SCOPING STUDY ON THE

Use of Artificial Intelligence in Climate Change Evaluations

February 2025

Technical Evaluation Reference Group
ADAPTATION FUND

CIF CLIMATE INVESTMENT FUNDS

IEO Independent Evaluation Office
GLOBAL ENVIRONMENT FACILITY

GREEN CLIMATE FUND | Independent Evaluation Unit

Acknowledgments

This scoping study was jointly commissioned by the evaluation units of the Adaptation Fund (AF), Climate Investment Funds (CIF), Global Environment Facility (GEF), and Green Climate Fund (GCF). The study was carried out by Elinor Bajraktari, an individual consultant. Special thanks are extended to Vladislav Arnaoudov (AF-TERG), Anupam Anand (GEF IEO), Brittney Melloy (CIF) and Yeonji Kim (GCF IEU) for their guidance and supervision during the entire process. Without their assistance, this study would not have been possible.

Table of Contents

Table of Figures

List of Acronyms

AEA	American Evaluation Association
AF	Adaptation Fund
AF-TERG	Technical Evaluation Reference Group of the Adaptation Fund
AI	Artificial Intelligence
AIDA	Artificial Intelligence for Development Analytics
CARMA	Carbon Monitoring for Action
CESM	Community Earth System Model
CIF	Climate Investment Funds
CME	Climate Model Emulator
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA	Evaluation Resource Center
GCF	Green Climate Fund
GDELT	Global Database of Events, Language and Tone
GEF	Global Environment Facility
GHG	Greenhouse Gas
GRAF	Global High-Resolution Atmospheric Forecasting
IEG	Independent Evaluation Group
IEO	Independent Evaluation Office
IEU	Independent Evaluation Unit
IFAD	International Fund for Agricultural Development
IRMCT	International Residual Mechanism for Criminal Tribunals
LLM	Large Language Model
ML	Machine Learning
NCAR	National Center for Atmospheric Research
NLP	Natural Language Processing
OIOS	United Nations Office of Internal Oversight Services
SDG	Sustainable Development Goals
UNDP	United Nations Development Programme
UNFPA	United Nations Population Fund
UNICEF	United Nations Children's Fund
WB	World Bank

I. Introduction

Artificial Intelligence (AI) has become an exciting and rapidly changing technology. Previously a phenomenon undergoing steady evolution alongside other components of computer science, the technology got a jumpstart with the launch of Open AI's ChatGPT in November 2022. This moment marked the emergence of generative AI – a significant milestone and leap forward in the capabilities of AI.[1] For the first time, a non-human entity was able to create completely new content, simply analysing, processing, or reorganizing existing data. Since that moment, AI has been going through accelerated and continued improvement, with numerous updates, versions, and models appearing almost every day. The generative capabilities of many models – be they in the form of text, images, voice, or other forms of media – are impressive and there is consensus among experts that this technology has potential to profoundly shape the future of many disciplines worldwide. From creative industries like art and literature to scientific research and problem-solving, to business and public sector processes, products and services, the applications of AI are vast and far-reaching. As the technology continues to improve, it is likely to have a transformative impact on the way we work, create, and interact with other technologies on a global scale.

One particularly promising application of AI is in the field of evaluation, where it has the potential to improve significantly the quality and efficiency of evaluation processes. Recognizing this opportunity, the four climate funds – Adaptation Fund (AF), Climate Investment Funds (CIF), Global Environment Facility (GEF), and Green Climate Fund (GCF) – decided to jointly commission a scoping study that explores the benefits, opportunities, and risks of the use of AI applications in program evaluations, with a particular focus on evaluations related to climate change.

1.1. Background

While AI has the potential to transform many areas of human endeavour, this potential is particularly remarkable in the area of public policy research. Within this field, program evaluation[2] is a function that stands to benefit significantly from the

1. While the term "AI" is a broad category of technologies that have been available for many years, in this paper the term "AI" is used to refer primarily to generative LLM models which have recently revolutionized the field of data analysis.
2. In the context of this study, the term "program evaluation" instead of simply "evaluation" is used because it is a narrower and more precise term than "evaluation," which can refer to the assessment of various entities and processes, such as health (evaluation of the health of an individual), building (evaluation of a building), and so on. The term "program evaluation" refers to the systematic assessment of the design, implementation, and outcomes of a specific program, policy or intervention. It should not be confused with specific uses of the term "program" for particular modalities in the context of certain organizations.

use of AI. As more and more data become available thanks to increasing availability and sophistication of data-producing technologies such as mobile devices and the Internet, governments will increasingly prioritize evidence-based and data-driven decision-making. Given the role of data as the engine of AI technologies, individuals and organizations will increasingly recognize the value of AI for analytical purposes and will seek to integrate it into evaluation processes. As the following sections of this report will demonstrate, this is a process that is already happening. The value of AI is particularly pertinent to evaluations related to climate change, where the complexity and scale of interventions demand the processing of large-scale data.

Being at the forefront of climate finance, AF, CIF, GEF, and GCF are committed to ensuring the effectiveness of their investments. The evaluation of the activities that are financed by these funds is a key instrument towards this objective. In light of this, the four funds have commissioned this scoping study to investigate the potential of using AI in program evaluations, with a focus on evaluations related to climate change activities.

BOX 1: Common AI categories

The terminology used in the AI field is complex due to the diversity of methods and approaches used. The following is a very brief explanation of the key terms that will be used further in this paper.

- **Machine Learning (ML) -** ML is a branch of AI. It focuses on algorithms and models that instruct computers to learn from large data and make predictions or decisions without being programmed for specific tasks.
- **Non-Generative AI -** Non-Generative AI analyses and reorganizes, without being able to generate any new content.
- **Generative AI -** Generative AI is an AI system that is capable of creating new content out of existing information. This includes text, images, voice, or other forms of data.
- **Natural Language Processing (NLP) -** NLP is a specialized sub-field of AI which uses ML techniques. It enables computers to understand, interpret, and generate language in a meaningful way. NLPs perform tasks like translation, sentiment analysis, and summarization of text. NLPs can be of both generative and non-generative in nature.
- **Large Language Models (LLMs) -** LLMs are considered as a subset of NLP. LLM models are typically trained on vast amounts datasets, which enables them to understand information and generate human-like text. Their operating logic is to predict the next word in a sentence, which allows them to generate coherent text based on a given input and instruction. LLMs are primarily of a generative nature, because they are able to generate new content.

1.2. Study objectives

The primary objectives of this study, established by the four climate funds, were to:

1. Explore the potential of using AI in program evaluations, focusing on evaluations related to climate change

2. Assess opportunities and risks associated with AI application across all stages of the evaluation process

3. Provide insights to inform future AI applications in program evaluations, especially in the area of climate change (including – if possible – specific cases where AI has been used successfully)

By addressing these objectives, this study seeks to contribute to the understanding of the funds' staff and decision-makers, as well as their partner (implementing) organizations, on how AI could be utilized to improve the quality, efficiency, and impact of program evaluations, especially in the area of climate change.

1.3. Methodology

The study employed a mixed-methods approach consisting of three main data collection components:

1. **Literature review:** The study started with a systematic review of the existing literature. A thorough search was conducted to identify any existing knowledge on the use of AI in program evaluations. Particular focus in the identification process was placed on climate change. The search was based on a set of inclusion and exclusion criteria, followed by the screening and assessment of the quality of studies, and ending with the selection of the most relevant works. Subsequently, the studies were analysed from the perspective of the use of AI in evaluations and the most pertinent information was synthesized. The resulting review, presented in this report, seeks to provide a basis for understanding the current state of AI use in evaluations.

2. **Semi-structured interviews:** The researcher conducted semi-structured interviews with 19 key stakeholders. These included staff members of the four funds and key international organizations who are directly involved with evaluations, and some key independent evaluators or representatives

of consulting firms or academic institutions who have written about or have experimented with the use of AI in evaluations. The Steering Committee[3] that was set up by the four climate funds for this study helped with the identification of interviewees and the organization of interviews. The interview guides were developed based on the findings of the literature review and the key research questions outlined in the study's Terms of Reference. All interviews were conducted remotely via video conferencing platforms.

The analysed information was synthesized to answer the key research questions and provide insights on the broader question about the potential of AI in program evaluations, with a focus on climate evaluations.

3. **Online survey:** Additionally, the researcher organized an online survey with the aim of reaching out to a broader range of stakeholders. While the interviews were more targeted to institutional representatives, the survey was more targeted to independent evaluators, particularly those working in the area of climate change. The survey questionnaire included both closed-ended and open-ended questions, which were designed to capture quantitative and qualitative data. The sample size and the contact information of the invitees were determined in consultation with the Steering Committee. The survey was administered using SurveyMonkey. A total of 32 responses were received. More detailed information on the profile of the survey respondents is provided in Annex II of this report.

The information collected from the three above-mentioned sources was analysed using a combination of qualitative (thematic coding and content analysis) and quantitative (descriptive and inferential statistics) techniques. The findings from the literature review, the interviews, and the survey were triangulated to identify key themes, patterns, and trends. The analysed information was synthesized to answer the key research questions and provide insights on the broader question about the potential of AI in program evaluations, with a focus on climate evaluations.

1.4. Cautionary note

Two cautionary points should be made in clear terms before the presentation of the findings of this study.

- Firstly, AI is an all-encompassing term that comprises a diverse array of tools,

3. This committee was comprised of one representative from each of the four climate funds that commissioned this study.

methods, and approaches. Each of them has its own characteristics, utility and performance level. The findings of this report are for the most part generalized statements about AI. While references to specific AI tools and methods are made throughout the report, many findings are presented in light of AI as a broader technology.

- Secondly, AI is currently evolving at a breathtaking pace. New models and platforms appear every day, while existing models get frequent updates. This presents a challenge – it is difficult for researchers and practitioners to keep up with such rapid change. For this reason, it is crucial to recognize that the findings and observations presented in this study are of temporary value. They are mostly applicable for the present (the time when this research was conducted, which was July-August 2024). Given that the capabilities and performance of AI systems are continuously improving, many of the challenges and limitations outlined in this report may become less relevant or even obsolete soon. However, it is still important to acknowledge that while some AI challenges identified in this report may be resolved by the industry, others will persist, and new ones will inevitably arise. This study highlights a fundamental characteristic of AI: independently of the model or version, it will need close human supervision and ongoing monitoring, assessment, and adaptation.

II. Literature review

The question at the centre of this study – ***"What is the potential of using AI in program evaluations, with a focus on climate evaluations, for all stages of the evaluation process?"*** – is also the key question that has driven the literature review presented in this section of the report. The literature on this topic is an emerging one. Most of the studies reviewed here – as can be seen in the bibliography list at the end of this report – bear the date of 2023 or 2024. The bulk of this research was prompted by the explosion of generative AI, following the launch of ChatGPT in November 2022.

A crucial source for the information presented in this review has been the Summer-Fall 2023 issue of the *"New Directions for Evaluation"* journal. This issue was dedicated to the exploration of the implications and opportunities of AI, particularly generative AI such as ChatGPT, for the field of evaluation. The ten articles constituting that issue examined foundational concepts in AI, potential applications and risks for the evaluation practice.

