INNOVATION & RESEARCH CAUCUS

# USING ARTIFICIAL INTELLIGENCE IN EVALUATION:

## A Rapid Evidence Review and Proposed Guidelines.

IRC Report No: 039

REPORT PREPARED BY

**Dr Ully-Yunita Nafizah**
Warwick University

**Dr Hamisu Salihu**
Warwick University

**Dr Halima Jibril**
Warwick University

Delivered with ESRC and Innovate UK

# CONTENTS

**Authors**

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

- Dr Ully-Yunita Nafizah – Warwick University
- Dr Hamisu Salihu – Warwick University
- Dr Halima Jibril – Warwick University

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 UK Research & Innovation (UKRI). We are very grateful to the project sponsors at UK Research & 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 UKRI.

We would also like to acknowledge and appreciate the efforts of the IRC Project Administration Team involved in proofreading and formatting, for their meticulous attention to detail and support.

**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 IUK, 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: Nafizah, UY., Salihu. H. and Jibril, H. November 2025. *Using Artificial Intelligence in Evaluation: A Rapid Evidence Review and Guidelines Development.* Oxford, UK: Innovation and Research Caucus

# Executive summary

This report presents a rapid review of the evidence base on the use of Artificial Intelligence (AI) tools in policy evaluation. It examines current and potential applications of AI globally and across different types of funding organisations and sectors. The report aims to provide UKRI with evidence-based insights to inform appropriate guidelines for the responsible and effective integration of AI into evaluation processes. The report draws on academic and grey literature 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.

## Use cases and reported benefits of AI in evaluation

The review found notable examples of AI across different phases of the funding evaluation cycle (Section 3). Here we highlight five areas where 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.

Given the limited evidence base, there is no clear consensus supporting a definitive shift to AI tools for specific evaluation functions. Most use cases remain experimental, with effectiveness and safety of AI tools dependent on context and adherence to best practices (outlined in Section 5). Human judgement remains essential in all decision-making functions.

## Risks and challenges of using AI in evaluation

Using AI tools in policy evaluation carries significant risks and challenges. We identified structural challenges which include 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.

## How can UKRI create the conditions for safe, transparent and responsible use of AI in evaluation?

Because the existing evidence on the effectiveness of AI in policy evaluation contexts is sparse, and comes predominantly from contexts outside R&I funding, this report recommends a measured and evidence-based approach to any integration of AI into UKRI evaluation activities. UKRI should consider:

1. Defining a clear but adaptable framework outlining appropriate use cases for AI within UKRI evaluations.
2. Adopting an experimental approach through carefully designed policy experiments and pilot schemes that test specific applications, capture any challenges, and generate lessons for wider adoption. This should preferably be done building on a stronger evidence base, for example after Phase 2 of

this project when evaluator interviews will provide richer, context-specific insights to inform any UKRI experimentation.

3. Ensuring data security and developing, where feasible, UKRI-specific internal AI systems and platforms.
4. Providing clear accountability structures for all AI-supported outputs in cases of unintended errors or harms arising from AI use.
5. Providing a framework that ensures strong human oversight and ethical safeguards, ensuring that AI tools complement, rather than replace, human judgement
6. Providing clear disclosure norms that incentivise transparency and accurate reporting of AI use (e.g., emphasising that disclosure enhances trust)
7. Maintaining open communication with evaluators on AI use throughout the evaluation process and creating opportunities for learnings and feedback loops
8. Investing in AI literacy and skills training for evaluators and periodically updating standards to reflect emerging best practices.

In addition to these governance and oversight responsibilities for UKRI, the review synthesises best practices and risk mitigation measures for UKRI evaluators including i) understanding of AI tools and assessing utilisation readiness ii) ensuring data privacy, confidentiality and security iii) ensuring transparency and compliance with ethical and legal frameworks iv) ensuring scientific rigour and reliability v) reflection, learning and capacity building, as well as guidelines related to the use of specific AI tools. These are detailed as Evaluator Checklists in Section 6.

## Gaps in the evidence base

The overall evidence base on the use of AI in evaluation is sparse and emerging, but there is even less evidence in the specific context of evaluating Research and Innovation policies and programmes in the UK and internationally. There is also limited discussion in the literature about the extent of transparency and disclosure of AI use in evaluation; most of the reviewed studies set out to explicitly incorporate and test the use of AI. We still need to understand more about the extent, benefits and challenges of 'everyday' adoption of AI tools in evaluation contexts. We also found little evidence on the use of AI to design evaluations.

## Next steps

To enhance our understanding of how AI tools are being used in evaluation, there is scope for conducting semi-structured interviews with staff in organisations actively trialling AI tools, consultancy firms experienced in their application, and evaluators within UKRI's own portfolio. These interviews could explore which tools are being used and for what evaluation tasks, the benefits gained compared to manual methods, the challenges and risks encountered, and how these are managed in relation to ethical and legal standards, including any policy and governance aspects within institutions. Interviews could also explore gaps in evaluators' skills and awareness, and examine attitudes around disclosing the use of AI tools, perceptions of how AI-produced outputs are received, and why evaluators may avoid disclosure.

# 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” (Russel et al., 2023). The advent of AI and 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; 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 (AOED, 2022; Wirjo et al., 2022; Yar et al., 2024). 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 and ethical and legal safeguards when AI is introduced into evaluation workflows.

This report presents a rapid review of the evidence on the use of AI tools in policy evaluation. The primary objective of the report is to provide UKRI with evidence-based insights to inform the development of appropriate guidelines for the responsible and effective integration of AI into its evaluation processes. To that end the review seeks to address 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? How do these vary by use cases and specific AI tools?[1]
4. What are the emerging best practices in the use of AI for evaluation?

The report proposes a funding evaluation cycle, based on the policy-making cycle (Cairney, 2023; IfG, 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[2]. The latter is the primary focus of the report, but earlier phases are considered to give a fuller picture of the potential benefits and risks of AI use across evaluation-related activities. Due to limited evidence on R&I evaluation contexts, the review draws widely from different policy contexts, including in development and trade policy. It combines insights from academic and grey literature and examines UK and international contexts.

The remainder of the report is organised as follows. Section 2 provides details of our approach to the evidence review, including our methodology, the scope of the review, and definitions of AI tools. It also introduces the funding evaluation cycle around which we organise the evidence. Section 3 provides findings from the evidence review on the use of AI across different stages of the funding evaluation cycle, highlighting both potential and actual use cases of specific AI tools. Section 4 discusses the key risks and challenges associated with AI in evaluation. Section 5 outlines emerging best practices and guidelines for effectively integrating specific AI tools. Section 6 synthesises the evidence to provide initial guidelines to support UKRI and its evaluators in the responsible use of AI. Section 7 summarises the findings and sets out next steps.

[1] A cost benefit analysis of AI use is outside the scope of this review.
[2] The use of AI in the implementation of evaluation findings is outside the scope of this review

# 2. Approach to the evidence review

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

## 2.1 AI-assisted literature discovery exercise

### 2.1.1 AI tool and motivation for use

To experiment and engage with the review's topic area, i.e., AI, we started 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 summarize relevant scholarly publications using semantic similarity techniques rather than purely keyword-based search engines. Our aim in doing this was to test and report, given the topic of the review, whether and how specialist AI tools can be useful in a research context relevant to many policy evaluation exercises (i.e., evidence reviews).

### 2.1.2 AI Prompt and output

We asked Elicit AI to conduct a review of academic articles, using its mid- advanced function, based on the simple 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, out of which it retained 25 of the most relevant articles after screening for articles that have 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?", and obtained similar outputs with 25 additional articles retained, making a total of 50 articles across both searches.

### 2.1.3 AI tool usefulness

Although the 50 studies included in the Elicit AI's reviews seemed relevant at first, only a few of them were ultimately included in this review: comparing our final reference list to that of Elicit AI returned 10 matches, representing about 8% of our reference list. This is because, upon further manual reviewing of the articles, we found that many of them focused on the *potential* uses of AI and in policy domains *outside of that related to economic and business programme interventions;* instead, the policy domains covered were predominantly in healthcare, electricity regulation and education. These articles may have been considered relevant for this review if they had focused on *actual* use of AI tools in policy evaluations.

The Elicit AI reviews were however 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 and engage with real time AI experimentations by evaluators. Thus, by returning only a few relevant articles, Elicit AI's outputs enhanced the efficiency of our manual review process.

