expertise.
- Document and disclose AI use transparently to support trust and reproducibility.

6. **Data governance, privacy and security:** It is important that sensitive data is processed only through secure or in-house systems under clear data-management rules (Golda et al., 2024; Respondent 4).

7. **Transparency and disclosure: AI** use should be transparently disclosed to evaluation colleagues and clients.

8. **Organisational governance and training** Good practice in AI use also requires responsible governance at the organisational level. The most advanced AI adopter interviewed described mitigating skills and quality risks through governance and training: "*we did quite a lot of training ... everyone also has to disclose when they have used Copilot ... that helps with the quality assurance process too*" (Respondent 7). Here, mandatory disclosure serves a dual purpose, supporting both transparency and quality assurance. This illustrates how organisational rules and disclosure norms can together create the conditions for responsible.

We also identified some AI Tool-specific challenges and good practices which we summarise in Appendix C.

7. Principles to guide AI use in evaluation

7.1. The case for a principles-based approach

A clear message from the interviews is that **guidance should be principles-based** rather than narrowly prescriptive, because detailed instructions date quickly in a fast-moving field. As one respondent argued, "*every guideline is outdated by the time it is written ... my main advice would be to make it more about how to learn ... what are the key principles that are going to allow you to continue learning how to use the next iteration*" (Respondent 8). Another corroborated this: "*you need to keep it at a high level ... closer to principles than guidelines ... you cannot be very prescriptive ... that advice will become dated very quickly*" (Respondent 6).

This steer, combined with the sparse and rapidly evolving evidence base, is why this report frames its guidance on AI use around durable principles supported, rather than detailed fixed rules.

7.2. 10 principles for AI use on evaluation

Drawing on the evidence above, AI use in evaluation should be guided by the following principles. We document the evidence supporting each principle.

1. **AI should support evaluator judgement, not replace it.**
   AI should be used to support evaluation tasks, rather than substitute for the judgement of qualified evaluators. Human evaluators should retain responsibility for interpretation, conclusions and any consequential decisions (Respondents 2, 3, 7; OECD, 2025).

2. **AI outputs should be checked and verified.**
   AI-generated outputs should be checked against source data, evaluator knowledge and, where relevant, human-coded results before being used in analysis or reporting (BIT, 2025; Raimondo et al., 2023A; Raimondo et al., 2025a).

3. **AI should be used only where it is appropriate to the task.**
   AI should be applied to tasks where there is evidence that it can perform reliably, such as interview transcription, coding, classification and summarisation. It should be avoided, or used with particular caution, where evidence suggests weaker performance, especially in final decision-making (Section 3; Patel et al., 2021).

4. **AI use should be proportionate to the data and evaluation context.**
   Automation should be considered only where the volume of data justifies it and the data is of sufficient quality. Manual approaches should remain available for smaller, lower-volume or sensitive datasets (Respondents 4, 6).

5. **LLMs should be used through an iterative prompt-and-validate process.** For analytical tasks such as coding or classification, evaluators should develop prompts iteratively, test them on representative samples, compare outputs with human-coded results, refine prompts, validate performance on a separate sample, and only then apply the tool more widely, with continued human review (BIT, 2025; Raimondo et al., 2023A; Raimondo et al., 2025a).

6. **AI use should be transparent and disclosed.** Evaluators and organisations should document and disclose when AI tools have been used. This is important for transparency, trust, reproducibility and quality assurance (Respondents 6, 7; Macnamara et al., 2024).

7. **AI-assisted analysis should be traceable back to the evidence.** Claims, classifications and summaries generated with AI should be traceable and clearly linked back to the underlying data or source material. This reflects the practice described by one respondent of extracting claims with metadata codes so that "every subsequent step can read this code and then reference it" (Respondent 8).

8. **AI use should follow clear data governance and security rules.** Sensitive data should only be processed through secure or in-house systems, and under clear data-management arrangements. This is particularly important where evaluation involves confidential, personal or commercially sensitive data (Golda et al., 2024; Respondent 4).

9. **Organisations should invest in the skills needed for responsible AI use.** Responsible AI use requires evaluators to have the skills to frame good prompts, understand tool limitations, identify bias or hallucination, critically appraise outputs, protect data, apply ethical judgement and know when not to use AI. Training and quality assurance are also important to ensure that AI complements, rather than erodes, core evaluation skills and expertise (Cekova et al., 2025; Mason, 2023; Respondent 7).

10. **AI use should meet relevant ethical, legal and professional obligations.** AI use in evaluation should be governed by clear rules on data privacy, confidentiality, security, transparency and disclosure. Evaluators should also use ethical and professional judgement in deciding whether, when and how AI tools should be used (Golda et al., 2024; Respondent 4; Respondents 6, 7; Macnamara et al., 2024; Cekova et al., 2025; Mason, 2023).

Appendix B provides additional guides for evaluators, prompting them to ask questions which should lead to upholding these principles. Evaluators and commissioners may adapt these questions to their context, and should expect to revise them as tools and evidence develop.

Finally, it is important that any large scale of intensive adoption of AI in UKRI evaluation should be approached experimentally and incrementally. In particular, AI use is best perhaps best introduced through UKRI policy experiments and pilot schemes supported, where feasible, by UKRI-specific internal AI systems to protect data, and deployed in conjunction with conventional tools and methods under human oversight rather than as a replacement for them.

7.3. Key responsibilities of UKRI as evaluation commissioner

As a funder considering AI use in its own evaluation activities, UKRI also has a responsibility to create the conditions for safe, transparent and responsible use. These relate primarily to governance and oversight as well as skills and training.

7.3.1. Governance and oversight: What UKRI should consider

- Defining a clear but adaptable framework outlining appropriate and inappropriate use cases for AI within UKRI evaluations.
    - Adopting an open but experimental approach through carefully designed pilots that test specific important applications, capture challenges and generate lessons.
- Ensuring data security and, where feasible, developing UKRI-specific internal AI systems and platforms so that sensitive data is not exposed to third-party tools.
- Providing clear accountability structures for AI-supported outputs, including for unintended errors or harms.
- Ensuring strong human oversight and ethical safeguards so that AI complements rather than replaces human judgement.
- Establishing clear disclosure norms that incentivise transparency and accurate reporting of AI use.
- Adopting a principles-base approach as outlined above, rather than overly prescriptive guidelines or checklists.

7.3.2. Skills and learning- UKRI should consider:

- Investing in AI literacy and skills training for evaluators, including the prompt-and-validate workflow and critical appraisal of outputs.
- Maintaining open communication with evaluators throughout the evaluation process and creating opportunities for learning and feedback loops.
- Periodically updating standards to reflect emerging best practices and new tools.
- Supporting the deliberate maintenance of core analytical skills, particularly for junior staff, to guard against skills erosion.

8. Conclusion and Implications for UKRI

This report set out to provide UKRI with evidence-based insights to inform the responsible and effective use of AI in evaluation. It brought together a rapid review of the academic and grey literature, organised around the evaluation cycle, and primary interviews with evaluators.

The evidence review found that AI use in evaluation-related activities remains at an early stage. Documented and effective use is currently concentrated in horizon scanning, administrative stages of proposal assessment, real-time monitoring and data collection, qualitative and quantitative analysis, and the summarisation of large documents. More consequential uses, such as final decision-making in proposal assessment, appear to be weaker and riskier areas for AI use. The review also identified a number of structural risks, including bias, data privacy, immature legal and ethical frameworks, and limited transparency and accountability. At the operational level, risks relate to analytical rigour, evaluator skills and interpretive quality.