This section is divided in three parts: (a) Application of AI in Evaluations; (b) AI in Climate Change Evaluations; and (c) Opportunities, Risks, and Mitigation Strategies.

2.1. Applications of AI in evaluations

The literature review conducted for this scoping study identified several key themes, which will be summarized in this section. At the high level, there is work, such as that of Nielsen, S. B. (2023), which looks at the extent at which AI has been applied in evaluations. One of Nielsen's main findings is that while evaluators have been relatively slow to adopt AI compared to other professions, recent evidence suggests increasing interest and application. Nielsen argues that soon AI may displace some human tasks (e.g., transcription, translation, screening of documents, coding of text), while augmenting others (e.g., research design, establishing evaluation criteria, critical synthesis). Consequently, according to Nielsen, evaluators will need to not only accept AI, but also train it for optimal use.

Other researchers have explored the benefits and challenges of applying AI in evaluations. An example of this work is the upcoming book of Leeuw and Bamberger (2025) which explores the evolving role of big data and AI in the field of evaluation and highlights their potential benefits, such as reduced costs and time for data collection and analysis, the ability to work with larger samples and conduct granular

analysis, and the ability to evaluate complex programs operating in dynamic contexts. One interesting area the authors explore is how AI can be used in collecting and repurposing administrative data, especially in areas such as climate change and the Sustainable Development Goals (SDGs). On the other hand, the authors emphasize the importance of theory-driven evaluations, with the theoretical model providing a framework for aligning big data variables with program objectives and hypothesized causal pathways and linkages.

Several studies have explored the way AI could be used in more specific evaluation tasks. One area in which AI has shown promise is in analysing qualitative data.

- For example, Kates and Wilson (2023) explore how LLMs like ChatGPT can be used for common evaluation tasks, such as proposal writing and evaluation planning. The authors claim that AI can significantly improve the quality and time-effectiveness of these processes by helping evaluators generate new (overlooked) ideas and produce first drafts. Similarly, Head, C. B. et al. (2023) looked at the viability of using LLMs for automating several evaluation tasks. They found that LLMs can have particular usefulness in tasks that involve the analysis of text – e.g., summarizing documents, comparing paragraphs or documents, extracting information from a long text or document, analysing themes and sentiment, generating new text, and so on.

- Sabarre, N. R. et al. (2023) conducted an experiment with three AI tools, which they used to code in the evaluation process. They compared the results of AI to manual processes and found that one of the tools (CoLoop)[4] produced themes which were remarkably similar to manual coding in a fraction of the time, while also showing an ability to respond to nuanced questions. The second tool (Avalanche)[5] generated more granular themes with high accuracy, but at a higher cost. The third tool (Rayyan)[6] was less contextually relevant to the researchers' needs. Based on this experiment, the authors argue that it is crucial to carefully consider the appropriateness and limitations of AI tools in specific evaluation contexts.

- Hitch (2023) explored the potential of AI tools like ChatGPT in reflexive thematic analysis[7] for qualitative health research. He tested the performance of ChatGPT in certain tasks such as identifying patterns and themes that are

4. CoLoop is a platform that allows for collaborative data exploration and machine learning. It enables multiple users to work jointly on data analysis and model development in real-time.
5. Avalanche is an open-source platform that allows researchers and developers to develop models that learn from the data over time without forgetting previously learned information.
6. Rayyan is a web-based platform that helps researchers screen and analyse large volumes of literature.
7. The "reflexive thematic analysis" is a type of qualitative research which is used by social researchers to identify and analyse patterns and themes within a given dataset.

not immediately obvious to human researchers. Hitch's testing showed that ChatGPT cannot fully replace human judgement in contextual interpretation and reflective deliberation, which are essential requirements of an evaluation. The author emphasizes that AI should be used to augment, not replace, the analytical process. He advises researchers to be vigilant and critical towards any outputs generated by AI.

- Morgan (2023) tested ChatGPT's ability to conduct an analysis of qualitative data by comparing its performance with his own manual analysis. He tested this for two datasets – one using Reflexive Thematic Analysis and the other using Iterative Thematic Inquiry. The author found that ChatGPT was quite successful in reproducing the themes he identified manually. AI was particularly effective for concrete, descriptive themes, but was less useful for interpretive, and especially subtle, themes. Although ChatGPT was simple to use and took much less time than the manual coding, Morgan argues that AI approaches should be seen as merely tools that cannot replace the analyst's intimate knowledge of the data.

The existing literature suggests that one particular evaluation task which AI seems to be well capable of handling is the summarization of text. For example, Balvir et al. (2024) have shown how AI can extract meaningful information from vast amounts of data and present it in a concise manner. This is a particularly valuable feature in the current situation of information overload. In another study, Shakil et al. (2024a) tested text summaries created by OpenAI's GPT models for conciseness, relevance, coherence and readability by comparing them with traditional methods. The authors found significant similarity between the two types, which demonstrated the potential of GPT models as useful tools for summarization. In another study, Shakil et al. (2024b) propose an approach for generating higher-quality summarization while minimizing the output of false information, often referred to as "hallucinations". To this end, they introduce a refinement process through tools like FactSumm,[8] QAGS,[9] SummaC,[10] ROUGE,[11] and GPT, which improves in an iterative way the accuracy of summaries. The researchers used statistical analysis to confirm the significance of improvements achieved in terms of factual consistency and reduction of hallucinations, particularly for abstractive and hybrid summaries. Through this work,

8. FactSumm is a model that uses machine learning techniques. It is used to assess the factual consistency of summaries. It identifies factual inaccuracies by comparing a generated summary with the source document.
9. QAGS, which stands for Question Answering and Generation for Summarization, is a metric that is used to evaluate the factual accuracy of text summaries.
10. SummaC, which stands for Summarization Consistency, is a model used to measure the consistency of summaries with their source documents. SummaC employs a pre-trained model fine-tuned for natural language inference (NLI) to compare the summary with the source text.
11. ROUGE, which stands for Recall-Oriented Understudy for Gisting Evaluation, is a tool used for assessing the quality of text summaries by comparing them to reference summaries.

the authors underline the need for more sophisticated summarization approaches that capture more effectively the nuances of language. They also emphasize the importance of factual consistency and context in assessing the quality of summarization tasks.

Researchers have also tested the use of AI in more complex tasks. One of these is the prediction of socio-economic outcomes based on the analysis of key input variables, typically in the form of large data. A good illustration of this is the work of Jean et al. (2016), which demonstrates an accurate, inexpensive and scalable method for estimating consumption expenditure and asset wealth in several countries in Africa by applying machine learning techniques to high-resolution satellite images. The researchers used a convolutional neural network,[12] pre-trained on night-time lights, as a proxy for economic activity. They fine-tuned it on daytime images to learn features that could explain up to 75 per cent of the variation in local-level economic outcomes from survey data. By leveraging information about lights, the researchers were able to overcome the limitations of survey data. With this model, they were able to demonstrate the strong predictive power of AI across multiple countries, outperforming models based solely on night-time lights. They also showed the possibility of generalizing across countries, which suggests the presence of common determinants of livelihoods that can be identified through satellite imagery.

In another study, Blumenstock et al. (2015) used a person's history of mobile phone use to infer their socioeconomic status. With this inferred individual-level information, they were able to reconstruct the distribution of wealth for an entire region or nation. The authors combined mobile phone metadata from Rwanda's largest mobile phone operator with follow-up phone surveys to train machine learning models that predict poverty and wealth with high accuracy. Through this approach, they were able to create a cost-effective method for gathering socioeconomic data in environments where traditional data collection is constrained by the lack of infrastructure and financial resources. At the same time, the study also revealed challenges related to protecting the privacy of surveyed individuals and the commercial interests of mobile operators when conducting such analyses.

Björkegren and Grissen (2020) created a model that can predict defaults on loans among borrowers in developing countries who have no formal financial history. A crucial aspect of this research was the application of machine learning to data on behavioural patterns extracted from mobile phone usage. The authors applied the model to approximately 5,500 behavioural indicators from mobile phone transaction

12. A Convolutional Neural Network (CNN) is a type of deep learning model which is used for image recognition, object detection, and other vision tasks.

records to predict repayment. The results of their model outperformed models based on information from the credit bureau. Their model also performed well for unbanked individuals who would be impossible to score with traditional methods. To make their predictions more stable over time, the authors used only "within-time period variation",[13] which helped them avoid common shocks that occur over time. With this approach, they reduced significantly the cost of screening individuals who were on the margin of the banking system and enabled the banks to come up with new forms of low-cost digital credit. At the same time, the study also outlined key challenges such as privacy issues, the potential for manipulation, and the need for the model to be re-estimated if phone sharing or the usage of multiple SIM cards changed over time.

The study found that machine learning provided more reliable estimates of the impacts of interventions targeted at key factors such as ex-combatants' confidence in government, social networks, security, and emotions, compared to employment status.

Samii et al. (2016) used machine learning to derive retrospective causal inferences, instead of using traditional methods such as standard regression analysis. The researchers in this case used AI to evaluate the success of anti-recidivism measures for former combatants in Colombia. The study found that machine learning provided more reliable estimates of the impacts of interventions targeted at key factors such as ex-combatants' confidence in government, social networks, security, and emotions, compared to employment status. With this research, the authors showed how machine learning can be employed to sort through a large number of potential causal factors to identify intervention priorities, especially for outcomes that are rare or appear after many years – key challenges in traditional evaluation approaches. Parthasarathy et al. (2017) have used NLP methods to examine the effects of inequalities that result from gender and status on deliberative influence in village assemblies in India. By applying AI to village assembly transcripts, the authors analysed key quantitative aspects of deliberative quality, such as floor time, agenda-setting power, and the state's responsiveness to citizen demands. The study showed that while the citizens had greater influence than the officials in the discussions, women faced significant disadvantages relative to the men. They were less likely to speak, set the agenda, and receive relevant responses from state officials, even when controlling for the topics raised. The authors also found that the presence of female

13. In econometric models, "within-time period variation" refers to the fluctuations that occur within a specific, defined time period, such as a day, a week, a month, or a year. It is used to estimate the effects of time-varying variables.

council presidents through gender quotas improved significantly the likelihood of women receiving a response by the state. AI in this case was crucial for determining the causal effects of key variables.

Another important topic that has been explored by several researchers is the interaction of AI with evaluation validity and ethics. An example of this is the work of Azzam, T. (2023), who has identified ways in which AI can be used to improve the validity of arguments and areas where the use of AI can create challenges. The study is organized around quantitative validity (internal validity, external validity, measurement validity) and qualitative trustworthiness (credibility, transferability, dependability). Azzam reports that AI's ability to establish internal validity for causal claims is limited by its lack of access to contextual information – especially, in aspects like identification of variables and participants, data collection approaches, and the situational context. However, when as far as external validity is concerned, Azzam finds that AI offers good potential for detecting patterns in large datasets. This is particularly helpful for assessing the replicability of findings to other contexts. Based on this, the author suggests that AI can be used in the development of "construct-validated measures"[14] for simple and observable phenomena. In particular, AI can contribute to qualitative analysis through sentiment analysis, keyword identification, and pattern matching across contexts. On the other hand, deeper meanings, lived experiences, and the constructivist paradigm[15] – which is central to qualitative research – currently remain challenging for AI. Azzam makes the case for a collaborative approach that leverages the strengths of both human judgement and AI's computational capabilities to generate assessments that have more robust validity.