## 2.2 Flexible rapid literature review approach:

Following our experimentation with Elicit AI which suggested the futility of gaining relevant insights from methods that rely exclusively on databases of peer-reviewed academic publications (such as systematic literature review methodologies) 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. We used 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 (e.g. World Bank, OECD and UN), evaluation networks, and government departments, to capture policy reports, guidance documents, and technical notes on the subject matter. Thus, recognising that much of the debate and innovation around AI in evaluation sits outside traditional academic publishing, we also

included grey literature. This encompassed 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/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 the snowballing techniques, 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 (adapted from the policy-making cycle) as an organising framework (see below).

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 inevitably missed some relevant studies or case examples (Gannan et al., 2010). The lack of standardized methodology during the process presents challenges for reproducibility relative to systematic literature reviews (Varker et al., 2015). However, our close collaboration with policymakers as end-users helps to strengthen the quality and relevance of the rapid evidence review, helping to balance some of its limitations (Raghunathan et al., 2022).

## 2.3. Scope of the 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, and our broader scope include studies with AI applications in

other funding and evaluation related activities including choosing intervention areas, proposal assessment, or monitoring and data collection; this informed the development of the funding evaluation cycle through which we organise the evidence. We limited the review to materials published in English.

## 2.4. The funding evaluation cycle

Here we introduce a funding evaluation cycle as a framework for organising evidence on the use of AI in evaluation (see Figure 1). 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), we develop the funding evaluation cycle which illustrates how AI can be integrated throughout the funding and evaluation process. The stages are:

- *Choosing Intervention Areas*

The first stage in the policy-making cycle 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. This process 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). They also help identify early potential problems or deviations and generate insights and lessons for future initiatives.

» *Impact Evaluation and Value for Money Evaluation*

Impact evaluation is defined as 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, meanwhile, assesses whether an intervention has used public resources efficiently, effectively, and economically to achieve its outcomes (HM Treasury, 2022). The results help policymakers understand which interventions deliver measurable benefits and inform future resource allocation and policy design.

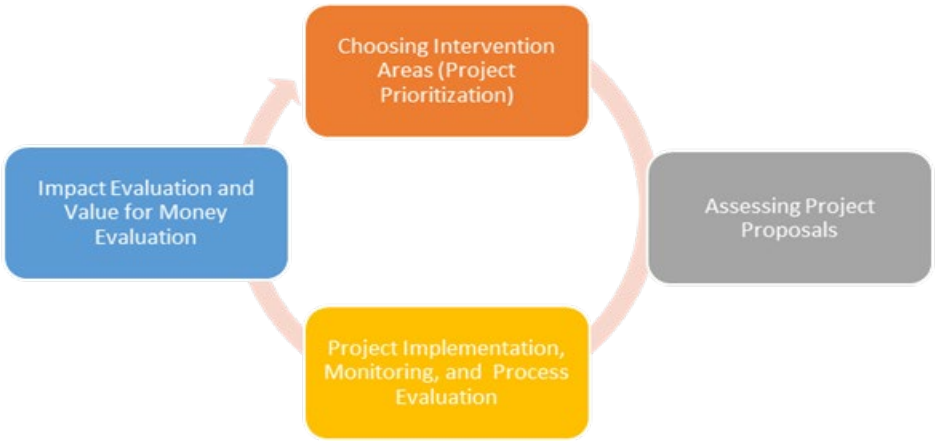


**Figure 1 Evaluation Cycle**

## 2.5. 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. Examples include: Grammarly and LanguageTool for grammar assistance; Google Translate for machine translation; and IBM Watson Assistant for customer-service 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, 2025). Evaluation examples include using GPT-4 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 (e.g., DALL-E), 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. Tools such as scikit-learn and TensorFlow support risk modelling, classification, and anomaly detection. 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 (Jasper et al., 2019).

### 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: Evidence on the use of AI in evaluation.

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

We thus present the findings of the evidence review in two parts:

- *Potential uses of AI* across the funding evaluation cycle: A large portion of the academic literature on the use of AI in evaluation focuses on the potential benefits of various AI tools, outlining how and why organisations might adopt them in funding decisions or in assessing funding impact. Some of these potential applications are already evidenced in practice (as discussed in later sections), while others remain largely theoretical. Given the rapidly evolving nature of AI, we review these potential uses to offer UKRI a forward-looking perspective - one that highlights not only what has been done to date but also what may be possible to explore in the future.
- *Actual use cases of AI* across the evaluation cycle: Here we discuss the rarer cases, predominantly from grey literature, which document how AI tools are being used in different stages of the evaluations and their effectiveness. Since most of the use cases are outside the R&I funding context, we outline in each

case the **potential learnings and implications of the evidence for AI use in UKRI evaluations**, focusing here on applications with the strongest evidence on effectiveness.

## 3.1 Potential uses of AI across the funding evaluation cycle

### 3.1.1. Choosing Intervention Areas (Project Prioritization)

At the first stage of funding evaluation cycle, AI tools could support in providing data-driven inputs that support the identification of key intervention areas. Through advanced analysis from a large dataset of both structured and unstructured evaluation data (e.g., project proposals, reports, survey texts, social media, sensor/satellite data etc), 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 (Cortés 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). AI tools could also help in estimating the expected costs and benefits of different policy options, which aids in determining the potential effects of suggested interventions (Wirjo et al., 2022).

### 3.1.2. 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). For instance, 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 Large Language model (LLM)-based tools to analyse a high volume of documents (Evaluation Task Force, 2025). Here, therefore, the role of AI could span the administrative and assessment phases of the proposal selection process.

In the administrative stage, Natural Language Processing (NLP) could help identify underdeveloped project proposals before they reach the assessment stage, allowing for their early removal from the selection process (Cortés et al., 2024); this could assist the initial quality control process for proposal assessment (Kousha and Thelwall, 2022). Similarly, Machine Learning could be used to detect administrative errors and to improve the accuracy of the administrative stage (Young et al., 2022). AI could also support detecting duplication of proposals, grouping proposals thematically (Jasper et al., 2019; Romberg and Escher, 2023), and identifying the best reviewers for project proposals based on topic classifications (Kousha and Thellwall, 2022). These functions may improve the efficiency and speed of proposal evaluations (Cortés et al., 2024).

During the assessment stage, Artificial Neural Networks (ANN) could assist in optimizing decisions by evaluating multiple criteria simultaneously (Huang et al., 2022; Koliousis et al., 2024). Likewise, Cognitive Computing Decision Support Systems (CCDS) could facilitate rational decision-making by leveraging two cognitive processes: the automatic system, which enables rapid pattern recognition, and the reflective system, which supports in-depth analytical reasoning through scenario analysis (Behera et al., 2023). In addition, Machine Learning tools could contribute to producing fairer decisions by appropriately handling sensitive variables such as race and gender using tools like the AI Fairness model (Rehill and Biddle, 2023).

### 3.1.3. Project Implementation, Monitoring and Process Evaluation

AI tools, either through generative AI or narrow AI (i.e., statistical AI, Natural Language Processing, or Computer Vision), could support in feeding input for an intervention, continuous monitoring, risk and trends 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 on-going projects, which allow funders to spot delays, anomalies, or feedback as the project occurs (Patel et al., 2021). This real-time analysis could help in performing real-time correction and early interventions to improve effectiveness of the on-going funded projects (Patel et al., 2021; ALP Consulting, 2025). This way, Machine Learning could detect irregularities and forecast potential project outcomes. Explainable Artificial Intelligence (XAI) methods

may enhance transparency and support the generation of actionable insights aligned with specific regulatory frameworks (de Carvalho and de Silva, 2021).

Using Machine Learning algorithms and Natural Language Processing (NLP) of automated data extracted from social media, mobile devices, sensors, and satellite imagery, AI could be used for real-time data collection, analyses, and evaluation using such evaluation methods as sentiment analysis (Yang et al., 2025). Also known as "subjectivity analysis", "opinion mining", and "appraisal extraction", sentiment analysis is "a Natural Language Processing and information extraction task that aims to obtain writer's feelings expressed in positive or negative comments, questions and requests, by analysing a large number of documents." (Mukherjee, S., & Bhattacharyya, 2013). It is a process used to determine the emotional tone or opinion expressed in text data, such as reviews, social media posts, or survey responses (Mejova, 2009; Sharma et al., 2025).

### 3.1.4. 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 visualizing of impact and generating actionable insights.

First, AI could advance the evaluation stage of funding evaluation cycle by providing accurate and faster analysis on the impact of policies (Wirjo et al., 2022). Generative AI, in particular, 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). In addition, AI could help in automating reporting and visualizing the summary of findings and insights (Tony Blair Institute for Global Change, 2024). As discussed previously, through aiding programme monitoring, ML tools could provide accurate data that could help in measuring the impact of policies (Wirjo et al., 2022).