The interviews complement and largely reinforce these findings. They suggest that evaluation is a profession in transition, with different AI tools being integrated at different stages of the evaluation cycle in a measured and pragmatic way. AI is not currently driving wholesale transformation of evaluation practice. Instead, the most effective reported uses are those that augment human capability, particularly where AI can help evaluators manage scale and improve efficiency, for example through transcription, coding and synthesis. Respondents also applied a pragmatic volume threshold in deciding when automation was useful. Across the interviews, human evaluators continued to exercise critical judgement, provide oversight and maintain methodological rigour.

There was clear optimism about the future role of AI in evaluation, but this was tempered by recognition of current limitations. Respondents raised concerns about quality, skills erosion and transparency. They also favoured a principles-based approach to guidance, rather than highly prescriptive rules, given the pace at which tools and practices are changing.

8.1 Triangulation of evidence

Table 5 sets out this triangulation more systematically, comparing what each strand of evidence shows across the main themes of the report. Overall, the literature and interview evidence are broadly aligned on where AI currently adds value in evaluation, but there are also important differences between documented capabilities, current practice and the perception of risks.

- Both evidence strands point to the clearest benefits of AI use in qualitative coding, classification and analysis, with gains concentrated in speed, scale and efficiency rather than consistently higher quality.

- There is also strong agreement that human oversight remains essential, and strong concerns about over-reliance and skills erosion, particularly for junior staff

- Interviewees reported more sophisticated approaches to evidence synthesis than earlier studies would suggest, particularly where AI is embedded in structured, traceable workflows with human verification. This most likely reflects the pace of model improvement between the 2023 experiments and our 2026 interviews.

- Quantitative impact analysis is comparatively well evidenced in the literature yet barely used by the evaluators we interviewed, which points to a potential opportunity that UKRI evaluators are currently not exploiting.

- Specific concerns over bias and equity are prominent in the literature but are not front of mind in practice, which argues for treating them as a commissioner-level requirement rather than leaving them to practitioner discretion.

8.2. Gaps in the evidence base

Some gaps remain in the evidence base, which future work could usefully address. These include limited evidence on:

- Using AI to design evaluations
- Development of disclosure norms and how disclosure is received by evaluation commissioners

Longer-term implications of AI adoption for the evaluation workforce and professional practice.

Table 5: Triangulation of evidence review and interview findings

Theme	Evidence from the literature review	Evidence from the interviews	Convergence/divergence
1. Overall maturity of AI use	AI use in policy evaluation remains early and has progressed more slowly than in other government functions (OECD, 2025). Documented cases assessing effectiveness against human effort are rare; most evidence concerns potential rather than actual use.	Adoption maturity varies significantly across evaluators, ranging from early experimentation with Copilot (Respondent 5) to a bespoke end-to-end survey and consultation pipeline (Respondent 7).	Current evaluation practice appears somewhat ahead of the documented maturity of AI use, likely reflecting publication lag.
2. The importance of human oversight and decision-making	Human judgement remains essential for consequential decisions, and AI judgement cannot yet replace human assessment (OECD, 2025). Use for final grant funding decisions should be avoided.	The importance of human oversight was unanimous. No respondent uses AI unsupervised, and human verification is built in at multiple stages (Respondents 3, 7).	Strongly aligned.
3. Using AI for qualitative coding, classification and interview analysis	Documented accuracy is high (and has been for some time), e.g., 94.5% on sentiment analysis with models as early as GPT-4 (Raimondo et al., 2023A), LLM categorisation of interview responses (BIT, 2023), and ML classification of proposals at 'la Caixa' (Cortés et al., 2024).	Qualitative coding, classification and interview analysis was the clearest reported gain, used by 7 of 9 respondents. For example, three weeks of manual coding across 31 transcripts was compressed to around a day (Respondent 8).	Strongly aligned.
4. Using AI for evidence synthesis	Earlier versions of ChatGPT produced some high-level insight but fabricated examples and evidence when synthesising six project evaluations (Raimondo et al., 2023B); synthesis quality was below human assessment.	Respondents see substantial value in synthesis at scale where the workflow is engineered for traceability, for example through claim-level extraction with metadata (Respondent 8). One respondent treats hallucination as a dated concern relating to earlier models.	Here, findings diverge. This likely reflects both model improvement between the 2023 experiments and our 2026 interviews, and the difference between single-prompt tests and structured pipelines with human verification. This suggests there may be value in re-testing newer models for synthesis.
5. Using AI for summarisation,	Earlier models (GPT-4o) performed well on summarisation, scoring highly for relevance,	There is widespread use, but of varying depth. Most respondents edit, structure	Evaluators are more conservative than the documented capability of AI would suggest. This perhaps reflects

Theme	Evidence from the literature review	Evidence from the interviews	Convergence/divergence
drafting and report writing	coherence and faithfulness (Raimondo et al., 2023A), but generative text was stylistically flat and homogenising (Sharma et al., 2025; Leitch and Chen, 2025).	and quality-assure their own text; only Respondents 7 and 8 use AI for drafting support, and Respondent 6 excludes it from reporting altogether.	stylistic preferences and a desire to retain authorship as a professional norm.
6. Using AI for quantitative impact analysis	Use here is well documented, for example using ML to quantify trade agreement impacts (Breinlich et al., 2021), causal forests for heterogeneous treatment effects (Abrell et al., 2022), and GPT-assisted econometric code and replication (Raimondo et al., 2023A).	This was the least used area of application, reported by only 2 of 9 respondents.	Findings here diverge, likely reflecting different contexts. The documented cases come from large institutions with in-house data science capacity focussed on macroeconomic analysis, whereas the interviewed UK R&I evaluators work at smaller scale. This could therefore be an area of underexplored or under-exploited opportunity for UKRI evaluators.
7. Using AI for real-time monitoring and data collection	Monitoring use cases are documented, including the NHS Covid-19 early warning system (Tony Blair Institute, 2024), compliance monitoring in finance (Atlan, 2025) and automated collection with sentiment analysis (HCLTech, n.d.).	AI is used for post-intervention data collection and survey analysis rather than continuous tracking; data collection is reported by 3 of 9 respondents.	Here there is some divergence, partly explained by context. The documented cases work in settings with direct operational data, whereas UKRI funding recipients operate with high autonomy and in contexts not immediately amenable to continuous monitoring.
8. Nature of the gains from using AI	Documented benefit is strongest on speed and volume, including over 50,000 consultation responses analysed in around two hours (DSIT, 2025) and a reported 70 per cent reduction in manual effort (HCLTech, n.d.). Evidence on improved validity is more mixed however, and gains are eroded outside the tools' capability boundary.	Respondents also cite speed, scale and depth, but views on quality relative to a skilled human range from doing "half the job" (Respondent 2) to matching humans in a blind classification test (Respondent 7).	The evidence here is aligned. Both evidence strands support clear benefits of AI in improved speed and efficiency, but not necessarily improved quality. The latter appears to depend on the specific evaluation task.
9. Concerns over skills erosion and over-reliance	The review suggests that sustained reliance on AI accelerates skill decay, sometimes without practitioners' awareness (Macnamara et al., 2024), and could be associated with	Skills erosion is also a concern particularly as it affects junior staff who may default to AI, and therefore may not develop analytical judgement	Findings are aligned, and complementary, with the literature identifying the mechanism of skills erosion and the interviews identifying specific risks to junior staff, which