Ferretti, S. (2023) has investigated the role prompts play in shaping AI results. Using four personality types – "pedantic", "I know it all", "meek", and "speech virtuoso" – the author demonstrates how different prompts can lead to different results and insights. For example, ChatGPT's "pedantic" trait can be leveraged to generate through detailed prompts high-quality outputs (concept notes, logical frameworks, questionnaires, training modules and other evaluation components). On the other hand, the "I know it all" trait can provide what might appear to be credible answers on any topic, but without true understanding or discernment. The "meek" trait allows ChatGPT to adopt any stance or perspective, which makes it a useful "thinking sparrow" for comparing approaches, assessing options, or roleplaying different arguments. As a "speech virtuoso", ChatGPT excels at language-related tasks like summarizing, simplifying, translating, and changing the register of text.

14. Construct-validated measures are used to measure with accuracy a specific theoretical construct or concept.
15. This paradigm is underpinned by the idea that knowledge and meaning are constructed by individuals through their repeated interactions with the world. They are shaped by their experiences, culture, and social contexts.

It can also perform various types of qualitative analysis (such as sentiment analysis, causality, etc.) and support the dissemination of results through audience-targeted products. This has the potential to democratize evaluations and bring in new voices, but also risks mechanizing the process if used uncritically. Ferretti's work highlights the importance of carefully crafting prompts to obtain meaningful and context-appropriate responses from AI. The author cautions evaluators to remain vigilant in fact-checking – identifying potential hallucinations at every step and critically examining the worldviews and biases ingrained in AI models.

Another area that has received some attention from researchers is that of evaluation education. Zach Tilton et al. (2023) have explored the implications of AI for evaluator and evaluation education. The authors experimented with three chatbot prototypes: a guidance counselor (answering questions about evaluation education programs), a teaching assistant (addressing specific evaluation theories), and a mentor (providing advice on fundamental evaluation practice). These chatbots were built through a combination of tools (like LangChain, Pinecone, TypeScript, OpenAI, Next.js)[16] and were provided pre-processed evaluation-specific documents. They performed better than expected, retrieving information accurately, synthesizing answers from multiple sources, and providing playful responses. However, the authors identified limitations. For example, some answers lacked depth and specificity. Nonetheless, the authors expect that improvements will come with more experimentation with data-retrieval methods, as well as larger context windows and more powerful models.

2.2. AI in climate evaluations

A key finding of this scoping study is that research on the use of AI specifically in climate evaluations[17] is almost non-existent – despite its potential to contribute significantly to the field. There seem to be no articles that have an exclusive focus on the evaluation of climate-related programs with the help of AI. There is, however, substantial research on the use of AI in climate research, which has direct relevance and implications for climate evaluations. Climate research and climate evaluations are connected by their shared reliance on extensive amounts of data and the need for modelling and prediction. Climate research is particularly data intensive. It often requires vast amounts of datasets from various sources, such as satellites, weather stations, ocean buoys, remote sensors, and other instruments like these.

16. LangChain is used for the development of applications that utilize LLMs. Pinecone is a managed vector database service that is used to handle the storage and retrieval of high-dimensional vector embeddings. Vector embeddings are representations of data (such as text, images, or audio) that capture semantic meanings in a format that can be easily processed by ML models. TypeScript is a programming language that is a superset of JavaScript. OpenAI is the company that has developed the GPT model. Next.js is a platform that enables the development of fast, scalable, and feature-rich web applications.
17. The term "climate evaluation" in this paper means the evaluation of climate change interventions or projects.

This data is crucial for understanding climate conditions and trends, as well as making predictions and informing policy decisions. Long-term climate data plays a critical role in the identification of trends and patterns, in tracking temperature or precipitation changes, and in monitoring other climate-related variables. It is also crucial for the assessment of climate change impacts. Also, data on greenhouse gas emissions, land use, and deforestation is key for our understanding of the drivers of climate change and for the formulation of mitigation and adaptation strategies. Furthermore, high-quality data are essential for the calibration and validation of climate models. Given the importance of data in climate research, AI has the potential to play a critical role in data analysis and the broader research process, especially, in areas like climate pattern recognition and predictive modelling. The features of AI that make it suitable for climate research are equally relevant for climate evaluations, as they provide value in assessing the effectiveness of climate policies, identifying areas for improvement, and making policies more evidence-based. For this reason, the focus of the following portion of the literature review is on the use of AI in climate research, with particular emphasis on data analysis, modelling and impact prediction – areas which have direct implications for climate evaluations.

The study by Changlani & Thakore (2023) provides an overview of the role that AI can play in climate research. The authors explore the use of a variety of AI applications in data collection and analysis, climate modelling, extreme weather prediction, and climate impact assessment. They focus in particular on the potential role of AI in the collection and processing of climate data. For example, the authors argue that in remote or hazardous environments climate specialists can leverage AI-powered sensors and drones to automate the data collection process. This will ensure more effective real-time monitoring and reporting of climate-related events. The authors also suggest that NPL approaches can be used to extract useful climate information from textual sources, such as scientific publications and climate reports. AI can also be used to keep datasets up-to-date and relevant. Machine learning can help researchers identify and correct data errors and outliers. It can also be used to integrate information from various sources into comprehensive datasets. AI algorithms can also be used in the analysis of satellite imagery, which helps identify and track climate-related phenomena such as hurricanes, wildfires, and melting ice sheets. AI tools can also help with the analysis and interpretation of remote sensing data, a task which manually is very hard to achieve. The authors also explore the ways in which AI can help researchers improve climate models. For example, deep learning can be used to increase the resolution of climate models to allow for more accurate simulations of climate processes like atmospheric or oceanic circulation. The authors also recommend the use of AI in data assimilation,[18] a process that combines observational data with model outputs to reduce errors and improve the model's performance. AI also has potential to help researchers refine the parameterization

of sub-grid scale processes[19] in climate models, which allows for more accurate parameterizations of phenomena like clouds and turbulence.

Additionally, according to the authors, quantum computing – when and if it comes to fruition on a significant scale – could potentially accelerate climate change simulations. This will enable predictions that are more comprehensive and accurate. AI can also play a role in improving the accuracy of climate predictions. A feature of machine learning is that it is good at recognizing complex patterns and relationships in the data. This ability makes it suitable for the identification of subtle climate patterns that traditional statistical approaches may miss. AI tools can also be useful in the calibration of climate models by using observed data to fine-tune these models to better match observations. They can also be used to create dynamic models that are adjusted and updated based on real-time data – this in turn will allow for more precise and timely predictions, especially in rapidly changing situations such as extreme weather events.

Changlani & Thakore complement their review of AI applications in climate modelling and analysis with the following key examples.

BOX 2: AI Applications In Climate Modelling And Analysis[20]

The following are examples of applications of AI in climate modelling and analysis, as reported by Changlani & Thakore:

- **IBM's GRAF model:** IBM's Global High-Resolution Atmospheric Forecasting (GRAF) model uses deep learning to improve short-term weather forecasts. By analysing vast datasets from various sources, GRAF can provide highly localized and accurate weather predictions, enhancing our ability to respond to extreme weather events.
- **Google's AI for weather forecasting:** Google is working on improving weather and climate forecasting using deep learning. Their AI-driven approach focuses on better predicting complex climate phenomena like rainfall patterns.
- **Microsoft's AI for Earth:** Microsoft's AI for Earth programme supports numerous climate-related projects. One example is the "LandCoverNet" project, which employs AI to map land cover changes, helping researchers track deforestation and urban expansion.
- **Carbon Monitoring for Action (CARMA):** CARMA utilizes data analytics to monitor global carbon emissions. It tracks the carbon output of thousands of power plants worldwide, providing valuable information for climate change mitigation efforts.

(continued)

18. Data assimilation is a method used to integrate real-world observational data with a computational model to improve the model's accuracy and predictive capabilities.
19. Parameterization of sub-grid scale processes is a technique used in numerical models, particularly in atmospheric, oceanic, and climate modelling, to represent the effects of physical processes that occur on scales smaller than the grid resolution of the model.
20. The list in the box is taken ad-verbatim from Changlani & Thakore (2023).

- **European Centre for Medium-Range Weather Forecasts (ECMWF):** ECMWF uses machine learning to improve weather forecasts, including predicting the tracks of hurricanes and cyclones more accurately.
- **Climate Trace:** The Climate Trace uses satellites, other remote sensing techniques, and artificial intelligence to deliver a detailed, independent look at global emissions, and make meaningful climate action faster and easier by mobilizing the global tech community to track greenhouse gas (GHG) emissions with unprecedented detail and speed and provide this data freely to the public.
- **NOAA's Climate Model Emulator (CME):** CME uses machine learning to emulate the Community Earth System Model (CESM), a complex climate model. CME offers faster and more efficient climate simulations, facilitating extensive model runs.
- **National Center for Atmospheric Research (NCAR) AI-enhanced climate models:** NCAR is actively researching the use of AI, particularly deep learning, to improve climate models. Their work involves developing AI-based parameterizations for climate model components, aiming to increase model accuracy.
- **NASA's GEOS:AI model:** NASA's Global Earth Observation System with Artificial Intelligence (GEOS:AI) incorporates machine learning into its climate models. This AI-driven model is designed to enhance the accuracy and efficiency of climate simulations.

Huntingford et al. (2019) argue that the complexity of the Earth climate system and the urgency of addressing global warming present a great opportunity for the application of AI in climate research. They point out that machine learning can be valuable in identifying complex patterns and connections in the vast amount of climate data, going beyond the limitations of traditional statistical analysis methods. The study presents an overview of several key machine learning algorithms and their potential applications, which include the use of neural networks to build ecological interaction equations and Gaussian processes for out-of-sample predictions of future climate states. The study also presents illustrative examples, such as the use of machine learning to understand the drivers of the 2018 summer drought in the UK (the so-called "warming hiatus") and improve climate model parameterizations. Additionally, the study emphasizes the role of AI in translating machine learning-derived insights into actionable information for decision-making, such as early warnings for extreme events like droughts. However, the authors urge caution for a careful selection of algorithms, thoughtful consideration of underlying assumptions, reproducibility, and collaboration between climate experts and AI specialists.

In the area of adaptation, Chen et al. (2023) have looked into AI's application in making meaning out of large volumes of data from satellites, sensors and other sources. One instance is precision agriculture wherein AI can help keep track of crop condi-

tions, predict yields, and optimize resource management that leads to reduced fertilizer and chemical use. In the context of natural resource management, AI can be used to monitor forests, predict deforestation risks, and support the restoration of ecosystems. The authors posit that AI can also be employed in monitoring water resource management through analysis of water quality, consumption patterns, and infrastructure performance. As for natural resource management, AI is capable of monitoring forests, predicting deforestation risks and facilitating ecosystem regeneration. In addition to this point, the authors posit that AI can be employed in monitoring water resources through analysis on water quality indicators, consumption trends as well as infrastructure performance.