Second, AI tools could enable new methods of impact evaluation using new types of data. For instance, AI tools could extract and analyse citizens' arguments and opinions, providing valuable insights into the perceived contributions of projects (Romberg and Escher, 2023). Additionally, Convolutional Neural Networks (CNNs) are effective in capturing public sentiment, which could be applied to policy evaluation (Yang et al., 2025). Furthermore, neural networks could assist in calculating context specific (e.g.,

country level or regional) cost-effectiveness of policies by analysing their costs and benefits (Mannarinni et al., 2022).

Third, AI could be used for generating actionable insights across various domains. Through analysing large, diverse datasets, AI-powered systems could provide real-time business and public policy insights (Vijayalakshmi & Thiyagarajan,2023). For instance, digital analytics frameworks have been used to generate consumer insights and create value-based outcomes (Gupta et al.2020). Also, Natural Language Processing and Machine Learning techniques allow for the extraction of client- or consumer-generated actionable insights for innovation (Asunmonu,2025), improving mutual government-citizen understanding (Pencheva et al., 2020), and overall social welfare (Rathore, 2024). Explainable AI could generate insights that improve evidence-based policies (De Carvalho and da Silva, 2021).

## 3.2. Evidence from use cases of AI across the evaluation cycle

### 3.2.1. Choosing Intervention Areas (Project Prioritization)

At the first stage of the funding evaluation cycle, the evidence suggests that AI tools support in providing data-driven inputs that support the identification of key intervention areas.

For example, an APEC Policy Brief by Wirjo et al. (2022) highlights the use of Natural Language Processing (NLP) technology developed by CitizenLab in Belgium, which helps civil servants process large volumes of data from digital public participation platforms. This tool enables the classification of public input and the clustering of similar contributions based on themes, demographic profiles, or geographic locations (OPSI, n.d. in Wirjo et al. (2022)). Since its launch in 2018, this feature has influenced several local administrations by providing automated analyses that strengthened their connection with citizens. One example is the city of Temse, which engaged residents in discussions about mobility and visualized their suggestions on a city map. This approach allowed the administration to pinpoint critical problem areas and better decide where to allocate resources. In the UK, the Department for Science, Innovation and Technology recently 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; the tool is reported to have achieved this within two hours while matching human accuracy, with positive implications for efficiency (DSIT, 2025).

Another example from Wirjo et al. (2022) highlights 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)). Similarly, Patel et al., (2021) highlights how the Victorian State Government in Australia uses a 'syndromic surveillance' programme by combining automated data capture with NLP to monitor reported symptoms and patient characteristics in hospitals. Here, AI serves as an early warning tool for detecting emerging public health issues.

#### Implications for UKRI

UKRI strategy teams may be engaged in setting funding agendas and determining the types of support provided to recipient organisations. Funding priorities often shift in response to changes in national policy. The evidence above suggests that AI tools can effectively support UKRI in agenda setting through the use of 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.

### 3.2.2. Assessing Project Proposals

Although, as discussed on Section 3.1.2, the literature highlights several potential uses of AI tools in the administrative and assessment stages of funding proposals, we found limited evidence of actual use. A use case from the Independent Evaluation Group (IEG) at World Bank highlights one successful experiment related to Generative Artificial Intelligence (Gen-AI) and Generative Pre-trained Transformer (GPT) for evaluation practices, particularly to conduct simple classification of proposals (Raimondo et al., 2023A). They tested both ChatGPT and the World Bank's enterprise version, m-AI (powered by GPT-3.5), to classify text data related to disaster risk reduction. When

comparing the AI results to manual classifications, ChatGPT achieved over 76% accuracy, while m-AI reached 57%.

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).

#### 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 of using LLMs, ML and GenAI to assist with pre-screening proposals against eligibility criteria, conducting simple thematic classifications. These applications, if appropriately implemented, could reduce administrative burden and streamline internal UKRI processes.

However, given ethical, legal and reliability challenges outlined in later sections, using AI tools for actual assessment and decision-making regarding grant outcomes should be avoided.

### 3.2.3. Project Implementation, Monitoring and Process Evaluation

Some emerging use cases illustrate AI application in programme monitoring. An example from Tony Blair Institute for Global Change (2024) shows how the UK NHS created the NHS Early Warning System during the Covid-19 pandemic to keep track of real-time and predicted patient demand and resource capacity, even down to the availability of specific beds. This allowed for real-time monitoring and proactive decision-making.

Other use cases illustrate AI applications that are not related to policy monitoring but are relevant to broader real-time monitoring efforts. For instance, AI tools using Machine Learning and NLP allow automation for compliance monitoring in finance by detecting risks, automating audits, and enforcing regulatory policies (Atlan, 2025). Another example is that of HCLTech's automation of gaming reviews for a global technology company, using Gen-AI to automate data collection and conduct sentiment analysis; this resulted in "a 70% reduction in manual efforts" as well as improvements in accuracy, optimisation of resources, reduction of turnaround time and refinement of program-wide complexity (HCLTech, n.d). Although this use case is from the private sector, the same technique could be applied to monitor large-scale citizen sentiment analysis for feedback in public policy context.

#### 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. It may therefore 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 could be employed to automate the collection and analysis of this data. This could provide real-time insights into which aspects of the programme are performing well, which may require improvement, and whether broader programme adjustments are needed to improve the likelihood of achieving intended objectives.

### 3.2.4. Impact Evaluation and Value for Money Evaluation

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

#### *AI use in quantitative impact analysis*

For quantitative impact analysis, Wirjo et al., (2022) highlights how World Bank developed a Machine Learning algorithm to quantify and evaluate the impact of trade agreements on trade flows (Breinlich et al., 2021). The algorithm enables data-driven

methods in identifying which intervention most strongly influences trade flows and quantifying the marginal impact of each selected intervention on trade outcomes. This allowed the estimation of specific features of trade agreements that drive better trade outcomes.

The same report also highlights the use case of AI to evaluate climate-related policies in the United Kingdom (Abrella et al., 2021). In particular, the use case highlights how the use of Machine Learning methods, i.e., causal forests, help in estimating heterogeneous treatment effects to reveal how policy effectiveness varies across different regions and economic context. This allowed better understanding on where and for whom carbon pricing works best for more targeted and effective climate design policy.

Another use case from World Bank IEG highlights how ChatGPT can be used to conduct econometric analysis to analyse the association between World Bank interventions and desired outcomes in the context of the World Bank's economic response to the pandemic (Raimondo et al., 2023A). They found that AI was effective at generating code, which made it easier to replicate the study's results.

***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.

***AI use in summaries, syntheses and developing 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[3], coherence[4], and faithfulness[5] of the generated text*".[6] 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

[3] Relevance measures whether the selected content from the source is the most important content following the prompt.
[4] Coherence measures the overall collective quality of the sentences: The response text should be built from sentence to sentence to a coherent body of information about a topic.
[5] Faithfulness measures whether the information generated is factually consistent with the information in the source or not.
[6] Details of the World Bank IEG guidance note and scoring models can be found here.

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).

### Implications for UKRI

In impact evaluation, the evidence suggests UKRI could effectively use ML and Gen-AI tools to support quantitative impact analysis, particularly the standardisation of software code to enable replication of econometric impact analysis. These use cases are well evidenced by the World Bank Group. UKRI may also effectively employ Gen-AI and other LLM-based tools to conduct qualitative impact assessments, such as using them to collect and analyse interview data, as demonstrated in experiments by Chopra and Haaland (2023) and the Behavioural Insights Team (BIT, 2023), and to conduct summaries of large documents.

Based on the evidence, it is not currently recommended that AI tools (Gen-AI and LLMs in particular) be used to produce evidence syntheses or evaluation reports. Persistent issues such as hallucination and the generation of inaccurate text mean these outputs require substantial human oversight and revision.

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 on the use of AI shows its effectiveness in horizon scanning activities, administrative stages of proposal assessments, real time monitoring and data collection, qualitative and quantitative impact analysis and producing summaries of large documents. Evidence syntheses and wider decision-making functions are fewer effective areas in which to incorporate AI. This list of functions is not exhaustive and experimentation will be required to identify other areas of responsible, safe and effective use. An experiment conducted among Boston Consulting Group employees found that, in activities undertaken by consultants that involve realistic, knowledge-intensive tasks (such as aspects of policy evaluation), AI improved performance when used within its "known capabilities." Specifically, consultants using AI completed 12% more tasks, 25% faster, and with 40% higher quality. However, in tasks outside AI's known capabilities, consultants who did not use AI made significantly fewer mistakes. While this evidence is interesting, it remains vague about which specific activities fall within the domain of AI's "known capabilities". Still, it suggests the potential to uncover different ways of using AI within UKRI evaluation contexts.

Table 1 summarises the use cases we identified for specific AI tools across the funding evaluation cycle and Table 2 presents the potential benefits associated with the use of specific tools.