Theme	Evidence from the literature review	Evidence from the interviews	Convergence/divergence
	cognitive offloading and weaker critical thinking (Gerlich, 2025).	(Respondent 4), and may lose the formative struggle of doing the work (Respondent 6).	points to the importance of training and supervision for this group.
10. Bias, privacy and legal frameworks	This is extensively documented: gender and race bias was found to be encoded and reproduced fluently (Nadeem et al., 2020; Bender et al., 2021), retention and re-use of data by third-party tools (Golda et al., 2024), and immature legal and ethical frameworks (Novelli et al., 2024).	Privacy is prominent and actively managed by respondents, with sensitive data kept manual or confined to ring-fenced internal tools. Bias and equity feature far less however, raised mainly by Respondents 3 and 9.	There is convergence of the nature of risks but divergence of emphasis. Bias and equity concerns are well evidenced in the literature review but are not front of mind for evaluators in practice. This may mean that issues around bias and equity should be handled as a commissioner-level requirement rather than leaving it to practitioner discretion.
11. Transparency and disclosure	This is usually framed technically and legally, around black-box models, explainability and contested accountability (Hassija et al., 2024; Novelli et al., 2024).	The interviews frame this as a cultural and practical issue. Six of eight organisations have a disclosure process, but embedded or low-level use routinely goes undeclared because staff do not recognise it as AI use, or are reluctant to say so (Respondents 1, 2, 4 and 6).	Both evidence strands recognise this issue but disclosure norms need to better address the practical, routine use of AI.
12. Form of guidance for the use of AI in evaluation	Existing funder and evaluation guidance is largely principles and good-practice based rather than overly prescriptive (RoRI, 2025; OECD, 2025).	Respondents expressed a clear preference for principles over prescriptive guidance, given the pace at which tools and practices are changing.	Both evidence strands align and this directly supports the principles-based approach set out in Section 7.

Source: Authors, comparing the evidence review (Section 3) with the interview findings (Section 4). Respondent numbers correspond to Table

REFERENCES

Abro, A. A., Talpur, M. S. H., & Jumani, A. K. (2023). *Natural Language Processing Challenges and Issues: A literature review*. *Gazi University Journal of Science*, 36(4), 1522–1536. https://doi.org/10.35378/gujs.1032517

Abuzar, M. (2025). University Students' Trust in AI: Examining Reliance and Strategies for Critical Engagement. *International Journal of Interactive Mobile Technologies*, *19*(7).

Ahmed, A. A. A., Agarwal, S., Kurniawan, I. G. A., Anantadjaya, S. P., & Krishnan, C. (2022). Business boosting through sentiment analysis using Artificial Intelligence approach. *International Journal of System Assurance Engineering and Management*, *13*(Suppl 1), 699-709.

Al Naqbi, H., Bahroun, Z. and Ahmed, V., 2024. Enhancing work productivity through generative artificial intelligence: A comprehensive literature review. *Sustainability*, 16(3), p.1166.

Alam, M. S., Mrida, M. S. H., & Rahman, M. A. (2025). Sentiment analysis in social media: How data science impacts public opinion knowledge integrates natural language processing (NLP) with artificial intelligence (AI). *American Journal of Scholarly Research and Innovation*, *4*(01), 63-100.

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., bin Saleh, K., Badreldin, H. A., Al Yami, M. S., & Albekairy, A. M. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education, 23*(1), Article 689. https://doi.org/10.1186/s12909-023-04698-z

ALP Consulting. (2025). *How Artificial Intelligence (AI) Can Be Used in Compliance?*. ALP. Available at: https://alp.consulting/artificial-intelligence-ai-in-compliance/?utm

Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., bin Saleh, K., Alowais, S. A., Alshaya, O. A., Rahman, I., Al Yami, M. S., & Albekairy, A. M. (2023). The emergent role of artificial intelligence, natural language processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy, 19*(8), 1236–1242. https://doi.org/10.1016/j.sapharm.2023.05.016

Al-Zahrani, A. M. (2024). Balancing act: Exploring the interplay between human judgment and artificial intelligence in problem-solving, creativity, and decision-making. *Igmin research*, *2*(3), 145-158.

Aninze, A., & Bhogal, J. (2024). Artificial Intelligence Life Cycle: The Detection and Mitigation of. In *Proceedings of the International Conference on AI Research*. Academic Conferences and publishing limited. Available at: https://www.researchgate.net/profile/Ashionye-Aninze/publication/386501524_Artificial_Intelligence_Life_Cycle_The_Detection_and_Mitigation_of_Bias/links/675b20afeea8d248be645ba8/Artificial-Intelligence-Life-Cycle-The-Detection-and-Mitigation-of-Bias.pdf

Arslan, B., Lehman, B., Tenison, C., Sparks, J. R., López, A. A., Gu, L., & Zapata-Rivera, D. (2024, October 7). *Opportunities and challenges of using generative AI to personalize educational assessment*. *Frontiers in Artificial Intelligence, 7*, Article 1460651. https://doi.org/10.3389/frai.2024.1460651

Asunmonu, A. A. (2025, April 7). *AI-Powered Consumer-Generated Insights for Product Innovation*. *Mikailalsys Journal of Advanced Engineering International, 2*(2), 129–142. https://doi.org/10.58578/mjaei.v2i2.5335

Atlan. (2025, 21 April). *AI for Compliance Monitoring in Finance: Future-Proofing Your Data Estate for Change*. Atlan.com. Retrieved from: https://atlan.com/know/ai-governance/ai-compliance-monitoring-finance/#ai-for-compliance-monitoring-in-finance-top-use-cases

Baclic, O., Tunis, M., Young, K., Doan, C., Swerdfeger, H., & Schonfeld, J. (2020). Challenges and opportunities for public health made possible by advances in natural language processing. *Canada Communicable Disease Report, 46*(6), 161–168. https://doi.org/10.14745/ccdr.v46i06a02

Bamberger, M. (2024). *Chapter 34: Using big data to strengthen evaluation*. In K. E. Newcomer & S. W. Mumford (Eds.), *Research handbook on program evaluation* (pp. 667–686). Edward Elgar Publishing. https://www.elgaronline.com/edcollchap/book/9781803928289/book-part-9781803928289-44.xml

Barcaui, A., & Monat, A. (2023). Who is better in project planning? Generative artificial intelligence or project managers? *Project Leadership and Society*, *4*, 100101.

Behera, R. K., Bala, P. K., Panigrahi, P. K., & Dasgupta, S. A. (2023). Adoption of cognitive computing decision support system in the assessment of health-care policymaking. *Journal of Systems and Information Technology*, *25*(4), 395-439.