“As for natural resource management, AI is capable of monitoring forests, predicting deforestation risks and facilitating ecosystem regeneration.”

Leo et al. (2020) have explored the use of AI in the assessment of climate change vulnerability. They argue that AI can help with the creation of a complex and detailed portrait of climate vulnerability, one that is comprehensive, data-driven, and human-centric. By combining satellite imagery with household survey microdata, AI can provide hyper-local insights into people, communities, and livelihoods, which allows for better targeted and more responsive program design and monitoring. The authors demonstrate this approach with case studies from the mapping of climate vulnerability in Malawi and Mali. They report large variation in vulnerability and its components, like sensitivity and adaptive capacity, between urban and rural areas (and even between neighboring administrative divisions). They conclude that AI has the potential to radically transform impact assessments of climate change programs, primarily through better measurement which not only leads to the construction of better baselines and endlines, but also does not overburden survey participants. The authors also stress the importance of tailoring indicators and models in a way that accounts for key contextual differences between regions and countries.

The work of Burke and Lobell (2017) demonstrates the potential of high-resolution satellite imagery in the assessment of maize yield variation and its determinants in the context of smallholder farms in western Kenya. By combining satellite images with one-meter resolution with extensive field sampling, the authors found that satellite estimates were largely in line with the survey measures, particularly for larger fields where errors in both field and satellite measures were reduced. The study showed that satellite measures can be useful in detecting positive yield responses to fertilizer and hybrid seed inputs, with statistically indistinguishable results from

surveys. This work suggests that the analysis of satellite images can make yield predictions which are as good as traditional survey methods.

Goldblatt et al. (2016) demonstrate how AI can be leveraged for a better understanding of urbanization patterns and dynamics, especially in developing countries where data is limited. The authors use AI for the classification of urban areas in a dataset consisting of 21,030 polygons across India that were manually labeled as "built-up" or "not built-up". These urban areas were detected through satellite imagery from the Google Earth Engine. The study highlights the potential of combining high-resolution satellite imagery, cloud-based computational platforms like the Google Earth Engine, and reliable ground-truth data for mapping the urbanization process at scale. In addition to the benefits, the authors also reported key challenges such as the lack of ground-truth data specifically developed for mapping urban areas. The study also found that larger training sets combined with vegetation indices improve classifier performance.

The 2022 report on the Artificial Intelligence Forum in New Zealand provides a good overview of the applications of AI in the area of environmental sustainability in Aotearoa, New Zealand. The report illustrated the use of AI technologies, which while still nascent are helping process big data, provide near real-time information, enhance predictions and modelling, support decision-making, and gain insights from historical data. The report outlined several examples on the use of AI. One of them is the experience of Lynker Analytics Consortium, a non-governmental group that applied machine learning models to aerial imagery in order to classify land cover into classes such as pasture, harvested land, and mature native forest. This automated monitoring system enabled the rapid assessment of the replanting status for greenhouse gas reporting. As another example, Safeswim, a public information service in New Zealand that provides real-time data on water quality and beach safety, used AI to combine real-time data on wastewater and stormwater networks with data-driven predictive models to forecast water quality at swimming sites in Auckland. In yet another example, Wildlife.AI, in collaboration with community groups and local students, designed a low-powered, open-source device that was used to record videos of ground-dwelling invertebrates and lizards. This data is used to prevent species loss by informing conservation efforts.

In their study of the potential of AI in the identification of climate-related disaster and security risks, Kim and Boulanin (2023) argue that recent advances in machine learning offer opportunities for understanding and predicting climate hazards, managing vulnerabilities and exposure to climate change, and detecting climate-related tensions. The authors single out AI's potential in helping overcome the data scarcity challenge in conflict-affected and fragile countries, which are often the most exposed to climate hazards. Lucas (2024) explores the critical role of AI in environmental mon-

itoring for rapid disaster response by leveraging AI's ability to process and analyse in near real-time vast amounts of data from various sources, such as satellite imagery, sensor networks, and social media. For example, the authors report the use of AI in the analysis of seismic data to predict the likelihood and intensity of earthquakes, as a way of enabling timely evacuations and other response measures. Similarly, AI can be used to monitor social media and identify and locate people in need of assistance during disasters.

Farley (2017) demonstrates the use of anonymized data from Mapbox's[21] traffic sensor network for the analysis of the impact of Hurricane Maria on the mobility and connectivity of the residents of Puerto Rico. The data was visualized to show movement patterns in different regions before and after the hurricane, which revealed the loss of critical infrastructure due to storm damage as well as difficulties faced by the population during slow recovery.[22] The data helped researchers track the long road to recovery and also raise some key questions about the factors contributing to the slow recovery in different areas, such as power supply, connectivity, or damages to the road infrastructure. With this study, Farley showcases the value of leveraging large-scale, anonymized data from sensor networks to gain insights into the impact of natural disasters on human mobility and infrastructure, which could potentially inform disaster response and recovery efforts.

Taghikhah et al. (2022) explore how AI can be used in predictive models – such as ML and computer vision – to locate and map bushfire vulnerabilities. They provide the example of ML algorithms which are applied to satellite images to assess vegetation distribution and identify changes. This enables researchers to automate precision operations for afforestation and manage bushfires and deforestation more effectively. The study provides insights into the complex behaviour of fire emergencies and helps responsible officials develop effective firefighting strategies.

In the area of mitigation, AI has shown potential for helping with the measurement and monitoring of carbon emissions. For example, Dannouni et al. (2023) explore how AI can help organizations to understand and to reduce their carbon footprints. They report the example of CO2 AI, a company that provides innovative SaaS platforms that leverage AI to enable organizations to map emissions across their value chain (from the production stage to the final product) and develop actionable insights to drive climate action. These platforms use AI to collect emissions data, match it to a

21. Mapbox is a location data platform that provides mapping, navigation, and location-based services to developers, businesses, and organizations.
22. Before the hurricane the data showed normal daily movement patterns - people commuting to work, traveling home, and relaxing on weekends. However, in the aftermath of Maria, the lack of electricity, communication networks, and accessible roads limited severely regional mobility, primarily to the capital city of San Juan. Even 35 days after the storm, mobility was only at 27 per cent of pre-storm levels.

company's activities and products, simulate potential solutions, and help the company build decarbonization action plans. In the renewables sector, Taghikhah et al. (2022) have shown how AI can be used to optimize power generation, as well as help with the planning of energy storage and the forecasting of demand. Machine learning algorithms can also be used to develop predictive models for wind power generation at different time scales. Taghikhah et al. also outline the usefulness of AI in identifying conservative energy consumers, learning their profiles and preferences, and developing strategies for motivating behaviour change for greater efficiencies.

2.3. Opportunities, risks and mitigation strategies

As shown throughout the previous two sections, the literature reviewed for this study clearly suggests that AI has the potential to facilitate significantly to the evaluation process, adding not only to the efficiency of the process, but also the quality of data and analysis. The majority of researchers and practitioners featured in this study believe that the application of AI tools in evaluations is inevitable; in fact, the speed of adoption is expected to increase at an accelerating rate. The implications of this will be transformative for the evaluation profession and field. Researchers like Head, C. B. et al. (2023) have estimated that as many as two-thirds of evaluation tasks could be affected by LLMs in the next five years.[23] They think that AI's full impact will likely come from tailored applications built on top of pre-trained LLMs (such as ChatGPT). These kinds of applications will rely on specialized "prompt engineering" to optimize evaluation tasks, fine-tune AI models on relevant documentation, and integrate LLMs into existing evaluation tools and processes. In a policy paper titled "Governing with Artificial Intelligence: Are governments ready?" by Ubaldi and Zapata (2024), the OECD explored the growing use of AI by various governments and the policy challenges that this new phenomenon entails. The study analysed a range of AI use cases in the public sector and identified specific tasks that had benefitted from the use of AI. The paper concluded that the responsible use of AI has the potential to substantially increase public sector productivity, make public policies and services more responsive to evolving citizen needs, and strengthen the government's accountability. For all the promise of the use of AI in evaluation, there are also substantial risks. Many of these risks have been identified in stark terms in the existing research. First, AI does not operate in a void. AI-generated outputs are largely dependent on the quality and representativeness of the data used for training, which raises concerns about bias, reliability, and scalability (Ferretti, S., 2023; Head, C. B. et al., 2023, Thornton 2023, Kim & Boulanin 2023). Thornton (2023) cautions that AI models are only as solid as

23. The authors shared a hypothetical scenario and list of evaluation activities with nine MERL experts and asked them to assess the potential impact of the use of LLMs on the level of effort (in practice, this is often measured in days needed for the evaluation team) required for the implementation of each activity. Respondents had the option of scoring the potential impact on a scale from 0 per cent (representing no impact of LLMs on the level of effort required to implement the activity) to 100 per cent (LLMs would replace all the work of the evaluation team in this activity).

the data they are trained on. As such, these models reflect the biases and structural inequalities that are present in the underlying data. Similarly, Head, C. B. et al. (2023) warn us that LLMs have major limitations and risks – models may hallucinate false information, reflect the biases which are ubiquitously present in the Internet data used for the training (e.g., overrepresenting English, wealthier and urban populations), and pose the "black box" interpretability challenge. In a similar vein, Kates and Wilson (2023) identify several challenges and risks associated with the use of AI in evaluations. These include the possibility of low-quality outputs, the sudden disruption of the evaluation market and professional roles, and the misalignment of AI with human values (the so-called "alignment problem").[24] The authors advise evaluators to get hands-on experience with various AI tools, use them transparently, think critically about AI's effects on their evaluation work, and maintain a clear perspective on the technology's potential implications for the profession. They conclude by calling for more empirical research on AI's impact on evaluation, as well as regular updates to professional guidelines that address the use of AI in various organizations and environments.

Based on his personal experience, Thornton, I. (2023) reports that AI excels in tasks that involve sorting, summarizing, or classifying data, but is less reliable in interpretation, extension, or validity checking. As he puts it, "there is not yet a one-size-fits-all distinction between what kinds of work are better done by AI and what kind of work is better done by humans. These technologies are too new to reliably conclude what complex tasks AI is genuinely better at than humans are. The determining factor is more often what kinds of error are acceptable, rather than how much." Thornton notes the emerging use of AI techniques in the evaluation field, while also highlighting the slow pace (so far) of adoption of AI in the evaluation practice. His article also covers important ethical considerations – in particular, AI's biases and the need for careful and responsible use of AI tools in evaluations. Thornton emphasizes the need for evaluators to understand AI in order to effectively use it or avoid using it. He also calls for further empirical work that assesses the quality and impact of AI contributions in the evaluation practice.