**Table 1 Summary of Evidence on the Use of Specific AI Tools and Relevant Use Case**

| Evaluation Cycle Phase | AI Tools | Use Case | Challenges/ Limitations (If any) | Sources |
|---|---|---|---|---|
| Choosing Intervention Areas (Project Prioritization) | Natural Language Processing (NLP) | Helps civil servants process large volumes of data from digital participation platforms in Belgium | » Classification of algorithms: the classification needs to support multiple languages and different administrative workflows<br>» Human adoption: maintain clear work flow policy | (OPSI, n.d.) in Wirjo et al., (2022) |
| Choosing Intervention Areas (Project Prioritization) | Big Data Analytics | Identify issues and behavior trends in the urban environment in Bulgaria | - | (Policy Cloud, n.d.) in Wirjo et al., (2022) |
| Assessing project proposal | Generative Artificial Intelligence (Gen-AI, GPT) | Classify text data | » Limitation to tackle more complex tasks | (Raimondo et al., 2023A) |
| Project Implementation, Monitoring, and Process Evaluation | Machine Learning and Big data Analytics | Keep track of real time demand and resources capacity in the context of Covid-19 in the UK | » Tackle different rules and regulations from each department<br>» Address privacy concern | (Tony Blair Institute for Global Change, 2024) |
| Project Implementation, Monitoring, and Process Evaluation | Machine Learning and Natural Language Processing (NLP) | Automation for compliance monitoring in finance | » Potential bias in AI model outcomes<br>» Concerns about data privacy and security<br>» Integration with legacy system<br>» High initial implementation cost | (Atlan, 2025) |

| Evaluation Cycle Phase | AI Tools | Use Case | Challenges/ Limitations (If any) | Sources |
|---|---|---|---|---|
| | | | » Tighter regulations and emerging AI-specific policies<br>» Ineffective metadata management processes | |
| Impact Evaluation | Machine Learning | Quantify and evaluate the impact of trade agreements on trade flow | » | (Breinlich et al., 2021) |
| Impact Evaluation | Machine Learning | Estimate heterogeneous treatment effect to evaluate carbon pricing policy effectiveness | » | (Abrella et al., 2021) |
| Impact Evaluation | Large Language Models (LLMs) | Conduct qualitative interviews. | » Design Challenge<br>» Potential algorithmic bias | (Chopra and Haaland, 2023) |
| Impact Evaluation | Large Language Models (LLMs) | Categorize qualitative interview responses | » AI limitations to deal with complex and high-level topics<br>» Difficult to standardize research-level differences | (BIT, 2023) |
| Impact Evaluation | Generative Artificial Intelligence (Gen-AI, GPT) | Analyse the associations between interventions and desired outcomes using econometric model | » | (Raimondo et al., 2023A) |

| Evaluation Cycle Phase | AI Tools | Use Case | Challenges/ Limitations (If any) | Sources |
|---|---|---|---|---|
| Impact Evaluation | Generative Artificial Intelligence (Gen-AI, GPT) | Conduct sentiment analysis by analyzing positive or negative association | » | (Raimondo et al., 2023A) |
| Impact Evaluation | Generative Artificial Intelligence (Gen-AI, GPT) | Conduct evaluation synthesis | » Limitation to synthesize evidence from multiple sources | (Raimondo et al., 2023B) |
| Impact Evaluation | Generative Artificial Intelligence (Gen-AI, GPT) | Produce high-level summaries of evaluation documents | » Limitation to synthesize evidence from multiple sources | (Raimondo et al., 2023A) |

Notes: 1) Not all use cases are directly relevant to R&I funding evaluation. 2) Table based on reports that identified the AI tool used.

**Table 2: Potential Benefits of Specific AI Tools for Evaluation**

| AI Approach/Tools | Benefits |
|---|---|
| **Natural Language Processing (NLP)** e.g., Google Translate, BioBART and Chatbots. | » Rapid identification, extraction, categorisation, and/or standardisation of data/information/insights from unstructured texts/documents (e.g., reports, surveys, feedback)<br>» Automates and speeds up literature reviews, proposal screening/evaluation, and richer qualitative analysis.<br>» Enhanced and automatic trend and text analyses (e.g., sentiment analysis in feedback) and topic detection. |
| **Large Language Models (LLMs)** e.g., GPT-4, Claude, LLama, BARD, and Gemini. | » Enables in-depth, scalable, large-scale text analyses.<br>» Can generate survey questions, proposals or literature reviews.<br>» Can rapidly summarise, translate research and policy documents (e.g. evaluation reports) and/or identify/synthesise findings quickly. |
| **Generative AI (GenAI)** e.g. Dovetail AI (Qual Research Insight Assistant), Klarity AI (Contract Review Assistant), Jasper AI (Marketing), DALL-E (images), Midjourney (images), ChatGPT (text), Copilot (code) | » Automates stakeholder communications, proposal triage, chatbot-based monitoring dashboards,<br>» Drafts report, and scenario generation<br>» Personalizes assessments and simulations (e.g., for training or engagement)<br>» Accelerates production of synthetic data for model training or scenario analysis |
| **Machine Learning** | » Enables real-time, adaptive analysis of quantitative and qualitative data<br>» Enables dynamic prioritization and enhances predictive capacity for proposal success, implementation risks, impact projections<br>» Reduces subjectivity and increases consistency compared to manual review |
| **Big Data Analytics** | » Integrates data from multiple sources for holistic project assessment<br>» Enables real-time monitoring, pattern, and risk detection<br>» Supports evidence-based decision making and efficient resource allocation.<br>» Improves predictive modelling for impact evaluation. |

Sources: See corresponding sources in Table 3.

# 4. Key risks and challenges of AI use across the evaluation cycle

Despite the considerable potential in the use of AI for evaluation purposes, the value or usefulness of AI depends on how well it is complemented by human agents or workers (Stephany & Teutloff, 2024). Moreover, there are inherent limitations, risks and challenges in using AI for evaluation and related purposes. These challenges must be understood and managed carefully so that policymakers can integrate AI effectively and responsibly in evaluation. Hence, in this section, we explore the risks and challenges associated with the use of AI in line with the UK Government Guidelines on AI use (Government Digital Service, 2025). In particular, the first three principles outlined in the Guidelines are relevant here; before using AI, civil servants and government organisations should: (i) know what AI is and what its limitations are, (ii) use AI lawfully, ethically and responsibly, and (iii) know how to use AI securely.

## 4.1 Structural risks and challenges

### 4.1.1. AI and Issues of Equity:

One of the goals of policymakers relate to issues of promotion of equity and social justice. Here, AI performs less effectively because policy-relevant concepts such as fairness, justice, and equity are issues that are inherently human in nature. AI tools are trained with data from the past, with unintentional bias rules which could penalise under-represented groups (Checco et al., 2021), and lead to the enforcement of existing discriminatory practices. This calls for caution on AI use in all aspects of the evaluation cycle.

A new field, AI Fairness, has developed to explore and offer mitigation strategies for *"the harms that can be done (particularly to already marginalised groups) by employing AI systems to make decisions*" (Rehill & Biddle, 2023, p.3). This is due to the increasing recognition that AI's ability to understand and interpret human realities, causality and cultural subtleties "*remains limited*" (Wirjo *et al*., 2022). These biases stem from various other factors such as a lack of diversity in datasets and in AI tool development teams, as well as existing societal biases and algorithmic design (Nadeem *et al*., 2020; Liu, 2024).

Studies have explored AI bias detection methods (Shrestha & Das, 2022), and mitigation strategies, such as implementing fairness in AI development, increasing diversity in data and teams, and improving training processes (Nadeem *et al*., 2020; Liu, 2024). However, addressing biases inherent in AI is complex and often involves trade-offs between fairness metrics and model accuracy (Aninze & Bhogal, 2024). Hence, researchers have emphasized the need for a multifaceted approach, including standardization, diverse representation in AI development, and understanding historical and political factors contributing to bias (Gebru, 2020; Marinucci *et al*., 2023). Evaluators must actively identify and mitigate structural biases, recognising that AI cannot independently ensure fairness or equity in evidence generation.

### 4.1.2 Data Privacy, Confidentiality, Safety and Security Concerns:

The use of AI for evaluation has raised significant concerns about data privacy and security (Patel et al., 2021; Paul, 2024; Golda *et al.* 2024). This has led some to conclude that GenAI is unsuitable for analysis involving very sensitive groups, issues, or data which may require confidentiality for safety and security (Flahavan,2024). Some of the consequences may include data misuse and privacy breaches (Kouha & Thelwall, 2022), espionage and misinformation. This underscores the need for appropriate frameworks and guidelines to safeguard against these risks (Ramezani *et al*. (2023). Evaluators should ensure that sensitive evaluation data should not enter AI systems without clear legal bases, secure environments, and explicit protections for vulnerable groups.