Belagatti, P. (2025, January 13). *Evaluating large language models: A complete guide*. SingleStore [Blog post]. Available at: https://www.singlestore.com/blog/complete-guide-to-evaluating-large-language-models/

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). *On the dangers of stochastic parrots: Can language models be too big?* In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), 610–623. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

BIT (Behavioural Insights Team (BIT), Department for Science, Innovation and Technology, & Department for Digital, Culture, Media and Sport). (2025, 23 April). *AI-assisted vs human-only evidence review: Results from a comparative study*. GOV.UK. [Online] Retrieved from: https://www.gov.uk/government/publications/ai-assisted-vs-human-only-evidence-review

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R. B., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellón, R.,

Chatterji, N. S., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). *On the opportunities and risks of foundation models* (arXiv preprint arXiv:2108.07258). CoRR. https://doi.org/10.48550/arXiv.2108.07258

Bravo, L., Hagh, A., Joseph, R., Kambe, H., Xiang, Y., & Vaessen, J. (2023, July 20). *Chapter 1: Machine learning applications in evaluation*. In *Machine Learning in Evaluative Synthesis: Lessons from Private Sector Evaluation in the World Bank Group* (IEG Methods and Evaluation Capacity Development Working Paper Series). Independent Evaluation Group, World Bank Group. [online] Available at: https://ieg.worldbankgroup.org/sites/default/files/Data/Evaluation/files/methods_paper-machine_learning.pdf

Breinlich, Holger, Corradi, Valentina, Rocha, Nadia, Ruta, Michele, Santos Silva, J.M.C., & Zylkin, Tom. (2021). *Machine Learning in International Trade Research: Evaluating the Impact of Trade Agreements*. (Policy Research Working Paper 9629). The World Bank. https://openknowledge.worldbank.org/handle/10986/35451

Cairney, P. (2023, 9 February). *What does policy making look lile.* Government UK. [Blog post[. Available at: https://publicpolicydesign.blog.gov.uk/2023/02/09/what-does-policymaking-look-like/

Cekova, D., Corsetti L., Ferretti, S. and Vaca, S. (2025). *Considerations and Practical Applications for using Artificial Intelligence (AI) in Evaluations. Technical Note.* CGIAR Independent Advisory and Evaluation Service (IAES). *Available at*: https://iaes.cgiar.org/sites/default/files/pdf/Considerations%20and%20Practical%20Applications%20for%20Using%20Artificial%20Intelligence%20AI%20in%20Evaluations.pdf

Chen, Z., Chen, C., Yang, G., He, X., Chi, X., Zeng, Z., & Chen, X. (2024). Research integrity in the era of artificial intelligence: Challenges and responses. *Medicine*, *103*(27), e38811.

Chopra, F. and Haaland, I. (2023) *Conducting qualitative interviews with AI*. CESifo Working Paper Series, No. 10666. Munich: CESifo.

Coffman, J., & Reid, C. (2024, June). *Chapter 24: Emerging trust-based evaluation approaches in philanthropy*. In K. E. Newcomer & S. W. Mumford (Eds.), *Research handbook on program evaluation* (pp. 667-686). Edward Elgar Publishing. https://www.elgaronline.com/edcollbook/book/9781803928289/9781803928289.xml

Cortés, C. C., Parra-Rojas, C., Pérez-Lozano, A., Arcara, F., Vargas-Sánchez, S., Fernández-Montenegro, R., Casado-Marín, D., Rondelli, & López-Verdeguer, I. (2024). AI-assisted prescreening of biomedical research proposals: ethical considerations and the pilot case of “la Caixa” Foundation. *Data & Policy*, *6*, e49.

De Carvalho, M. S., & Da Silva, G. L. (2021, September). Inside the black box: using Explainable AI to improve Evidence-Based Policies. In *2021 IEEE 23rd Conference on Business Informatics (CBI)* (Vol. 2, pp. 57-64). IEEE.

Department for Science, Innovation and Technology DSIT (2025). Ground-breaking use of AI saves taxpayers' money and delivers greater government efficiency. Press release, 16th October 2025. Available: https://www.gov.uk/government/news/ground-breaking-use-of-ai-saves-taxpayers-money-and-delivers-greater-government-efficiency. [Accessed 28/11/2025).

Dhyani, P. (2025). *Natural language processing: Challenges and applications*. Jellyfish Technologies. [Blog Post]. Retrieved from https://www.jellyfishtechnologies.com/natural-language-processing-challenges-and-applications/

Djunaedi, H. (2024). AI as employee performance evaluation: An innovative approach in human resource development. *Power System Technology*, *48*(1), 2008-2021.

Evaluation Task Force. (2025). *Guidance on the Impact Evaluation of AI Interventions.* Gov.uk [Online] Available at: https://www.gov.uk/government/publications/the-magenta-book/guidance-on-the-impact-evaluation-of-ai-interventions-html?utm

Ferrario, A., & Loi, M. (2022, June). How explainability contributes to trust in AI. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 1457-1466).

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative ai. *Business & Information Systems Engineering*, *66*(1), 111-126.

Flahavan, E. (2024). *Using Large Language Models to unlock value in unstructured text data*. [online] Medium. Available at: https://medium.com/@ed.flahavan/using-large-language-models-to-unlock-value-in-unstructured-text-data-164ae19d5661

Fleurence, R., Bian, J., Wang, X., Xu, H., Dawoud, D., Higashi, M., & Chhatwal, J. (2024). Generative AI for Health Technology Assessment: Opportunities, Challenges, and Policy Considerations. *arXiv preprint arXiv:2407.11054*.

Ganann, R., Ciliska, D., & Thomas, H. (2010). Expediting systematic reviews: methods and implications of rapid reviews. *Implementation Science, 5*, 56.

Garcia, F., Cibelushi, C., (2023, 2 February). *Building the AI Taxonomy with Innovate UK KTN*. The Data City. Available at: https://thedatacity.com/blog/building-the-ai-taxonomy-with-ktn/

Gebru, T. (2020). Race and gender. *The Oxford handbook of ethics of AI*, *4*, 253.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, *15*(1), 6.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020, May 30). *A review of challenges and opportunities in machine learning for health*. *AMIA Joint Summits on Translational Science Proceedings*, 2020, 191–200. https://pmc.ncbi.nlm.nih.gov/articles/PMC7233077/?utm

Golda, A., Mekonen, K., Pandey, A., Singh, A., Hassija, V., Chamola, V., & Sikdar, B. (2024). Privacy and security concerns in generative AI: a comprehensive survey. *IEEE Access*, *12*, 48126-48144.

Google Cloud. (n.d.). *What is big data?* Google Cloud. [Blog Post. Retrieved from https://cloud.google.com/learn/what-is-big-data

Government Digital Service. (2025). *Consult*. AI.GOV. [Online]. UK.Retrieved from: https://ai.gov.uk/projects/consult/

Government Digital Service. (2025, February 10). *Artificial Intelligence Playbook for the UK Government* (ISBN 9781036688745) [PDF]. UK Government. Retrieved from: https://assets.publishing.service.gov.uk/media/67aca2f7e400ae62338324bd/AI_Playbook_for_the_UK_Government__12_02_.pdf

Gupta, S., Leszkiewicz, A., Kumar, V., Bijmolt, T., & Potapov, D. (2020). Digital analytics: Modeling for insights and new methods. *Journal of interactive marketing*, *51*(1), 26-43.

Hadan, H., Mogavi, R. H., Zhang-Kennedy, L., & Nacke, L. E. (2025). Who is Responsible When AI Fails? Mapping Causes, Entities, and Consequences of AI Privacy and Ethical Incidents. *arXiv preprint arXiv:2504.01029*.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... & Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, *16*(1), 45-74.