Hitch (2023) focuses on the ethical dilemmas that result from the use of AI in qualitative research, such as the issue of data privacy and the potential biases built in the underlying algorithms. The author highlights the importance of being mindful and critically aware when using AI in reflexive thematic analysis, seeing it more like a tool that helps with certain tasks, rather than a replacement of human judgement and expertise. Hitch makes the case for the development of strong guidelines for the ethical use of AI in qualitative research and predicts a gradual transition towards AI-aug-

24. The alignment problem refers to the challenge of ensuring that the goals, behaviours, and decision-making processes of AI systems align with the intentions, values, and safety requirements of humans.

mented approaches in the field.

Another key angle of AI that researchers have explored is its propensity for cultural biases.

Christou's work (2024) focuses on the main risks and limitations of AI. His main concerns include the potential loss of human interpretive and evaluative skills, the elevated risk of biases, and key ethical considerations related to the use of data that underlies AI models. Focusing on thematic analyses, the author provides researchers, especially young analysts, with advice on how to use and document AI tools in each phase of the process. Christou emphasizes the importance of maintaining a balance between AI-driven analysis and human interpretation to ensure rigor and contextual relevance. The author also puts forward an illustrative table that outlines the opportunities, risks, and criteria that researchers should consider when using AI tools in each stage of the thematic analysis. Christou concludes by emphasizing the importance of not letting AI overshadow the analyst's critical evaluative and interpretive skills. Instead, the author calls for the use of AI as an aid in thematic analysis, a tool that enhances the depth and breadth of the analysis, provided that certain quality criteria are adhered to.

Another key angle of AI that researchers have explored is its propensity for cultural biases. An example of this is the work of Vinuesa et al. (2020). These researchers have examined the interaction of cultural values and regional (context-specific) needs with the way AI systems are developed and used. The authors examine how AI solutions designed without contextual knowledge or engagement with the local communities, especially in the context of conflict-affected and fragile nations, exclude key variables and have unintended negative consequences. For instance, an AI system optimized for narrow climate change goals will fail to account for social-ecological aspects of the land, which are crucial to indigenous groups. Similarly, that model might also ignore endangered species valued by conservationists. This lack of cultural and contextual understanding may generate resistance from the stakeholders or might even lead to outright conflict. In another study, focused on the ethical use of AI and its impact on public goods, Reid (2023) argues that AI may perpetuate collateral damage, such as algorithmic bias and new forms of racism and prejudice. The author invites evaluators to ask critical questions about who benefits and who loses from the use of AI in evaluation.

Thus, an effective integration of AI in the evaluation process requires a careful consideration of the context, AI's limitations, and its ethical implications. As Nielsen, S. B. (2023) has argued, while AI has the potential to affect significantly the evaluation industry, it is unlikely to disrupt entirely the need of the evaluation process for spe-

cialist knowledge and judgement. Evaluators bring to the process a crucial understanding of theory, causality, validity, ethics, equity, and judgement. Azzam (2023) identifies the collaboration of humans and AI as key to a productive approach. AI's potential is unleashed when AI's pattern recognition and analytical power is combined with human skepticism, contextual understanding, and critical perspective. The author provides several examples of this. For example, AI can be used to identify statistical anomalies, while humans interpret biases. AI can also act as a second coder in qualitative analysis, while humans verify its trustworthiness. Or as another option, AI can be used to detect hidden patterns, while humans make meaning out of them. Azzam argues that ongoing testing, research and experimentation are needed to fully understand and realize the benefits of AI in evaluation, while at the same time actively identifying and mitigating potential risks.

There is also one peculiar AI risk directly related to the environment and climate change – this is the impact resulting from the growing needs of AI's computational infrastructure for energy and natural resources (Chen et al., 2023; Cowls et al., 2021; Dannouni et al., 2023; Kaack et al., 2020). The 2022 OECD report "Measuring the Environmental Impacts of AI Compute and Applications: The AI Footprint" examines the direct and indirect environmental effects of the use of AI. The report raises concerns about AI's environmental impacts, primarily through greenhouse gas emissions, and recommends the establishment of a comprehensive measurement framework that allows for the benchmarking of national AI initiatives and their environmental footprint. Cowls et al. (2021) have made an attempt to assess the carbon footprint of AI research and the factors that influence AI emissions (this was before the emergence of Open AI's ChatGPT). They found that this carbon footprint is significant and emphasize the need for more research and evidence on the trade-off between emissions from AI research and the efficiency gains enabled by AI. Similarly, Dannouni et al. (2023) have estimated that the cloud and hyperscale data centres that support the AI systems may account for 0.1 per cent to 0.2 per cent of global greenhouse gas emissions. Furthermore, while water-based cooling remains the most energy-efficient option for the data centres, it puts pressure on global water resources.

The existing literature shows that AI comes with its own set of risks and challenges. Researchers recommend a multi-dimensional approach for addressing them. One suggestion is to train AI on more diverse and representative data. This will help reduce bias and the risk of excluding certain groups. Another suggestion is more rigorous validation of AI outputs. Another one is the development of stronger ethical frameworks and the active involvement of stakeholders in the assessment of results. Researchers also recommend the use of AI in a context-appropriate manner, recognizing that human judgement and critical thinking will still have to play a vital role. Other suggestions include regulating AI's environmental footprint, as

well as promoting transparency by sharing advancements and data with the public. Ultimately, a combination of technical improvements, oversight, thoughtful application, human involvement, and openness, will help evaluators and organizations better able to navigate risks and mitigate AI's potential downsides.

The mitigating actions they propose include rigorous testing and validation and greater human-AI collaboration to critically review and interpret the outputs generated by AI.

The research reviewed for this study provides several specific recommendations. Head et al. (2023) raise the need for the evaluation community to engage more critically with LLMs, assess them more systematically for bias, exclusion, risk, and societal impact, and ensure that the affected communities have a voice in their design, validation, and accountability. The mitigating actions they propose include rigorous testing and validation and greater human-AI collaboration to critically review and interpret the outputs generated by AI. They also make the case for the development and validation of AI tools based on more diverse datasets which represent more diverse evaluation contexts. As Thornton (2023) put it, *"for evaluators, AI can serve as both an object (a tool applied in evaluations) and a subject (something to be evaluated itself)."* Thornton outlines some methods that can be used in the assessment of AI models. They include qualitative approaches that assess factors like user satisfaction and trust, mixed methods that incorporate user feedback and biometrics, and quantitative metrics that compare the performance of models. Sabarre et al., (2023) maintain that evaluators must thoroughly assess the evaluation context and needs before choosing an AI tool. Ferretti (2023) emphasizes the importance of human judgement and critical thinking in evaluation, even when using advanced AI tools.

In a paper by Raftree and Tilton (2023) presented to the 2023 American Evaluation Association (AEA) Conference, the authors discuss several concerns about the use of AI. In particular, they highlight the increasing need for guidance on the use of AI in evaluations. Similarly, Reid (2023) makes a strong case for the formulation of ethical guidelines for AI's use in evaluations, whereas Montrosse-Moorhead (2023) emphasizes the importance of engaging stakeholders in the design and implementation of evaluations that are aided by AI in order to make sure that their values and concerns are taken into account. Montrosse-Moorhead also calls for the establishment of guidelines and protocols for the responsible use of AI, including requirements for human oversight and validation. Based on a synthesis of the articles in the special issue of the "New Directions for Evaluation" journal, Montrosse-Moorhead, B. (2023) pro-

poses eight criteria domains for evaluating the use of AI in evaluation practice. These domains are organized into two categories: (i) those related to the conceptualization and implementation of AI in evaluation practice; and (ii) those focused on outcomes resulting from the use of AI in evaluation. Through these criteria, Montrosse-Moorhead proposes a framework for assessing the appropriateness, effectiveness, and impact of the use of AI in evaluations. Another study that focuses on the need for guidance is that by Tilton et al. (2023). The authors highlight the importance of the context for the use of AI in evaluations and make a call for comprehensive training and education for evaluators on the appropriate use of AI.

Motivated by concerns about the risks of AI, an article by Kaack et al. (2020) proposes several policy measures in two areas: (i) regulating the impact of the AI infrastructure on emissions; and (ii) facilitating the sharing of data with the public. In the first area, the authors recommend economic incentives and regulatory requirements for emission reductions, including transparency and reporting requirements for emissions and energy consumption related to AI operations (this includes life-cycle impacts and externalities). In the second area, the authors recommend the sharing of data in the public domain and the establishment of processes that enable the incorporation of stakeholder feedback throughout the process of deploying AI. The authors also advocate for the development of standards and best practices to guide decisions on when and how AI should be used. This includes standards around data collection, management, and sharing that take into account privacy concerns and data control considerations.

In conclusion, most researchers reviewed in this study suggest that evaluators and organizations must embrace AI, albeit critically. There are researchers like Nielsen, S. B. (2023) who make an explicit case for this. In the same vein, the article by Sabarre et al. (2023) makes the case for small evaluation businesses to embrace AI tools to increase their value and remain competitive in the marketplace. Similarly, Raftree and Tilton (2023) emphasize that, despite the risks, there is a need for the evaluation community to proactively engage with these technologies to avoid the risk of becoming less relevant as data scientists and computational statisticians increasingly take on evaluation work.

While there are no doubts about the usefulness of AI in the evaluation process and the inevitable trend of its increasing use by the evaluation community, a crucial advice that emerges from the review of this literature is the need for a careful and balanced approach in the adoption process that involves close human supervision. Furthermore, this approach should be gradual, through a process of continuous verification, improvement and adaptation. As in other fields of human endeavour, individual evaluators, evaluation organizations, and educational institutions must gradually adapt to the changes brought about by AI. They should embrace a culture

of continuous improvement and adaptation. As Sabarre, N. R. et al. (2023) suggest in their article for the special edition of the “New Directions for Evaluation” journal, the process for the adoption of AI in evaluation will require regular assessments of its effectiveness and appropriateness, adjustments along the way and the continued incorporation of new developments and best practices as they emerge.

III. Experience of international organizations and evaluators

In addition to the literature review, this scoping study also collected information about the experiences of the international organizations and individual evaluators with the use of AI in the evaluation process.

3.1. Experience of international organizations

This section synthesizes the experiences and insights of several international organizations in leveraging AI for evaluation-related activities. These are the organizations which at the time of the study were identified as having experimented with and used AI in evaluations.

Climate Investment Funds (CIF)

The CIF team has not yet used AI extensively in evaluations – only for basic tasks such as using ChatGPT to generate evaluation summaries or convert text to bullet points. The CIF team sees potential for the use of AI in synthesizing evidence; for example, in diagnostic work to quickly understand a country's context and policies, assessing the alignment of proposed projects with CIF country plans, and synthesizing insights from a set of evaluation reports on topics like stakeholder engagement or transformational change (e.g., thematic reviews across large corpuses of evaluations). However, the CIF team has noted that the AI outputs often introduce extraneous information or miss nuances, which requires careful human review.

Some key challenges identified by the CIF team include data security concerns connected to the uploading of sensitive/internal documents in external AI platforms,[25] the limitations in the interoperability of the AI tools across the various Multilateral Development Banks or UN agencies,[26] and the rapid pace of change in AI tools, which makes their examination and testing quickly obsolete.