### 4.1.3 Legal and Ethical challenges

The deployment of AI tools in policy evaluations in the UK faces several ethical and legal challenges. The UK government has developed guidelines for responsible AI development and use in the public sector. The document aims "*to guide the safe, responsible and effective use of Artificial Intelligence (AI) in government organisations*." (Government Digital Service, 2025). The objective is to ensure the maintenance of 'public trust', protection of 'individual rights' and fostering of 'equitable societal progress' (ibid).

The legal landscape governing the use of AI in evaluation in the UK is evolving but is still a complex terrain to navigate. Globally, current AI governance frameworks have

been described as "inadequate" (Hadan *et al.*,2025). This is mainly due to lack of empirical grounding in real-world incidents, fragmented coverage, weak enforceability, pace of technological change outstripping regulation etc; this creates uncertainty for practitioners or those wanting to use AI for evaluations.

One of the primary concerns relates to ensuring compliance with the UK General Data Protection Regulation (UK GDPR) and Data Protection Act 2018. Both require stringent data protection standards. Also, as proposed by Cortés et al. (2024), an important ethical question for policymakers to consider in all aspects of the evaluation cycle may relate to the conditions under which any automation of the evaluation processes is: (a) socially acceptable (b) fair and (c) reliable. This has led some to emphasize the need for the development of ethical frameworks, human oversight, and algorithmic transparency in the use of AI (see Cortés et al., 2024; Kouha and Thelwall. 2022). Evaluation teams must integrate legal compliance, ethical oversight, and transparency measures from the outset rather than treating them as afterthoughts.

### 4.1.4 Issues of Transparency and Accountability

The literature points to other ethical issues relating to transparency and accountability in the use of AI in research contexts (see Romberg and Escher, 2024). Practitioners need to provide clear explanations on the details of their use of AI which can enhance users' trust in its outputs and reliability (Pieters, 2011). In this regard, Ferrario & Loi (2022) posit that "*In order to trust AI, we must trust AI users not to trust AI completely*". Equally, accountability and transparency are crucial to both the development and deployment of AI. In instances where errors or unintended outcomes arise from the use of AI systems, appropriate mechanisms must exist to ensure individuals or organisations are held accountable (Novelli *et al.*, 2024). Such measures are vital to discouraging unethical practices and ensuring that those responsible for any harm caused by AI technologies are held to account (e.g., Romberg and Escher, 2023). Where AI informs policy judgements, clear documentation and traceability are essential to enable accountability when decisions go wrong.

## 4.2 Operational challenges

### 4.2.1. Scientific Rigour, Validity and Reliability:

Rigour in the use of AI in evaluation is often used in the sense of methodological rigour, i.e., whether the methods of research and analysis are applied correctly or not. However, rigour can broadly entail ensuring that the entire AI use cycle adheres to robust scientific principles where the epistemic, conceptual, interpretative, and reporting processes are valid and reliable (Olteanu *et al*., 2025). Ensuring scientific rigour, validity, and reliability in AI-driven evaluations is critical to maintain the integrity of outputs.

Yet, there are real challenges in ensuring adherence to scientific rigour especially when using GenAI for research and evaluation (Fleurence et al., 2024; Olteanu *et al*., 2025). This includes data fabrication, hallucinations, and algorithmic biases (Chen *et al*., 2024). The opaque nature of many AI models, often referred to as the “black box” problem, also means that GenAI outputs are often generated without background details on methods, processes and reasoning (Hassija *et al*., 2024). This raises concerns about the quality (validity) of data and analyses from GenAI outputs and whether findings based on these can be relied upon for evaluation purposes. AI-generated findings should always be corroborated by human analytical review to ensure valid, reliable and contextually grounded evaluation outputs.

### 4.2.2. Overreliance and Deskilling:

Overreliance on AI use in evaluation presents with risks associated with deferring uncritically to AI outputs which usually overlook context. This can result in the erosion of crucial evaluation and critical thinking skills as well as underemphasis on human judgement in evaluation (Al-Zahrani, 2024; Gerlich, 2025). For instance, a theoretical perspective on AI assistants proposes that heavy use of such systems might accelerate skill decay among experts and hinder skill acquisition among novices, arguing that radiologists reliant on image-classification AI may have fewer opportunities to exercise their diagnostic judgment, which might lead to cognitive atrophy over time (Macnamara *et al*., 2024).

Also, a captive study of reliance on AI dialogue systems found that heavy AI usage correlated with weaker critical thinking skills in educational settings (Zhai *et al*. 2024). In

addition, as pointed out earlier, AI can perpetuate the marginalisation of already marginalised groups. To mitigate these risks, researchers suggest implementing AI literacy training, promoting critical engagement with AI technologies, and developing strategies for responsible AI use (Mason, 2023; Abuzar *et al*., 2025). AI should support, not replace, evaluators' critical judgement; capacity-building is essential to prevent professional deskilling especially among early and mid-career staff.

#### 4.2.3 Stilted outputs and monotonous machine tone and Style:

Studies compare Gen-AI and human writing styles, examining factors like complexity, readability, coherence, flow and sentiment (Sharma *et al*., 2025). AI literary and prose styles raise questions about artistic value and style preservation (Leitch & Chen, 2025). Concerns have often been raised by some over the non-contextual, stilted, dry and monotonous AI writing style (Lee, 2024; Nilep, 2024). For instance, Safaei and Longo (2023) notes AI's lack of context and ministerial styles. Also, the Behavioural Insights Team's (BIT) (2025) experiment found that "*the initial draft of the AI output was... stilted [and] ...required more revisions than the 'human' version*." However, customizable AI tools can, with the insertion of appropriate prompts or command codes, preserve individual, administrative or organisational styles. AI-assisted writing must be adapted to organisational tone, voice and audience needs; human editing remains vital to ensure clarity and usability. Thus, as indicated earlier, AI should augment, not substitute, human evaluative judgement. Hence, evaluators should ensure targeted prompt design, oversight, and iterative review are done to ensure that outputs remain contextually grounded, credible, and aligned with evaluation standards.

# 5. Emerging best practices and proposed guidelines from the literature

## 5.1 The use of LLMs in evaluation

Effectively harnessing the potential of AI for evaluation requires ongoing experimentation, learning, and adaptation, with the main question often asked by evaluators being: "*How can I know if it performs well*?" (Raimondo et al., 2025a).

In response to this question, Raimondo et al., (2025b) has produced a guidance note intended to synthesise current insights on the integration of AI into evaluation practice, with a main focus on Generative AI (GPT) and LLMs. Recognising "the prompting and validation loop" as the most critical factor in attaining satisfactory outcomes according to their chosen LLMs evaluation metrics , the World Bank IEG team provides the following iterative processes in Figure 2, which resulted from their experiments with using LLMs in various stages of the evaluation processes:

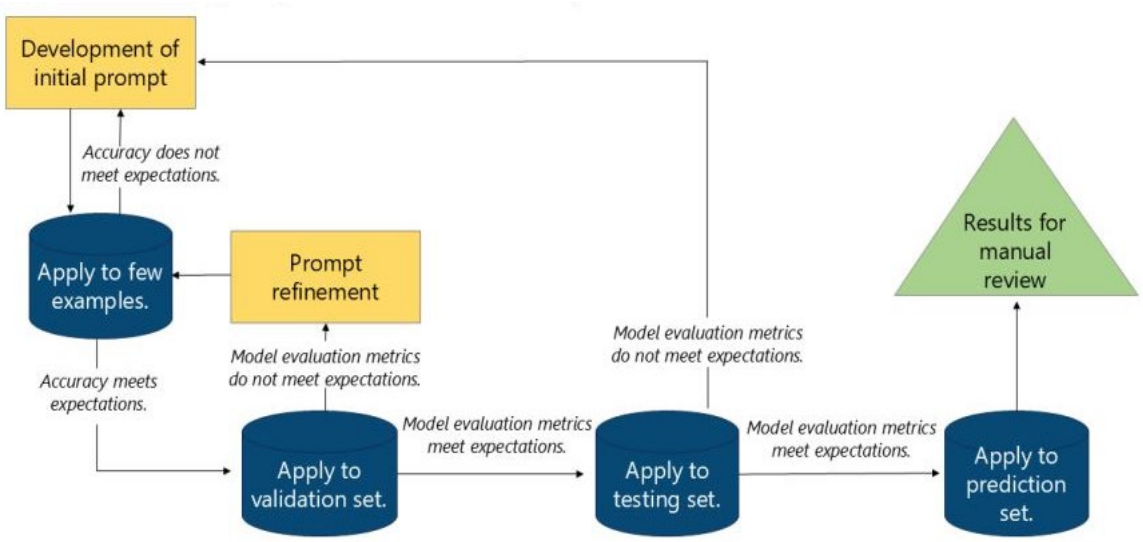


**Figure 2: Prompting and Validation Loop**

Source: Raimondo et al. (2025b)

In line with standard Machine Learning protocols, World Bank IEG suggest that datasets should be divided into 4 subsets: training, validation, testing, and prediction subsets. A small number of human-annotated examples from the training set are embedded within prompts to guide model responses. The prompt is then applied to the validation set to assess performance against predefined metrics; if outcomes are suboptimal, the prompt is iteratively refined.