HCLTech. (n.d.). *GenAI-powered sentiment analyzer reduces manual effort by 70 percent*.HCLTECH.COM. [Online]. Available at: https://www.hcltech.com/case-study/genai-powered-sentiment-analyzer-reduces-manual-effort-by-70-percent

He, L., Omranian, S., McRoy, S., & Zheng, K. (2025, May). Using large language models for sentiment analysis of health-related social media data: empirical evaluation and practical tips. In *AMIA Annual Symposium Proceedings* (Vol. 2024, p. 503).

Hinge, P., Salunkhe, H., & Boralkar, M. (2023, May). Artificial intelligence (ai) in hrm (human resources management): A sentiment analysis approach. In *International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022)* (pp. 557-568). Atlantis Press.

Hirschberg, J., & Manning, C. D. (2015, July 17). *Advances in natural language processing*. *Science, 349* (6245), 261–266. https://doi.org/10.1126/science.aaa8685

HM Teasury. (2022). The Green book. Gov.UK [Online]. Available at: https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-government/the-green-book-2020

Huang, W., Zhang, L., & Wu, X. (2022, June). Achieving counterfactual fairness for causal badit. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, No. 6, pp. 6952-6959).

Institute for Government (IfG). (2024, 19 June). *Understanding Policy Making*. Retrieved from: https://www.instituteforgovernment.org.uk/publication/understanding-policy-making

J. Abrella, M. Kosch, and S. Rausch. (2022) How Effective Is Carbon Pricing? A Machine Learning Approach to Policy Evaluation. *Journal of Environmental Economics and Management.* 112 . https://doi.org/10.1016/j.jeem.2021.102589

Jacob, S. (2024). Navigating the challenges of policy evaluation. *Canadian Public Administration, 67*(2), 282–290. https://doi.org/10.1111/capa.12571

Jacob, S. (2025). *Artificial intelligence and the future of evaluation: From augmented to automated evaluation*. *Digital Government: Research and Practice, 6*(1), Article 10. https://doi.org/10.1145/3696009

Jasper, P., Haldrup, S. V., & Mikhaylov, S. J. (2019, March). *Artificial intelligence and machine learning methods for programme evaluations in Global Affairs Canada: Literature review*. Oxford Policy Management. https://www.opml.co.uk/sites/default/files/migrated_bolt_files/a3489-ai-and-ml-for-evaluations.pdf

Jasper, P., Vester Haldrup, S., & Mikhaylov, S. J. (2019). Artificial intelligence and machine learning methods for programme evaluations in global affairs Canada. *Literature Review*.

Khan, R., Shrivastava, P., Kapoor, A., Tiwari, A., & Mittal, A. (2020). Social media analysis with AI: sentiment analysis techniques for the analysis of twitter covid-19 data. *J. Crit. Rev*, *7*(9), 2761-2774.

Koliousis, I., Al-Surmi, A., & Bashiri, M. (2024). Artificial intelligence and policy making; can small municipalities enable digital transformation? *International Journal of Production Economics*, *274*, 109324.

Kong, F., Li, Y., Nassif, H., Fiez, T., Henao, R., & Chakrabarti, S. (2023, August). Neural insights for digital marketing content design. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 4320-4332).

Kousha, K., & Thelwall, M. (2022). Artificial intelligence technologies to support research assessment: A review. *arXiv preprint arXiv:2212.06574*.

Kumar, P. (2024). Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, *57*(10), 260.

Kuo, K. Y. (2025). Utilizing AI and Analytics for Public Opinion Analysis in Taiwan's Government Policy-Making. In *AI, Analytics and Strategic Decision-Making* (pp. 218-248). Routledge.

Lee, S. J. (2024). Analyzing the use of ai writing assistants in generating texts with standard american english conventions: A case study of chatgpt and bard. *The CATESOL Journal*, *35*(1).

Leitch, A., & Chen, C. (2025, April). Unlimited Editions: Documenting Human Style in AI Art Generation. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-9).

LinkedIn. (2023, March 8). *What benefits and challenges arise from using natural language processing in text analytics?* LinkedIn Advice. https://www.linkedin.com/advice/1/what-benefits-challenges-using-natural-language

Liu, A., & Sun, M. (2023, December 1). *From voices to validity: Leveraging large language models (LLMs) for textual analysis of policy stakeholder interviews* (Version 1) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2312.01202

Liu, Y. (2024). Unveiling bias in artificial intelligence: Exploring causes and strategies for mitigation. *Appl. Comput. Eng*, *76*, 124-133.

Loor-Torres, R., Duran, M., Toro-Tobon, D., Mateo Chavez, M., Ponce, O., Soto Jacome, C., Segura Torres, D., Algarin Perneth, S., Montori, V., Golembiewski, E., Borras Osorio, M., Fan, J. W., Singh Ospina, N., Wu, Y., & Brito, J. P. (2024). *A systematic review of natural language processing methods and applications in thyroidology*. *Mayo Clinic Proceedings: Digital Health, 2*(2), 270–279. https://doi.org/10.1016/j.mcpdig.2024.03.007

Macnamara, B. N., Berber, I., Çavuşoğlu, M. C., Krupinski, E. A., Nallapareddy, N., Nelson, N. E., Smith, P. J., Wilson-Delfosse, A. L., & Ray, S. (2024). Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cognitive Research: Principles and Implications, 9*(1), Article 46. https://doi.org/10.1186/s41235-024-00572-8

Mannarini, G., Posa, F., Bossy, T., Massemin, L., Fernandez-Castanon, J., Chavdarova, T., ... & Hartley, M. A. (2022). What If…? Pandemic policy-decision-support to guide a cost-benefit-optimised, country-specific response. *PLOS Global Public Health*, *2*(8), e0000721.

Marinucci, L., Mazzuca, C., & Gangemi, A. (2023). Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender. *AI & SOCIETY*, *38*(2), 747-761.

Mason, S. (2023). Finding a safe zone in the highlands: Exploring evaluator competencies in the world of AI. *New Directions for Evaluation*, *2023*(178-179), 11-22.

Meier, P. (2015). Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response. CRC Press.

Mejova, Y. (2009). Sentiment analysis: An overview. *University of Iowa, Computer Science Department*, *5*, 1-34.

Mistry, J. (2024). *Simplified guide to big data: Analytics, benefits and challenges*. Ace Infoway. https://www.aceinfoway.com/blog/big-data-analytics-benefits-and-challenges/

Mukhamediev, R. I., Popova, Y., Kuchin, Y., Zaitseva, E., Kalimoldayev, A., Symagulov, A., Levashenko, V., Abdoldina, F., Gopejenko, V., Yakunin, K., Muhamedijeva, E., & Yelis, M. (2022). Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges. *Mathematics, 10*(15), 2552.

Mukherjee, S., & Bhattacharyya, P. (2013). Sentiment analysis: A literature survey. *arXiv preprint arXiv:1304.4520*.

Mungalpara, J. (2023, April 27). *Evaluation methods in Natural Language Processing (NLP): Part-1*. Medium. https://jaimin-ml2001.medium.com/evaluation-methods-in-natural-language-processing-nlp-part-1-ffd39c90c04f

Nadeem, A., Abedin, B., & Marjanovic, O. (2020). Gender bias in AI: A review of contributing factors and mitigating strategies. *Aisel*. Available at: https://aisel.aisnet.org/acis2020/27/

Nilep, C. (2024). *AI and new forms of knowledge work: How writing education can help equip students for an uncertain future*. ToXiv. Retrieved from http://toxiv.ilas.nagoya-u.ac.jp/2024/NILEP2024.pdf

Novelli, C., Taddeo, M., & Floridi, L. (2024). Accountability in artificial intelligence: What it is and how it works. *AI & Society*, *39*(4), 1871-1882.