The CIF team emphasized the importance of developing practical protocols and guidance for evaluators on how to use AI in the evaluation process responsibly. They also stressed the need for building the capacity of evaluators and evaluation offices

25. The MDBs that CIF works with have different disclosure policies and security considerations.
26. For example, the World Bank's internal "MyAI" tool can only access and analyse World Bank documents and not those of other MDBs.

to understand the potential and the limitations of individual AI tools.

Global Environment Facility (GEF)

GEF's Independent Evaluation Office (GEF IEO) has been experimenting with AI and machine learning tools – particularly supervised classification and regression models – for over 15 years now. The team has prioritized the analysis of geospatial data, including satellite images, for the assessment of changes in land use, forest cover, and biodiversity as a result of interventions funded/supported by GEF. More recently, the team has started experimenting with the new-generation generative AI. The following is a brief overview of GEF IEO's experience in this area.

- GEF IEO reported that the use of AI algorithms has generally improved the accuracy of climate change models and has enabled the organization to better estimate changes and relate them to its contributions – even for interventions not directly supported by GEF (for example, quantifying the carbon sequestered in GEF-supported areas).

- The team has used regression models to predict the sequestration of carbon under various scenarios. These models have provided valuable ex ante information for the assessment of the effectiveness of GEF projects and setting realistic project targets. For example, a key finding that GEF IEO has identified through this process is that a project's maximum impact is seen about five years after project completion.

- GEF IEO has also combined AI-analysed satellite data with survey data from sources like the World Bank's Living Standards Measurement Study (LSMS) and USAID's Demographic and Health Surveys (DHS) to identify key factors of success (e.g., access to electricity).

- More recently, GEF IEO has been using generative AI tools, particularly in code optimization, debugging, and the analysis of qualitative data. Other use cases include summarizing documents, transcribing videos, translating documents in multiple languages, and designing communication materials using tools such as Midjourney and DALL·E3.

- In terms of tools, the GEF IEO team uses Bard/Gemini27 (pro version) and ChatGPT as the main generative AI chatbots, along with more specialized models like Aurora for satellite data analysis, and have been testing the

27. Previously called Bard, Gemini is Google's next-generation LLM that competes with other advanced AI models in the market, such as OpenAI's GPT.

potential of GEOAI models such as Clay.

- The GEF IEO team has conducted a comparative experiment which showed that generative AI could perform well in the identification (extraction) of socio-economic benefits from project documents. However, it also missed certain key aspects when compared to manual coding.

- The GEF IEO team also noted that while AI has reduced time spent on coding and enabled more innovative data analysis, it also requires continuous learning and experimentation to stay up to date with the rapid developments in the field. As a counteracting approach to AI biases and hallucinations, the team emphasized the need for human oversight, knowledge of the specific domain/context where AI is applied, and careful engineering of prompts. The team also emphasized the need to train the AI and work with it in an iterative way, while being hands-on and experimental in setting up and using these tools.

- Regarding data privacy concerns, the GEF IEO team advised against uploading sensitive, non-public data to cloud-based AI platforms. The team suggested using APIs to access AI models without sending data to external AI servers. The team highlighted the World Bank's approach of using Azure to train OpenAI models in-house on its own documents.

- For those evaluators/teams who don't have sufficient resources for proprietary AI systems, the GEF IEO team recommends anonymizing the data before sending it to AI servers and exercising elevated caution when applying AI to sensitive information.

International Fund for Agricultural Development (IFAD)

Over the past six to eight months, IFAD's Independent Office of Evaluation has been actively exploring ways of using AI for the improvement of the quality and efficiency of the organization's evaluation work. Their approach has focused on portable AI solutions across evaluations. IFAD sees significant potential for AI to strengthen the evidence and insights of evaluations. But the team also recognizes that realizing this potential will require continued exploration, learning, and investment. The following are some key insights from IFAD's experience.

- IFAD has engaged with other UN agencies and IFIs to understand their experiences with AI based on their specific needs, resources, and capacities. This has helped IFAD in designing its own strategy and approach.
- IFAD started with simple AI solutions and has been gradually progressing towards more complex ones, with a focus on "learning by doing" and

combining evaluation expertise with data science skills.

- IFAD has tested the applicability of AI tools in four evaluation areas:
  - **Intervention targeting:** IFAD has used ChatGPT for the development of a QGIS map which is used to assess the geographic relevance of the targeting of its interventions. The AI approach produced the map within 12 minutes, which was much faster than the traditional method. The output was validated by a GIS specialist, who confirmed the accuracy of the result.
  - **Document synthesis:** IFAD has tested the ability of ChatGPT to synthesize information from multiple documents by having it answer ten evaluation questions based on 14 supervision reports and a mid-term review. The process took only 42 minutes (2 minutes for data preparation and 40 minutes for processing). This test confirmed that AI can help consolidate large information to address a specific query.
  - **Text analysis:** IFAD used AI to extract specific information from project documents (in this case, to identify evidence of environmental impacts and associated mitigation measures). The AI results were manually validated by the team with spot checks. The test confirmed ChatGPT's potential to quickly identify key themes and evidence from large text-based datasets.
  - **Analysis of non-lending activities:** IFAD used ChatGPT to extract and analyse information about non-lending activities from various documents. ChatGPT created a structured database that enabled the generation of descriptive statistics, graphs, and tables. This test demonstrated the ability of AI to process very fast complex project data and visualize it with efficiency.
- IFAD reported challenges related to API request limits, hardware limitations, performance degradation during long runs, and token[28] limit management. Solutions suggested by the IFAD team include using corporate API subscriptions, cloud computing platforms, built-in rest periods for the AI, tools for monitoring token consumption, and breaking down (chunking) large documents.
- The following are key lessons identified by IFAD:
  - Without proper and well-structured prompts, AI tends to hallucinate. To prevent misunderstandings and to ensure consistency, it is important to

28. A token is a basic unit of text that large language models process. Tokens can be as small as individual characters or as large as entire words, depending on how the text is tokenized.

make clear and specific requests and avoid conflicting terms. This is a time-consuming, but essential process.

- AI may show reduced performance (getting tired) if it is run extensively. To mitigate this risk, a "sleep function" can be used in the script to give the system a break and prevent performance degradation.

- Large-size data and local hardware limitations may slow down processing times and reduce scalability. This can be mitigated by leveraging cloud computing platforms like AWS and Azure, which help improve scalability.

- When processed in one go, large documents may exceed the token or API request limits. This can be addressed by splitting large documents into manageable pieces, keeping each chunk within the token limits.

- Another challenge is the limited number of APIs which individual users can request per day and minutes. This limitation can be overcome by using a corporate subscription.

- The iterative prompt engineering process is time-consuming. Thus, it is crucial to find the right balance between automation and manual input.

- It is also important to ensure that the maximum output is less than the total available tokens minus the sum of the prompt and input tokens. The TikToken library[29] can be used to monitor the consumption of tokens and prevent information leakage due to token overflow.

- IFAD's team emphasized the importance of validating AI-generated results through methods like spot checks and comparison with human expert opinions as a way of ensuring accuracy. Human supervision is seen as essential.

- While AI offers significant potential for enhancing the quality and efficiency of evaluations, IFAD emphasizes the importance of maintaining high-quality standards and investing in human capacities.

United Nations Children's Fund (UNICEF)

UNICEF's evaluation office reported that it is in the early stages of using AI. It is mainly focused on the synthesis of qualitative data from past evaluations using machine

29. The TikToken library is a Python package used for tokenizing text specifically for OpenAI's language models, such as GPT-3 and GPT-4.

learning and natural language processing. It aims to use AI to mine the organization's evidence base to inform evaluation design and reduce data collection efforts. UNICEF is moving cautiously in this area and is not using generative AI models. The following are key points from UNICEF's experience.

- UNICEF's evaluation office has built a small data pod and has hired a data scientist to support AI applications. The challenges it faces are due to the predominantly qualitative nature of their evaluations and the lack of highly structured and machine-readable data.

- AI has so far been primarily used for the extraction of insights from a large database of past evaluations, assessments, and studies. The aim has been to inform the design of new evaluations and avoid duplicating data collection. This work has involved techniques like text mining, NLP, and supervised machine learning to classify and summarize needed information.

- AI is seen as a potentially useful tool for making the organization's evaluations more efficient, cost-effective, and better informed by past evidence. However, UNICEF emphasizes the importance of human judgement and validation of any insights generated with the help of AI.

- UNICEF is cautious about the use of generative AI models like ChatGPT because of concerns about possible hallucination and biases. For data analysis the team prefers to use open-source tools and libraries like Python.

- UNICEF highlighted the importance of structuring evaluation data in machine-readable formats and standardizing evaluation reports to make them more conducive to AI analysis. UNICEF is working on developing data governance frameworks and ethical guidelines for the use of AI.

- Other UNICEF divisions, such as the data science team and the innovation units, are more advanced in using AI and big data for monitoring, real-time analysis, and geospatial applications. The evaluation office is learning from their experience.

United Nations Development Programme (UNDP)

UNDP's Independent Evaluation Office has made a lot of progress in developing and using AI tools for evaluations. A unique experience of UNDP is the development of its own proprietary search and extraction tool called AIDA (Artificial Intelligence for Development Analytics) which is powered by AI. The following are key points from UNDP's experience.

- UNDP started using AI in 2021 when it developed AIDA (Artificial Intelligence for Development Analytics). This system uses machine learning to extract findings, conclusions and recommendations, enable searching and filtering by various criteria (e.g., country, region, theme, time period, evaluation quality), and provide summaries and insights from over 6,500 evaluation reports which are stored in UNDP's Evaluation Resource Center (ERC) database. AIDA includes features like translation of non-English reports, sentiment analysis at the sentence and paragraph level, and a taxonomy to label paragraphs by topic.

- The development of AIDA involved extensive work on data preparation, model training, and validation. UNDP worked with external researchers to refine its sentiment analysis models and adjust them to the context of the UN. AIDA is not seen as exclusive, but as a complement to other external AI tools like Copilot (used, for example, for drafting text).

- The IEO team reported that the uptake of AIDA has thus far been good – including by UNDP country offices which use it to formulate projects/programs based on evidence derived from evaluation reports. The system is also used by UNDP staff and evaluators to inform the design of evaluations (e.g., by identifying gaps in evidence or trends), support the synthesis of evidence, and generate insights for decision-making. AIDA has been shared with other UNDP teams (e.g., accelerator labs), which use it as a knowledge management tool.

- UNDP continues to refine and expand AIDA (e.g., by integrating more data sources), while carefully monitoring risks related to data security and privacy. AIDA's architecture has the potential for further expansion to support other UN entities, but currently it is only used by UNDP. UNDP is exploring the idea of a single UN-wide repository for evaluations that could be partially analysed with the help of AI.

- For data analytics, UNDP uses other tools in the public domain, such as Power Query.[30] The UNDP IEO team has also set up an internal version of ChatGPT (Microsoft's version) that staff can use for tasks like improving writing or extracting key insights, on the proviso that sensitive data is not uploaded.

- UNDP emphasizes the importance of using AI tools responsibly and ethically in evaluation tasks. The focus is on applications where the risks are lower, but the

30. Power Query is a data transformation tool available in Microsoft Excel, Power BI, and other Microsoft products.

efficiency gains significant. The UNDP IEO team stresses the need for building evaluators' capacities and strengthening the guidance on the use of AI.