Once satisfactory results are achieved, the optimised prompt is tested on a separate testing set to evaluate generalisability without further modification. Should test performance remain inadequate, the process is restarted with a new testing subset to

prevent data leakage. If acceptable, the prompt is deployed with manual relevance checks and ideally a sample-based metric evaluation to inform future improvements.

With the division of datasets into four subsets, particular good practices recommended by the World Bank team are:

i. ***Representative Sampling:***
   To enhance the generalisability of prompts applied to prediction data, it is essential that evaluators understand the distribution of the input dataset such as its degree of homogeneity or heterogeneity and to include representative observations accordingly. A representative sample ensures semantic diversity and reduces model bias, thereby facilitating better generalisability and interpretability.

ii. ***Initial Prompt Development:***
   Evaluators should also understand that proper prompting for instruction-tuned LLMs typically includes a defined persona (for the model to adopt for e.g., evaluation analyst), clear task instructions, relevant contextual texts, and explicit response format requirements. In fact, in many models, to ensure clarity and alignment, evaluators must not only adopt the correct model-specific template but also be able to effectively deconstruct complex tasks into step-by-step instructions (e.g., via chain-of-thought prompting). Still, prompt formats require frequent testing and refinement.

iii. ***Model Evaluation:***
   In using AI for evaluation, evaluators should know that the manual review of outputs remains essential. AI performance should be assessed for *faithfulness, relevance, and coherence*. Evaluation strategies must define context-specific metric thresholds, agreed upon by evaluation stakeholders, and supported by clear annotation guidelines (e.g. via codebooks). In classification tasks, confusion matrices offer diagnostic insights and support prompt refinement, especially where minimising false negatives is prioritised, as in structured literature reviews.

iv. ***Prompt Refinement:***
   In the evaluation process, evaluators should refine prompts constantly and this should be guided by validation results. Where performance falls below

predefined thresholds set by evaluators, errors should be analysed and the prompt revised accordingly. This is often by clarifying instructions or addressing incorrect model assumptions. Confusion matrices are valuable for identifying error patterns, such as false positives or negatives. In any case, prompts should remain concise, because excessive modifications risk overfitting and reduced generalisability.

These emerging guidelines are in line with findings from BIT's experiment (BIT, 2025), which highlighted the importance of developing precise, explicit prompts to optimise the relevance and accuracy of AI outputs. BIT also recommended rigorous manual oversight to identify and correct errors, and noted that the performance and reliability of AI tools can vary significantly by topic, limiting the generalisability of the findings.

## 5.2 Evaluator Skills and Attitudes for effective use of AI

Cekova et al. (2025) identify eight essential skills and attitudes which evaluators should cultivate to engage meaningfully and effectively with AI tools in the evaluation practice. Among these are:

- **Human metacognition:** Evaluators draw on their own experience, judgement, and intuition to sense when something "does not feel right". This means checking whether AI outputs align with field knowledge and contextual realities. Human insight remains the quality filter.
- **Ethical awareness**: AI can easily overlook power, equity and inclusion issues. Evaluators therefore must actively question who benefits, who is erased, and whether generated content risks reinforcing harmful stereotypes or bias.
- **Politeness**: AI responds better to clear, respectful dialogue, much like a colleague would. Maintaining a constructive tone improves clarity and keeps the interaction efficient, even when results frustrate.
- **Patience and persistence**: AI conversations often require iteration. Evaluators must be willing to refine prompts, change direction when results stall, and remain strategic in getting from initial drafts to usable insight.

- **Balance**: Providing too much or too little detail can derail the usefulness of AI responses. Breaking complex tasks into smaller questions, while keeping the bigger picture in view, leads to more relevant outputs.
- **Stance awareness**: Evaluators must be clear about their purpose and analytical lens, whether they are emphasising equity, scalability, technical rigour, or scepticism about a programme theory. AI performs better when given that framing.
- **Thoughtfulness:** Before prompting, evaluators take a moment to clarify what they truly need. This prevents shallow or scattered outputs and helps the AI produce responses that advance evaluation goals rather than busywork.
- **Adaptability**: Evaluators should be willing to adjust their approach in response to what is working or failing in the exchange, much like in real-time facilitation. Learning from the pattern of interaction improves outcomes.

## 5.3 Emerging lessons for Evaluators Using AI

In their article entitled "Understanding the Evaluations Role in Measuring the Impact of AI Interventions Across Government"[7], staff from the UK Department for Business and Trade (DBT) identified four “valuable lessons regarding the AI system evaluation”. Here, we reframe these to serve as broad good practice guidelines for evaluators seeking to explicitly incorporate AI tools in the evaluation process.

(i) ***Engage stakeholders early and often***: Involving individuals from diverse teams, professions, and backgrounds before using AI can introduce a wider range of perspectives, enabling the identification of risks, benefits, and challenges that may not have been evident at the outset of the evaluation.

(ii) ***Approach AI systems flexibly and with an open mind***: Given the novelty and inherent uncertainties of AI systems, evaluators should critically reassess assumptions and remain open to unexpected outcomes or evolving system behaviours.

7 https://digitaltrade.blog.gov.uk/2025/04/14/understanding-the-evaluations-role-in-measuring-the-impact-of-ai-interventions-across-government/

(iii) ***Combine multiple methods*:** Integrating quantitative data with qualitative evidence can contribute to more comprehensive evaluations, as can combining and comparing outputs from AI tools with those from traditional evaluation approaches.

(iv) ***A consistent and co-ordinated communication plan*:** Clear and ongoing communication ensures that stakeholders are informed about evaluation activities which incorporate AI activities and are better equipped to engage with Ai outputs.

As with use cases and benefits (Section 2), specific AI tools are associated with certain risks and challenges when used in an evaluation context. These are summarised in Table 3.

**Table 3: Challenges and Best Practices for using specific AI tools for evaluation**

| AI Approach/Tools | Challenges/Risks | Best Practices | Sources |
|---|---|---|---|
| **Natural Language Processing (NLP)** e.g., Google Translate, BioBART and Chatbots. | » Bias in training data affects results/outputs. | » Use domain-specific fine-tuning and evaluation | Hirschberg & Manning (2015); Loor-Torres *et al*. (2024); Baclic *et al*. (2020); Abro *et al*. (2023); LinkedIn, 2023); Stryker & Holdsworth (2024); Dhyani (2025) |
| | » Accuracy issues in domain-specific texts. | » Use diverse, representative training data. | |
| | » NLP "grapples with the complexity, ambiguity, and variability of natural language, which can make interpretation, representation, and evaluation difficult" (LinkedIn, 2023). | » Regularly audit and update models used for evaluation for fairness and accuracy. | |
| | » NLP requires large, annotated/labelled, high-quality datasets | » Combine outputs with careful human/expert review for validation. | |
| | » Struggles with contextualisation especially in low-resource languages. | | |
| | » Potential privacy and compliance concerns when processing sensitive data | | |
| **Large Language Models (LLMs)** e.g., GPT-4, Claude, LLama, BARD, and Gemini. | » Hallucinations (plausible but incorrect outputs) and fabricated citations | » Use multi-metric and human-in-the-loop evaluation to verify performance/outputs with primary sources or manually. | Alqahtani *et al*. (2023); Alowais *et al*. (2023); Zhai *et al*. (2024); Bender *et al*. (2021); Bommasani *et al*. (2021); Belagatti (2025) |
| | » Risk of over-reliance may affect evaluators' critical thinking or cognitive abilities (Zhai *et al*., 2024) | » Transparently document model use and limits | |
| | » High carbon footprint (Bender *et al*., 2021). | » Regularly test with new data and real-world scenarios | |
| | | » Use LLMs as complementary drafting aids, not replacements or final authority. | |
| | | » Choose energy-efficient models. | |