Observatory of Public Sector Information (OPSI). (n.d.). *Unlocking the Potential of Crowdsourcing for Public Decision-making with Artificial Intelligence.* OPSI.org. [Online]. Available at: https://oecd-opsi.org/innovations/unlocking-the-potential-of-crowdsourcing-for-public-decision-making-with-artificial-intelligence/

OECD (2025), Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions, OECD Publishing, Paris, https://doi.org/10.1787/795de142-en.

Olteanu, A., Blodgett, S. L., Balayn, A., Wang, A., Diaz, F., Calmon, F. D. P., ... & Barocas, S. (2025). Rigor in AI: Doing Rigorous AI Work Requires a Broader, Responsible AI-Informed Conception of Rigor. *arXiv preprint arXiv:2506.14652*.

Organisation for Economic Co-operation and Development (OECD). (2021). *Applying evaluation criteria thoughtfully*. OECD Publishing. https://doi.org/10.1787/543e84ed-en

Patel, J., Manetti, M., Mendelsohn, M., Mills, S., Felden, F., Littig, L. and Rocha, M. (2021, 5 April): AI Brings Science to the Art of Policymaking. Boston Consulting Group. Available at: https://web-assets-pdf.bcg.com/prod/how-artificial-intelligence-can-shape-policy-making.pdf

Paul, J. (2024, November). Privacy and data security concerns in AI. *ResearchGate.*

Pencheva, I., Esteve, M., & Mikhaylov, S. J. (2020). Big Data and AI–A transformational shift for government: So, what next for research? *Public Policy and Administration*, *35*(1), 24-44.

Pieters, W. (2011). Explanation and trust: what to tell the user in security and AI?. *Ethics and information technology*, *13*(1), 53-64.

Policy Cloud. (n.d.). *Urban Policy Making through Analysis of Crowdsourced Data. Policycloud.eu.* [Online]. Available at: https://policycloud.eu/pilots/urban-policy-making-throughanalysis-crowdsourced-data

Qian, C., Mathur, N., Zakaria, N. H., Arora, R., Gupta, V., & Ali, M. (2022). Understanding public opinions on social media for financial sentiment analysis using AI-based techniques. *Information Processing & Management*, *59*(6), 103098.

Raimondo, E., Ziulu., V., Anuj., H (2025a). *Balancing innovation and rigor: Guidance on how thoughtfully integrate AI in evaluation* [Blog post]. World Bank Independent Evaluation Group. Available at:https://ieg.worldbankgroup.org/blog/balancing-innovation-and-rigor-guidance-how-thoughtfully-integrate-ai-evaluation

Raimondo, E., Ziulu., V., Anuj., H. (2023B, August 20). *Unfulfilled Promises: Using GPT for Synthesis Tasks.* IEG World Bank Group. [Blog post]. Available at: https://ieg.worldbankgroup.org/blog/unfulfilled-promises-using-gpt-synthetic-tasks

Raimondo, E., Ziulu., V., Anuj., H. (2025b). *Balancing innovation and rigor: Guidance for the thoughtful integration of artificial intelligence for evaluation (Guidance Note)*. World Bank. Available at: https://documents1.worldbank.org/curated/en/099136005132515321/pdf/IDU-dccd6e52-4ee3-4294-a264-28fda8a94a49.pdf

Raimondo, E., Ziulu., V., Anuj., H., (2023A, August 23). *Fulfilled Promises: Using GPT for Analytical Analysis Tasks.* IEG World Bank Group. [Blog post]. Available at: https://ieg.worldbankgroup.org/blog/fulfilled-promises-using-gpt-analytical-tasks

Rathore, M. (2024). The Role of Artificial Intelligence in Social Welfare: Harnessing AI For Positive Societal Impact. In *AI in the Social and Business World: A Comprehensive Approach* (pp. 277-293). Bentham Science Publishers.

Rehill, P., & Biddle, N. (2023). Fairness implications of heterogeneous treatment effect estimation with machine learning methods in policy-making. *arXiv preprint arXiv:2309.00805*.

Romberg, J., & Escher, T. (2024). Making sense of citizens' input through artificial intelligence: a review of methods for computational text analysis to support the evaluation of contributions in public participation. *Digital Government: Research and Practice*, *5*(1), 1-30.

RoRI, Research on Research Institute; Newman-Griffis, Denis; Buckley Woods, Helen; Youyou, Wu; Thelwall, Mike; Holm, Jon (2025). Funding by Algorithm - A handbook for

responsible uses of AI and machine learning by research funders (ISBN 978-1-7397102-2-4). Research on Research Institute. Book. https://doi.org/10.6084/m9.figshare.29041715.v1

Russel, S., Perset, K., Grobelnik, M..(2023, 29 November). *Updates to the OECD's definition of an AI system explained*. [Online] Retrieved from: https://oecd.ai/en/wonk/ai-system-definition-update.

Safaei, J. and Longo, J. (2023). *When artificial intelligence meets real public administration: Experimenting with ChatGPT in writing briefing notes*. [online] Available at: https://www.researchgate.net/publication/361152677_When_artificial_intelligence_meets_real_public_administration

Sharma, N. A., Ali, A. S., & Kabir, M. A. (2025). A review of sentiment analysis: tasks, applications, and deep learning techniques. *International journal of data science and analytics*, *19*(3), 351-388.

Sharma, T., Sachdev, P., & Kumari, S. (2025). Unveiling Writing Styles: A Comparative Analysis of AI-Generated and Human Generated Content. *AIJR Proceedings*, 193-212.

Shrestha, S., & Das, S. (2022). Exploring gender biases in ML and AI academic research through systematic literature review. *Frontiers in artificial intelligence*, *5*, p. 976838.

Stephany, F., & Teutloff, O. (2024). What is the price of a skill? The value of complementarity. *Research Policy*, *53*(1), 104898.

Stryker, C., & Holdsworth, J. (2024, August 11). *What is NLP (natural language processing)?* IBM. Available at: https://www.ibm.com/think/topics/natural-language-processing

Susaiyah, A., Härmä, A., & Petković, M. (2023). Schema-Driven Actionable Insight Generation and Smart Recommendation. *arXiv preprint arXiv:2307.13176*.

Thomson Reuters (2025, January 21). *The rise of large language models in automatic evaluation: Why we still need humans in the loop.* Thomson Reuters. [Online] Available at: https://www.thomsonreuters.com/en-us/posts/innovation/the-rise-of-large-language-models-in-automatic-evaluation-why-we-still-need-humans-in-the-loop/

Toloka Team. (2023, August 27). *Difference between AI, ML, LLM, and Generative AI*. Toloka Blog. https://toloka.ai/blog/difference-between-ai-ml-llm-and-generative-ai/

Tony Blair Institute for Global Change. (2024, 20 May). *Governing in the Age of AI – a New model to transform the state.* Institute Global. [Online] Available at: https://institute.global/insights/politics-and-governance/governing-in-the-age-of-ai-a-new-model-to-transform-the-state#a-new-model-of-government-the-impact-of-ai-today

UK Evaluation Society (2025). AI in Evaluation: Good Practice Guidelines for Practitioners. UKES publication.

ul Zahra, A., & Sadiq, A. H. B. (2025). Reimagining gender narratives: A sentiment analysis of evolving representation of female celebrities through x posts. *Contemporary Journal of Social Science Review*, *3*(1), 93-121.