United Nations Office of Internal Oversight Services (OIOS)

The OIOS team has tested various AI tools for tasks like document search, data analysis and synthesis as part of several evaluation pilots. The following are key points from OIOS's experience.

- OIOS has tested the use of AI in the evaluations of ***MONUSCO*** (UN peacekeeping mission in the Democratic Republic of Congo) and ***IRMCT*** (International Residual Mechanism for Criminal Tribunals). NLP was used in the evaluation of ***MINURSO*** (UN Mission for Referendum in Western Sahara).

- During the planning phase of the MONUSCO evaluation, the OIOS team tested some AI tools available publicly for searching documents (UNDP's AIDA, RDiscovery, Consensus)[31] and analysing documents (Petal, Lateral, Scholarcy, AI LYZE, ChatGPT).[32] While these tools helped identify and analyse some documents, the team encountered challenges with too many or non-specific results, privacy concerns, and hallucination.

- For the IRMCT pilot, the OIOS team developed jointly with the UN Operations and Crisis Centre (UNOCC) a customized AI tool, which was used to analyse 70 structured interview transcripts. The tool was developed on a secure UN server and included prompts engineered to generate summaries and insights, as well as sentiment analysis. The AI-generated output was compared with analysis done manually with NVivo to assess its accuracy.

- For the MINURSO evaluation, the OIOS team used NLP techniques for the sentiment analysis of Security Council speeches and bilateral communications related to Western Sahara. The team also conducted a thematic network analysis using the Global Database of Events, Language and Tone (GDELT).[33] Furthermore, the team used the spaCy NLP[34] library to parse and classify events from 31,800 daily situation reports, automating geocoding to map incidents in Western Sahara.

31. RDiscovery and Consensus are tools that use AI and machine learning to manage and analyse large volumes of data.
32. Petal, Lateral, Scholarcy, and AI LYZE are tools that leverage AI for various aspects of research, data analysis, and information management.
33. The GDELT Project monitors the world's broadcast, print, and web news in over 100 languages and identifies people, locations, organizations, themes, sources, emotions, counts, quotes, images and events features in news. It serves as a free open platform for computing on the entire world. (https://www.gdeltproject.org/)
34. spaCy is an open-source Python library for advanced NLP. It is used in production environments, providing a fast, reliable, and efficient way to work with large volumes of text data.

- The OIOS team has drawn the following key lessons from this experience: (i) AI has potential to support various evaluation stages and processes, but the tools need to be carefully piloted, tailored to context, and validated; (ii) the available commercial AI packages come with different strengths and limitations; (iii) it is necessary to further pilot AI tools alongside traditional methods to assess their reliability and accuracy; (iv) significant programming, machine learning, and analytical skills, as well as financial investments, are required for customized tools; and (v) ethical considerations around data privacy and explaining AI use in evaluation reports are critical (including adherence with the UN Secretariat guidelines on responsible AI use).

United Nations Population Fund (UNFPA)[35]

UNFPA has accumulated some good experience with the piloting of AI in evaluations. To guide this work, the organization has prepared a comprehensive strategy (2023-2025) for leveraging generative AI in evaluations. The document places emphasis on ethics and responsible use of AI. The following are key features of UNFPA's experience.

- UNFPA's journey began with a needs assessment across the evaluation lifecycle. This assessment mapped several use cases of AI. It also included a study on "solutions exploration" based on UNFPA-approved Google tools and platforms (Duet AI,[36] Bard/Gemini).

- The UNFPA team identified and prioritized five use cases for AI in evaluation: information synthesis, initial report writing, inclusive dissemination, evidence synthesis, and qualitative data analysis.

- The team developed a pioneering *"Strategy for Gen-AI powered evaluation function at UNFPA"* to optimize evaluation processes and products, while ensuring ethical and responsible use of generative AI. The strategy provides a framework for recognizing and mitigating the challenges and risks emanating from the use of AI (such as data protection and privacy, mitigation of bias, accuracy and reliability, transparency and accountability, and overreliance on AI tools). It is aligned with the UNEG ethical principles for AI use in evaluation.

- The strategy outlines key principles, including a demand-driven approach, diversification and innovation of generative AI tools, upholding quality and credibility, adhering to an ethical and human rights-based approach, and

35. The Team Lead for Communications, Knowledge Management, and AI at the UNFPA Evaluation Office co-chairs the UN Evaluation Group working on data and AI.
36. Duet AI is an AI-powered tool developed by Google to assist with real-time collaboration in Google Workspace applications such as Google Docs, Sheets, Slides, and Meet.

promoting generative AI capacity, especially in the Global South.

- UNFPA is in the process of implementing the strategy through a phased approach. Under this framework, the organization is focusing on the development of customized generative AI solutions, change management and communication, an iterative and adaptive approach to digital transformation, and the issue of long-term sustainability.

- The team noted that it is taking a slow, but intentional, approach in rolling out the use of AI in the evaluation practice. The focus is on staff training and capacity building. The team is integrating AI considerations into its evaluation quality assessment framework.

- Two major generative AI pilots are underway: (i) Using AILYZE for the qualitative analysis of 75 country program documents to assess strategic shifts and accelerators across two strategic plan periods, and (ii) Using generative AI for report extraction and coding/synthesis in an inter-agency meta-synthesis of evaluation reports on youth education and employment. The pilots involved close human-AI collaboration and validation, which creates challenges for evaluation timelines.

- UNFPA pays particular attention to the ethical use of AI and risk mitigation. Key measures to this end include contractual clauses, human-AI collaboration, strong human oversight, data protection, and transparency about AI use in evaluation reports.

- The team emphasized the importance of building AI capacities, aligning analytical frameworks for AI use, allocating time for accuracy checks and validation, including AI disclaimers and methodology explanations in reports, and assessing efficiency gains before any scaling up.

United Nations Secretariat

In July 2023, the UN Secretariat issued guidance on the "Responsible Use of Publicly Available Generative Artificial Intelligence (AI) Tools". The guidance recognizes the potential of AI tools to enhance productivity. It emphasizes the need for caution and responsible use. Further, it underscores the importance of protecting sensitive information and personal data, warning against uploading such content to public AI platforms. The guidance stresses that AI-generated content can be unreliable and might perpetuate biases or create security vulnerabilities. It recommends thorough verification of all AI outputs. The guidance reminds UN staff to remain fully accountable for any content produced using these tools. For IT specialists, the guidance offers

additional considerations regarding cost absorption, human oversight, transparency, and risk assessment when implementing AI in UN projects.

World Bank

The World Bank Group has created an internal mAI tool, which is widely used within the organization for research purposes, but also in the evaluation practice.[37] With specific regard to evaluation tools, the World Bank's Independent Evaluation Group (IEG) has mostly focused on non-generative AI techniques such as NLP and computer vision. Over the past three years, IEG has made significant progress in incorporating these techniques into evaluations, starting with simpler applications and gradually increasing complexity to build trust and understanding. More recently, the IEG has also been experimenting with generative models, which includes ChatGPT, GPT-4 via API, (powered by GPT-3.5), and Google's Bard/Gemini, to understand their potential and limitations for various evaluation tasks. These experiments were designed with different user profiles in mind (e.g., data scientists, analysts, and evaluation managers) and compared AI-generated outputs to those produced manually by staff members.[38] The following are some key insights from the experiments of the IEG.

- The following are applications which were identified as successful:
    - Document summarization: ChatGPT was able to summarize large reports based on key sections, although token limits created challenges.
    - Coding: ChatGPT excelled at generating code, adding comments, and summarizing scripts when given specific instructions.
    - Sentiment analysis: GPT-4 outperformed IEG's best model in classifying the sentiment of sentences, achieving 94.5 per cent accuracy compared to 86.8 per cent.
    - Econometric analysis: ChatGPT generated the "R code" for multivariate regression analysis, replicating study results when provided with specific instructions.
    - Text classification: ChatGPT and GPT-4's API achieved a 76 per cent accuracy in classifying text related to disaster risk reduction.
- The following are applications which were considered less successful:

37. mAI is not specifically created for evaluation purposes. It is accessible to all World Bank staff and is designed to protect confidential data by preventing its use in training public GPT models.
38. In her article "What has the World Bank's Internal Evaluation Group learned from experimenting with NLP?", Richards (2023) discusses the World Bank's experiments. Another team from the World Bank's IEG explored potential applications, benefits, and limitations of GPT and generative AI in evaluation practice by conducting nine experiments covering various stages of the evaluation process, user profiles, and output types, and comparing the AI-generated results with outputs produced by humans (Raimondo et al. 2023).

- Generating synthetic images for geospatial analysis: Open AI's DALL-E[39] struggled to generate sufficient urban images for data augmentation, with limited output per prompt and difficulty assessing real images.

- Literature reviews: ChatGPT and mAI provided seemingly "plausible" responses. However, verifying the information's authenticity and mitigating the risk of hallucinations proved challenging. In particular, ChatGPT generated seemingly convincing but fabricated references.

- Evaluative synthesis: ChatGPT produced a well-written synthesis of project evaluations but fabricated the evidence and examples.

The IEG team advises WB staff to not input private or confidential information into the public versions of GPT. It also cautions them against using AI for complex tasks, where the risk of incorrect or hallucinated answers is elevated and harder to detect. The team recommends starting fresh chat sessions for each prompt, validating outputs against reliable sources, and ensuring human supervision and verification throughout the process.

The IEG has increased its data science and AI team and has organized training, publications, and conferences to help the evaluation teams understand how AI technologies can be leveraged to answer evaluation questions. According to IEG, the process is driven by evaluation questions rather than data alone, and the results are combined with other methodological approaches for triangulation.

Going forward, WB's IEG is developing a multi-year plan on how to responsibly integrate generative AI into its evaluation work, focusing on infrastructure, capacity building, and customized tools. They will continue experimenting and testing, while prioritizing data security and the purposeful use of AI to enhance rather than drive evaluations. WB has also developed a practical AI framework, which can be used to assess AI readiness across six key areas: (i) data and infrastructure; (ii) human capital and workforce; (iii) innovation and entrepreneurship; (iv) policy and regulation; (v) ethics and values; and (vi) future readiness. The framework is complemented by a simple guidance on the responsible use of generative AI, emphasizing ethical considerations, data governance, risk management, inclusivity, and capacity building.

3.2. Experience of evaluators

This study's data collection effort also included a survey with evaluation experts. The

39. DALL-E is an advanced AI model developed by OpenAI that generates images from textual descriptions.

purpose of the survey was to understand the extent and manner in which evaluators are using AI in evaluations. The survey involved evaluation staff/experts in the climate funds and international organizations and independent evaluators who are involved with these organizations. Using a purposive sampling approach that targeted circa 50 key evaluation experts, the survey received a total of 32 responses. It should be emphasized here that the selected sample is not representative of the evaluators' community, as the targeting focused on those evaluators likely to have tinkered with AI. However, as the survey showed (see Figure 1 below), even within this specifically tailored sample, only 32 per cent of participants reported to having made significant use of AI in their evaluative work. The box below provides a profile of the survey participants. More detailed information about the survey participants can be found in Annex II of this report.