**Table 3: Challenges and Best Practices for using specific AI tools for evaluation (cont…)**

| AI Approach/Tools | Challenges/Risks | Best Practices | Sources |
|---|---|---|---|
| **Generative AI (GenAI)** e.g. Dovetail AI (Qual Research Insight Assistant), Klarity AI (Contract Review Assistant), Jasper AI (Marketing), DALL-E (images), Midjourney (images), ChatGPT (text), Copilot (code) | » Misinformation, hallucinatory, or biased content generation | » Adopt composite evaluation approaches (human & automated) | Zhai *et al*. (2024); Arslan *et al*. (2024) |
| | » Lack of control and explainability in outputs. | » Align Gen-AI use with ethical frameworks and guidelines. | |
| | » Ethical issues: Potential for bias in evaluation and privacy concern | » Require documentation and traceability for generated outputs. | |
| | » Risks of users over-trusting suggestions, reducing scrutiny and affecting quality of evaluation. | | |
| **Machine Learning** | » Model bias from training data | » Iterative human-machine collaboration and reviewer oversight | Ghassemi *et* al. (2020); Mukhamediev *et al.* (2022) Bravo *et al*. (2023) |
| | » Requires labelled training data. | » Regular retraining with updated and diverse data | |
| | » Lack of transparency or interpretability and explainability in model decisions | » Use explainable/interpretable ML models/approaches (e.g. SHAP, LIME) where possible; and document assumptions and parameters. | |
| | » Overfitting or drift when underlying data changes | | |
| **Big Data Analytics** | » Data quality/integrity issues, including missing or biased data, and data veracity issues. | » Ensure compliance with GDPR | Coffman & Reid (2024); Google Cloud (n d.); Mistry (2024). |
| | » Complexity in harmonising disparate datasets | » Establish rigorous data management and validation processes/ techniques | |
| | » Data privacy and governance concerns | » Clearly define objectives and outcome metrics for analysis | |
| | | » Combine with traditional data sources | |
| | | » Big data requires robust infrastructure | |

# 6. Applying the findings: Towards guidelines for the safe and responsible use of AI in UKRI evaluations

Considering the key risks, challenges and best practices identified in this review, we delineate what we see as UKRI responsibilities, including providing systems for effective governance and oversight of AI use and potentially providing training and capacity building activities for evaluators. We also propose checklists which UKRI could provide as guidance to evaluators planning to use AI for evaluation.

Because the existing evidence on the effectiveness of AI in policy evaluation contexts is sparse and comes predominantly from contexts outside R&I funding, we recommend exercising considerable caution in exploring opportunities to incorporate AI in UKRI evaluation. It is recommended that the use of AI tools in UKRI evaluations should be explored:

- Preferably as UKRI policy experiments explicitly designed to test specific applications and approaches through pilot schemes and to document associated challenges.
- Preferably after Phase 2 of this project, during which we will interview evaluators about their experiences of using AI tools in R&I evaluation contexts. This will provide a richer contextual evidence base for UKRI experimentation compared to relying on this evidence review alone.
- Where applicable, through the use of UKRI-specific internal AI agents and platforms to protect data privacy.
- In conjunction with conventional tools and under strong human oversight, since the evidence review suggests that AI tools are best used to complement rather than substitute human effort and existing processes.

## 6.1 Key responsibilities of UKRI in commissioning evaluations with AI use

To create an enabling environment for evaluators to use AI transparently, safely and responsibly, we propose two key UKRI responsibilities as a starting point: ensuring effective governance and oversight, and supporting evaluators in developing the requisite skills. These guidelines are intended to help alleviate some of the structural and operational challenges identified in Section 4, particularly those that fall outside the immediate control of evaluators, or where UKRI can usefully collaborate with evaluators to address them.

**1. To ensure effective governance and oversight of AI use in evaluation, UKRI should consider:**

- Defining appropriate use cases for AI in evaluation (Section 3 provides some guidance)
- Ensuring human oversight is built into all AI-supported evaluation processes; AI must complement, not replace, human judgement (the Evaluator Checklists below ensure evaluators are aware of this requirement).
- Establishing and maintaining a clear accountability framework for all AI-supported outputs and decisions, including processes for accountability in cases of unintended errors or harms arising from AI use.
- Requiring transparent reporting of AI methods, assumptions, and limitations across evaluations.
- Setting up mechanisms to trace when, where, and how AI tools are used.
- Ensuring that ethical review procedures consider fairness, equity, representation, and risks of harm in evaluations involving AI.
- Providing opportunities for clear, consistent communication with evaluators about when and how AI tools are used.
- Establishing transparency and disclosure policies, with systems to incentivise accurate reporting of AI use (e.g., emphasising that disclosure enhances trust).
- Creating a system which allows learnings and feedback loops on the effectiveness of using AI, for example through an internal repository or an AI community of practice.

**2. To ensure UKRI Evaluators have the requisite AI skills, UKRI should consider:**

- Offering or requiring AI literacy training for evaluators expected to use AI, ensuring evaluators develop key mindsets and skills highlighted in Section 5. This training could also extend to evaluators within UKRI involved in reviewing commissioned evaluations that use AI methods, where it would be useful to have AI literacy that enables effective assessment of AI methodologies.
- Ensuring that training builds evaluator capacity to engage critically with AI while limiting overreliance and preventing skill decay or hindered skill acquisition.
- Periodically reviewing and updating governance, training, and ethical standards to reflect new developments in AI evaluation practices, drawing wherever possible from evaluator reflections and feedback on AI use.

### 6.2 Using AI in Evaluation: A Checklist for UKRI Evaluators

We provide two potential checklists relevant to UKRI evaluators intending to use AI in evaluation. The first is applicable across different types of AI use and relates to ensuring that evaluators understand the uses and limitations of AI, are transparent in their use of AI, comply with ethical and legal requirements, adopt best practices to maximise scientific rigour, reflect on any learnings and feedback to UKRI. This checklist is based on integrating insights from Sections 4 and 5. The second checklist complements this by focussing on specific AI tools and the best practices in applying them for evaluation; this checklist is primarily based on Table 3.

Both of these checklists are non-prescriptive, and UKRI or other evaluation commissioners may adapt them, for example by prioritising certain elements of the checklists based on the context of a specific evaluation or the type of evaluation activity that AI assists.

#### Checklist 1: Responsible Use of AI in Evaluation

1. **Understanding of AI tools and Utilisation Readiness**
   - We have clearly informed UKRI stakeholders of our intentions or plans to incorporate AI in evaluation, and we will maintain consistent communication on AI use throughout the project.
   - We have a clear understanding of what the proposed AI tools do, how they work, and their known limitations.
   - We have identified a clear purpose for using AI in this task (e.g., summarising data, analysing text).
   - The task is appropriate for AI assistance (i.e., it involves repetitive or large-scale data processing rather than subjective judgement).
   - We have checked whether suitable human expertise, data and digital infrastructure are in place to support AI use.
   - The AI tool or model (e.g., NLP, ML, LLM) has been selected based on clear functionality and suitability for the task.
   - The model's training data and version are known and documented.
   - Where possible, an explainable or interpretable model has been chosen (e.g., ML models with SHAP/LIME).
   - Human oversight mechanisms are in place to review outputs and make final judgements.

**2. Ensuring Data Privacy, Confidentiality and Security**

- Data sources are reliable, representative and relevant to the evaluation question.
- Sensitive or personal data have been anonymised or protected in line with UK GDPR and the Data Protection Act.
- Data storage and transfer comply with established legal standards.
- Sensitive evaluation data will not enter AI systems without clear legal bases, secure environments, and explicit protections for vulnerable groups.
- Potential biases or gaps in the dataset have been assessed (e.g., under-representation of certain groups or regions).
- All data sources and processing steps have been documented

**3. Ensuring Transparency and Compliance with Ethical and Legal Frameworks**

- We have considered potential impacts of AI use on fairness, equity and inclusion.
- If AI is used in decision-making (e.g., scoring, classification), results have been cross-checked by humans.
- All AI use has been disclosed in evaluation documentation and reports, including which tools were used and for what purposes.
- Any automated processes that could affect outcomes are transparent to stakeholders.
- Clear responsibility is assigned for all outputs; we understand that AI tools do not replace accountability by evaluators.
- The use of AI to automate evaluation processes is compliant with UK Government and UKRI's legal guidelines related to data protection,
- The use of AI to automate evaluation processes considers any ethical guidelines, including ensuring the intended use of AI is socially acceptable, fair and reliable.

**4. Ensuring Scientific Rigour and Reliability**

- AI outputs are checked against predefined quality criteria (e.g., validity, reliability, relevance).
- Where feasible, confusion matrices or similar diagnostic tools are used to assess classification accuracy.
- Hallucinations, fabricated data or unsupported claims have been removed.
- The overall analytical process will remain methodologically sound and consistent with UKRI's evaluation standards.