Varker, T., Forbes, D., Dell, L., Weston, A., Merlin, T., Hodson, S., & O'Donnell, M. (2015). Rapid evidence assessment: increasing the transparency of an emerging methodology. *Journal of Evaluation in Clinical Practice, 21*(6), 1199–1204.

Vijayalakshmi, M. C., & Thiyagarajan, M. (2023). Intelligent business insights generation. *Available at SSRN 4457774*.

Wagstaff, T., Lee, S., Jankin, S., Lindsay, N., Gupta, P. and Alfonzetti, M. (2025, January). *The National Development Strategies of Africa and South Asia*. [online] OPML.co.uk. Available at: https://www.opml.co.uk/sites/default/files/2025-04/a6079-national-development-strategies.pdf

Wirjo, A., Calizo Jr., S., Niño Vasquez, G., & San Andres, E. A. (2022, November). *Artificial intelligence in economic policymaking* (APEC Policy Support Unit Report No. 52; APEC# 222-SE-01.18) [PDF]. Asia-Pacific Economic Cooperation. [Online]. Available at: https://www.apec.org/docs/default-source/publications/2022/11/artificial-intelligence-in-economic-policymaking/222_psu_artificial-intelligence-in-economic-policymaking.pdf

Yang, W., Lin, Y., Xue, H., & Wang, J. (2025, April). Research on stock market sentiment analysis and prediction method based on convolutional neural network. In *Proceedings of the 2025 International Conference on Machine Learning and Neural Networks* (pp. 91-96).

Young, Z., & Steele, R. (2022). Empirical evaluation of performance degradation of machine learning-based predictive models–A case study in healthcare information systems. *International Journal of Information Management Data Insights*, *2*(1), 100070.

Zhai, C., Wibowo, S., & Li, L. D. (2024). *The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review*. *Smart Learning Environments, 11*, Article 28. https://doi.org/10.1186/s40561-024-00316-7

APPENDIX

Appendix A: Evaluator Skills and Attitudes for Effective Engagement with AI

The following skills and attitudes support effective and responsible engagement with AI tools in evaluation. They are synthesised from the literature on evaluator competencies in an AI-enabled environment (Cekova et al., 2025; Mason, 2023) and from the practices reported by interview respondents. They are intended as a developmental guide rather than a fixed standard.

Table A1. Evaluator Skills and Attitudes for Effective Engagement with AI

Skill or Attitude	Description
Clear prompting and instruction	Framing precise, well-contextualised prompts and instructions, providing relevant background and worked examples, and refining iteratively.
Critical appraisal and verification	Treating AI outputs as drafts to be checked against source data and domain knowledge, and validating against human-coded results before scaling.
Awareness of limitations	Understanding the failure modes of specific tools, including bias, hallucination, inconsistency and weak performance on synthesis.
Data stewardship	Safeguarding privacy, confidentiality and security when selecting tools and deciding what data may be processed and where.
Ethical and professional judgement	Exercising judgement about appropriate use, including when not to use AI, and upholding research integrity and professional norms.
Maintaining core expertise	Continuing to develop and exercise core analytical skills so that AI complements rather than erodes expertise, particularly for junior staff.
Transparency and documentation	Documenting and disclosing AI use to support trust, auditability and reproducibility.
Continuous learning	Keeping pace with new tools and evidence, and focusing on transferable principles rather than tool-specific routines.

Source: Authors, synthesised from Cekova et al. (2025), Mason (2023) and interview evidence.

Appendix B: Additional Guide for the Responsible use of AI in an Evaluation

This guide operationalises the 10 principles identified, for an individual evaluation. It is intended as an optional aid to be adapted to context and should be revised as tools and evidence develop.

1. Understanding and readiness

- Have we identified the known limitations and failure modes of the tool for this task, including bias, hallucination and inconsistency?
- Have we considered whether AI should not be used for this task, rather than only how it should be used?
- Have we confirmed that evaluators using the tool have the skills needed to prompt, assess and verify outputs appropriately?

2. Data privacy, confidentiality and security

- Does the data include commercially confidential, sensitive organisational or unpublished information, as well as personal data?
- Have we confirmed whether data entered into the tool may be stored, reused or used for model training?
- Are there clear rules on what data can and cannot be entered into external AI tools?

3. Transparency and disclosure

- Have we documented the prompts, tool version, settings and key steps used to produce the outputs?
- Have we recorded where AI materially shaped the analysis, not only where it was used administratively?
- Is the documentation sufficient for another evaluator to understand or reproduce the process?

4. Ethical and legal compliance, with human oversight

- Who is accountable for the AI-assisted outputs and final evaluation judgements?
- Have we checked that AI-generated claims, summaries or classifications can be traced back to the underlying evidence?
- Have we ensured that AI has not introduced unsupported claims, overinterpretation or distortions of participant meaning?
- Have we defined what level of human review is required before AI-assisted outputs can be used in reporting?

5. Validation, quality assurance and traceability

- Have we tested the AI approach on a small representative sample before applying it more widely?
- Have we compared AI outputs with human-coded results or source data before scaling up?
- Have we refined the prompt or process where errors, ambiguities or omissions were identified?
- Have we validated the final prompt or process on a separate sample?
- Can each AI-assisted claim, code, classification or summary be audited back to the source material?

6. Reflection, learning and capacity building

- Have we recorded what worked well, what did not, and where human correction was needed?
- Have we considered whether AI use is supporting evaluator capability or creating overreliance?
- Have we identified any training or quality assurance needs before using AI in similar tasks again?

Appendix C: Tool-specific Considerations: Challenges and Good Practice

This appendix summarises the principal challenges and good practices associated with each category of AI tool discussed in this report. It is intended as a quick reference to accompany the principles in Section 7 and should be read alongside the risks in Section 5.

Table A2: Challenges and good practice for using specific AI tools in evaluation

AI tool	Key challenges	Good practice
Natural Language Processing (NLP)	Sensitivity to language, context and domain terms; risk of misclassification; quality depends on input data.	Validate classifications against human coding; tune to domain vocabulary; use on well-structured tasks such as extraction and categorisation.
Large Language Models (LLMs)	Hallucination and fabricated content; weak performance on evidence synthesis; inconsistency across runs.	Use the prompt-and-validate loop; verify outputs against sources; favour summarisation and coding over synthesis; avoid final decisions.
Generative AI (GenAI)	Stilted or formulaic outputs; limited interpretive nuance; privacy and IP concerns.	Use for drafting and scenario support, not final narrative; edit substantively; control data inputs and disclosure.
Machine Learning (ML)	Bias from training data; opacity of predictions; need for staged implementation.	Test for disparate outcomes; use explainable methods where possible; pilot before scaling; retain human oversight of decisions.
Big Data Analytics (BDA)	Data integration and quality issues; privacy and security risks; infrastructure demands.	Ensure secure data governance; combine with traditional methods; validate insights before use in decisions.

Source: Authors, synthesised from sources cited in Sections 3, 5 and 6.

Appendix D: AI Maturity Level Classification

The AI maturity level recorded for each respondent in Table 1 was assigned by interviewer on the basis of the interview evidence, rather than being self-reported. Each respondent was assessed against five observable indicators evident in the transcripts:

- Breadth of use across the evaluation lifecycle (the number of stages/tasks at which AI is used).
- Tool sophistication (general-purpose assistants, configured tools, or bespoke and in-house systems).
- Workflow integration and reproducibility (ad hoc experimentation, routine use, or systematised and traceable workflows).
- Governance, validation and quality assurance (none, manual checking, or formal verification, training and disclosure).
- Reflexivity and capability (limited awareness, considered working norms, or the ability to design, build or teach).

The resulting levels are ordinal and ascending. Table A3 below sets out the defining characteristics of each level. This ordering is consistent with the breadth of use reported in Table 4, where the Limited respondent reports the narrowest use and the more advanced respondents progressively wider use across the lifecycle.

Table A3: AI maturity level classification used in Table 1

Level	Defining characteristics
Limited	Minimal hands-on use; engagement largely conceptual or confined to peripheral tasks such as occasional summarisation or classification; pronounced caution; general-purpose tools only; no embedded workflows.
Beginner	Active but unsystematic experimentation across a few stages using a single general-purpose assistant; no fixed or repeatable workflows; benefits and limitations still being explored.
Intermediate	Routine use across several stages with off-the-shelf tools; established working norms and human checking; a clear sense of appropriate and inappropriate uses.
Advanced	Broad use across most stages, supported by explicit decision rules (for example volume thresholds), validation against manual results, and organisational understanding of the technology.
Very advanced	Use spanning the full lifecycle, supported by bespoke or in-house systems and by formal governance such as human-verification workflows, training and disclosure requirements.
Expert	Designs and builds novel, traceable AI workflows from first principles; deep technical fluency; able to innovate methodologically and to articulate transferable principles for others.

Source: Authors.

Appendix E: Potential uses of AI Across the Funding Evaluation Cycle

- Choosing intervention areas (project prioritisation)

At the first stage of the funding evaluation cycle, AI tools could provide data-driven inputs that support the identification of key intervention areas. Through advanced analysis of large datasets of both structured and unstructured evaluation data (for example project proposals, reports, survey texts, social media, and sensor or satellite data), AI tools such as NLP enable the identification of patterns and trends that may be difficult, if not impossible, for human analysts to detect within a reasonable timeframe (Cortes et al., 2024; Koliousis et al., 2024). This could assist in the identification and prioritisation of the most critical policy issues to help guide focused policy interventions (Wirjo et al., 2022). Similarly, the ability of AI tools to conduct scenario analysis and forecast policy outcomes can help in identifying patterns and anticipating potential challenges (Patel et al., 2021), and in estimating the expected costs and benefits of different policy options (Wirjo et al., 2022).

- Assessing project proposals

AI-driven analysis could generate data insights to support policy decision-making during the project proposal assessment stage (Patel et al., 2021). A recent policy report from the UK Evaluation Task Force explores the potential of AI systems to assist in reviewing grant applications and the potential use of LLM-based tools to analyse a high volume of documents (Evaluation Task Force, 2025). In the administrative stage, NLP could help identify underdeveloped project proposals before they reach the assessment stage (Cortes et al., 2024) and assist initial quality control (Kousha & Thelwall, 2022). Machine Learning could detect administrative errors and improve accuracy (Young et al., 2022), support the detection of duplicate proposals, group proposals thematically (Jasper et al., 2019; Romberg & Escher, 2024), and identify the best reviewers for proposals based on topic classifications (Kousha & Thelwall, 2022). During the assessment stage, Artificial Neural Networks could assist in optimising decisions by evaluating multiple criteria simultaneously (Huang et al., 2022; Koliousis et al., 2024), Cognitive Computing Decision Support Systems could facilitate rational decision-making (Behera et al., 2023), and ML tools could contribute to fairer decisions by appropriately handling sensitive variables such as race and gender (Rehill & Biddle, 2023).

- Project implementation, monitoring and process evaluation

AI tools, either through generative AI or narrow AI (statistical AI, NLP or computer vision), could support input for an intervention, continuous monitoring, risk and trend identification, predictive analysis, and compliance checks (Tony Blair Institute for Global Change, 2024; ALP Consulting, 2025). For instance, AI tools allow automation and real-time monitoring of ongoing projects, which allows funders to spot delays, anomalies or feedback as the project occurs (Patel et al., 2021). This real-time analysis could support correction and early intervention to improve effectiveness (ALP Consulting, 2025). Explainable Artificial Intelligence methods may enhance transparency and support the generation of actionable insights aligned with specific

regulatory frameworks (de Carvalho & da Silva, 2021). Using ML and NLP of automated data extracted from social media, mobile devices, sensors and satellite imagery, AI could be used for real-time data collection, analysis and evaluation through methods such as sentiment analysis (Yang et al., 2025), which aims to determine the emotional tone or opinion expressed in text data (Mejova, 2009; Mukherjee & Bhattacharyya, 2013; Sharma et al., 2025).

- Impact evaluation and value for money evaluation

AI tools have the potential to transform impact evaluation by speeding up traditional methods, enabling new types of impact evaluation, aiding the visualisation of impact and generating actionable insights. First, AI could provide accurate and faster analysis of the impact of policies (Wirjo et al., 2022). Generative AI could automate the analysis of large volumes of real-world data, enabling a comprehensive assessment of the broader impact of funded projects (Fleurence et al., 2024), and could help automate reporting and visualisation (Tony Blair Institute for Global Change, 2024). Second, AI could enable new methods using new types of data. AI tools could extract and analyse citizens' arguments and opinions (Romberg & Escher, 2024); Convolutional Neural Networks are effective in capturing public sentiment (Yang et al., 2025); and neural networks could assist in calculating context-specific cost-effectiveness of policies (Mannarini et al., 2022). Third, AI could be used for generating actionable insights across various domains (Vijayalakshmi & Thiyagarajan, 2023; Gupta et al., 2020; Asunmonu, 2025; Pencheva et al., 2020; Rathore, 2024; De Carvalho & da Silva, 2021).

Table A4. Potential benefits of specific AI tools for evaluation

AI approach / tool	Potential benefits
Natural Language Processing (NLP)	Rapid identification, extraction, categorisation and standardisation of information from unstructured texts; automates and speeds up literature reviews, proposal screening and qualitative analysis; enhances trend and text analysis (for example sentiment analysis) and topic detection.
Large Language Models (LLMs)	Enables in-depth, scalable text analysis; can generate survey questions, proposals or literature reviews; can rapidly summarise and translate research and policy documents and identify or synthesise findings quickly.
Generative AI (GenAI)	Automates stakeholder communications, proposal triage and monitoring dashboards; drafts reports and supports scenario generation; personalises assessments and simulations; accelerates production of synthetic data for model training or scenario analysis.
Machine Learning (ML)	Enables real-time, adaptive analysis of quantitative and qualitative data; supports dynamic prioritisation and predictive capacity for proposal success, implementation risks and impact projections; reduces subjectivity and increases consistency relative to manual review.
Big Data Analytics (BDA)	Integrates data from multiple sources for holistic project assessment; enables real-time monitoring, pattern and risk detection; supports evidence-based decision making and efficient resource allocation; improves predictive modelling for impact evaluation.

Sources: See corresponding sources in text above and Table 2

www.ircaucus.ac.uk

Email info@ircaucus.ac.uk Twitter @IRCaucus

INNOVATION & RESEARCH CAUCUS

Delivered with ESRC and Innovate UK