- The tone, clarity and style of AI-generated text are reviewed and edited to meet UKRI communication standards.
- All AI-generated outputs are always corroborated by human analytical review to ensure valid, reliable and contextually grounded evaluation outputs.
- AI-generated outputs are compared with human or traditional methods to check consistency.
- Any limitations, anomalies or biases observed are recorded and communicated.
- AI outputs are not accepted uncritically; they are reviewed manually for faithfulness, coherence and factual accuracy.
- Any final conclusions are based on evaluator judgement.

**5. Reflection, Learning and Capacity Building**

- We have reflected on how AI affected efficiency, quality and inclusiveness of the evaluation.
- We have reflected on our role, values and evaluative purpose when engaging with AI.
- Lessons learned about prompts, AI tools or validation methods have been documented for future evaluations and shared with UKRI.
- We have identified any training or skill development needs for evaluators which can enhance more critical engagement with AI tools
- Stakeholders and team members have been informed of AI use and invited to provide feedback.
- We have contributed to UKRI's internal learning on responsible AI use in evaluation.

Checklist 2: Additional checklist for the responsible use of specific AI tools in Evaluation

**1. Using Natural Language Processing (NLP):** *e.g., Google Translate, BioBART and Chatbots.*

- We will use diverse, representative datasets to reduce bias.
- We will apply domain-specific fine-tuning for policy/evaluation terminology.
- We will maintain privacy controls when processing sensitive text.
- We will validate AI interpretations with human expert review.
- We will continuously audit performance for accuracy and fairness.
- We will document training data limitations and assumptions.

**2. Using Large Language Models (LLMs) and other Generative AI tools:** *e.g., GPT-4, Claude, and Gemini, ChatGPT, Copilot*

- We will treat outputs as drafts subject to expert verification.
- We will check for hallucinations and fabricated evidence.
- We will use structured *prompt-validation loops* with clear metrics
- We will disclose LLM use in evaluation outputs.
- We will document use to ensure traceability of any generated analysis.
- We will ensure targeted prompt design, human oversight, and iterative review to ensure that outputs remain contextually grounded, credible, and aligned with evaluation standards.
- We will maintain full documentation of generated text, images, code, or summaries.
- We will review output for stilted style, logical gaps, and bias before use.

**3. Using Machine Learning (ML):** *(e.g., SHAP, LIME)*

- We will confirm that training data are relevant to the intervention/population
- We will routinely retrain models and monitor them for model drift.
- We will use explainable ML tools (e.g., SHAP, LIME) for accountability.
- We will pair predictions with human interpretation.
- We keep a clear record of AI tools and model parameters used.
- We will evaluate performance for fairness across demographic groups.
- We will pilot test before fully adopting.

**4. Using Big Data Analytics (BDA)**

- We will validate data quality, completeness, and integration assumptions.
- We will ensure GDPR-compliant acquisition and governance of data.
- We will build robust infrastructure to handle sensitive data.
- We will cross-check automated insights with primary/traditional data.
- We will define clear objectives and metrics for any analysis.
- We will apply strong security and metadata management.

# 7. Summary and next steps

## 7.1. Summary

This report has reviewed the evidence base on the use of Artificial Intelligence (AI) tools in policy evaluation. Drawing on academic and grey literature, the review highlighted both potential AI applications and practical use cases, mostly outside Research and Innovation evaluation but relevant across policy contexts. Using a funding evaluation cycle, the review considers AI use in selecting interventions, assessing proposals, monitoring programmes, and conducting impact evaluations. We found that tools and approaches such as Natural Language Processing, Big Data Analytics, Machine Learning, and Large Language Models (e.g., GPT) have been applied to assist in prioritising intervention areas, proposal assessments, and both quantitative and qualitative impact assessments. These applications of AI have the potential to confer benefits related to efficiency and scalability of evaluation processes.

The review also presented evidence on the risks and challenges of AI use in evaluation contexts. We identify structural and systemic concerns relating to equity, fairness, ethical and legal frameworks, data privacy, transparency and accountability. We also identify operational and performance related risks, including uncertain levels of rigour and reliability and quality of outputs from generative tools, and the potential overreliance on AI which could lead to human deskilling. Emerging best practices from the literature suggest heavily involved human oversight, iterative validation of approaches e.g., prompts for Gen-AI, and development of the skills and attitudes of evaluators in terms of their engagement with AI tools.

Publicly available evidence on the use and effectiveness of AI in Research and Innovation evaluation contexts is rare, as is evidence on ‘everyday’ applications of AI in evaluations (i.e., cases where testing or experimenting with AI tools is not the objective of the evaluation).

The review has also developed a set of initial guidelines for UKRI and its evaluators based on the key benefits, risks and challenges identified. These guidelines, summarised in the Executive Summary and detailed in Section 6, should be refined based on a more robust evidence base developed within the context of R&I funding and evaluation.

## 7.2. Next Steps

To address the evidence gaps, semi-structured interviews could be conducted especially with staff in organisations that are actively trialling the use of AI tools in evaluations and with consultancy companies that are experienced in the application of AI in evaluation. It would be useful to also gain insights from UKRI's own portfolio of evaluators. We envisage potential interviews could help in answering the following questions:

1. What specific AI tools or models do they use, and for which tasks in the evaluation process?
2. Compared to manual methods, what concrete benefits do these organisations gain from using particular AI tools or models?
3. What challenges or risks do they face in adopting these AI tools or models?
4. How do organisations and staff manage these challenges while balancing the use of innovative AI tools with adherence to ethical principles and guidelines?
5. Do organisations and staff disclose the use of AI tools or models in their reports, and specify the outputs generated? If yes, how are these outputs received or assessed by line managers and clients? If not, what are the reasons for not reporting them?

# 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.

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

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_source=chatgpt.com

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.

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_source=chatgpt.com#choosing-the-evaluation-approach

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*.

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_source=chatgpt.com

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, 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. (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/

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

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*.

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.

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.

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

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 (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. (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

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*.

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

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/

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.

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 I: Evaluator's conversational skills and attitudes for better engagement with AI

| Skill | What it is About | Inner Thinking |
|---|---|---|
| **Human metacognition** | Using uniquely human judgment, intuition, and lived experience to guide the conversation and evaluate responses critically. | • "Does this align with my experience?" • "What contextual knowledge am I bringing that's missing here?" • "Something feels off about this conclusion—what could be wrong?" • "What background insights from me may enrich this analysis?" |
| **Ethical awareness** | Bring conscious attention to ethical dimensions that AI may miss or handle poorly, including considerations of bias, impact, and representation. | • "Are there ethical implications being overlooked here?" • "Whose perspectives may be marginalised?" • "Is the AI unintentionally steering toward a particular worldview? How might this analysis impact vulnerable groups?" |
| **Politeness** | Maintain a constructive conversational approach that research shows yields better results, without unnecessary deference or formality. | • "How would I phrase this to a knowledgeable colleague?" • "Am I expressing myself clearly without being unnecessarily demanding?" • "Is my frustration affecting the quality of our exchange?" |
| **Patience and persistence** | Balance continued effort with strategic pivots when needed, recognising when to push forward and when to change course. | • "Is this approach getting us closer to what I need, or should we try something different?" • "What small adjustments might improve our direction?" • "When should I step back and reconsider our approach entirely?" |
| **Balance** | Find the right level of detail, context, and direction for productive exchange, including when to break complex problems into manageable parts. | • "Have I provided enough context without overwhelming?" • "Should I break this down into smaller questions or maintain the broader view?" |

| | | |
|---|---|---|
| | | • “Am I getting lost in details when a simpler approach might work better?" |
| **Stance awareness** | Be aware of the position, intentionality, needs, objectives, and the paradigm from which the conversation is being approached. | • "Have I clarified what I'm really trying to accomplish?” • “Does this direction serve my actual purpose?” • “Am I true to my intent and approach?" |
| **Thoughtfulness** | Take time to consider what is really wanting to be known or accomplished, recognising that careful consideration of one’s own input dramatically impacts output quality. | • "What am I truly trying to understand?” • “Is this the right question to get me closer to my goal?” • “Have I taken enough time to frame this request effectively?” • “How could a more considered approach yield better insights?" |
| **Adaptability** | Adjust the approach based on what is working and what is not, learning from the conversation patterns that emerge. | • "What is most effective in our exchange so far?” • “Which approaches are yielding the best insights?” • “How can I modify my approach based on what I am learning about this conversation’s dynamics?" |

Source: Cekova *et al*. (2025)

www.ircaucus.ac.uk

Email info@ircauscus.co.uk Twitter @IRCaucus

INNOVATION & RESEARCH CAUCUS

Delivered with ESRC and Innovate UK