Box 3: Profile of survey participants

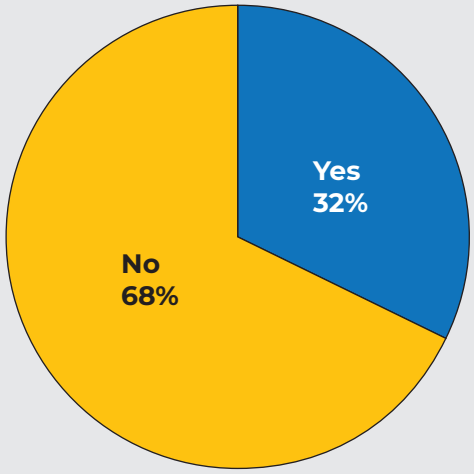
Among other things, the survey organized for this study collected information about the background and experience of participating professionals in the field of evaluation. The following is a brief summary, with more information provided in Annex II of this report.

- The majority of respondents (47 per cent) work as evaluators in NGOs or international organizations. About 38 per cent of them are program officers or managers.
- Most respondents have between 2-5 years (39 per cent) or more than 10 years (32 per cent) of experience with evaluations. They work primarily for international organizations (45 per cent), government agencies (26 per cent), and NGOs (10 per cent).
- The main areas in which survey respondents conduct evaluations are climate change (39 per cent) and environment (29 per cent). About 61 per cent of respondents have conducted evaluations of initiatives or projects funded by the climate funds (AF, CIF, GEF, or GCF).

The following is a summary of the results of the survey. The interpretation of results is presented in percentage terms, reporting the feedback of only those participants who responded to the survey questions (i.e., excluding those who did not respond to the question).

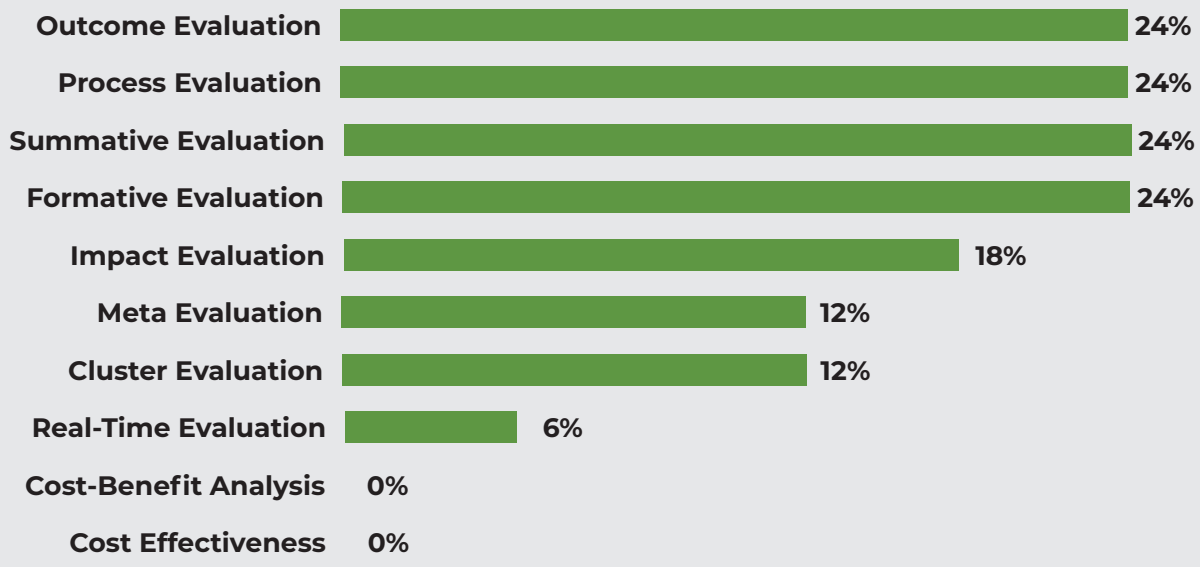
As can be seen in the figure below, the majority of survey respondents (approximately 68 per cent) reported to not have used AI in any form in relation to their evaluation work. This confirms one of the key points of the literature review – that while AI is starting to gain traction in the field of evaluation, it is still not widely adopted by practitioners.

FIGURE 1: Use of AI by survey respondents



Among those who have used AI in evaluation, its application has been distributed across various types of evaluations. The figure below shows that formative, summative, process, and outcome evaluations are the main types of evaluation where AI is used – with 24 per cent of responses each. Impact evaluations were identified by 18 per cent of respondents, whereas cluster evaluations and meta-evaluations have been used by 12 per cent each. The least extent of use of AI was reported for the task of real-time evaluation, which was selected by only 6 per cent of respondents.

FIGURE 2: Use of AI by type of evaluation



When used in evaluations, the survey shows that AI is most commonly applied to climate change adaptation projects, with 40 per cent of respondents selecting this area (figure below). Water resource management and agriculture and food security are also significant areas of focus, each selected by 28 per cent of respondents. Climate change mitigation is another key area, noted by 22 per cent of respondents. Fewer applications of AI are noted in the areas of disaster risk reduction and renewable energy, each with 11 per cent of respondents.

FIGURE 3: Use of AI in thematic areas

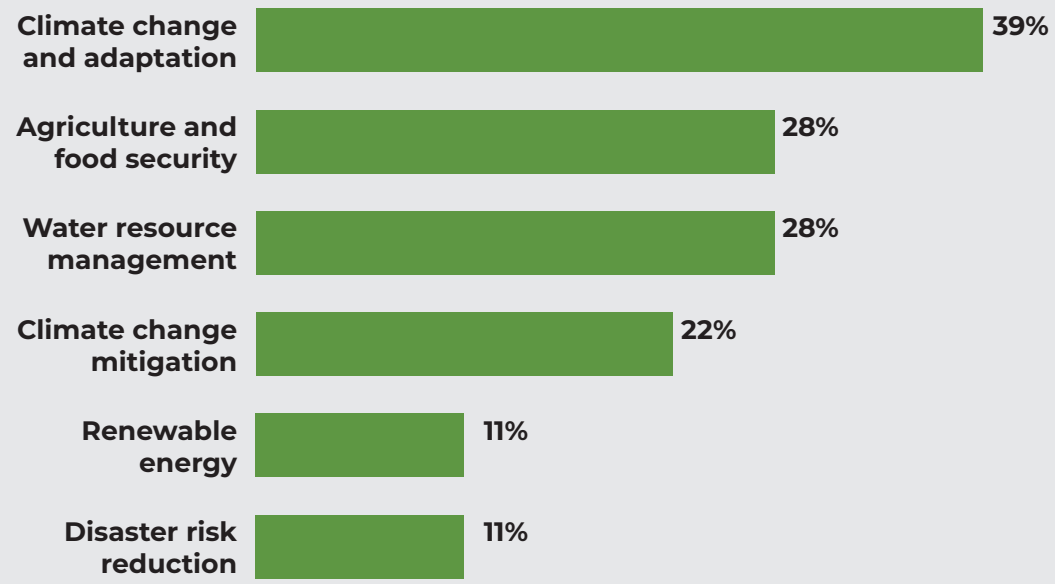


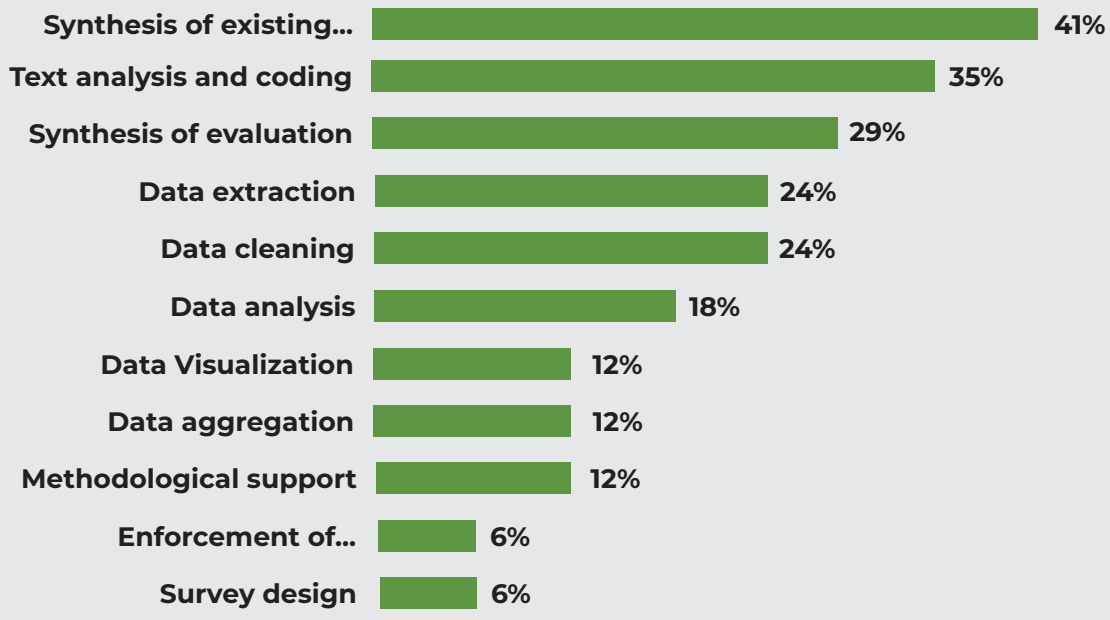
Figure 4 shows the evaluation phases in which survey respondents have used AI. The "Use and Dissemination" phase (which includes reporting) has the most use of AI with 67 per cent. The "Design" phase also sees substantial use of AI, with 50 per cent. The "Implementation" and "Quality Control" phases are equally represented, each with 42 per cent of respondents.

FIGURE 4: Use of AI in evaluation phases

A survey question explored the types of AI tools evaluators use in their work. The most common tool indicated by the respondents is by far ChatGPT. Other tools include Claude by Anthropic, Bard/Gemini by Google, and some custom-built applications that leverage APIs[40] from public AI models. Some respondents mentioned various advanced AI models such as Transformers, BERT, RoBERTa, or specific applications like SNA and XCBoost-Model,[41] which suggests that there is a more specialized use of AI among more technically-savvy evaluators. Additionally, some responses indicate that some internal/proprietary AI tools are also in use (e.g., the mAI platform developed by the WB Group).

The figure below shows the various uses of AI in evaluation. The most common applications are the synthesis of existing evidence and knowledge (41 per cent) and text analysis and coding (35 per cent). The synthesis of evaluation results is also a notable use case, reported by 29 per cent of respondents. Data cleaning and data extraction are two additional tasks that were reported by survey participants – each selected by 24 per cent of respondents. Data analysis was selected by 18 per cent of participants. Less frequently AI is used for methodological support, data aggregation, and data visualization, each selected by 12 per cent of respondents. A smaller proportion indicated the use of AI for survey design and enforcement of standards and quality control, with both at 6 per cent.

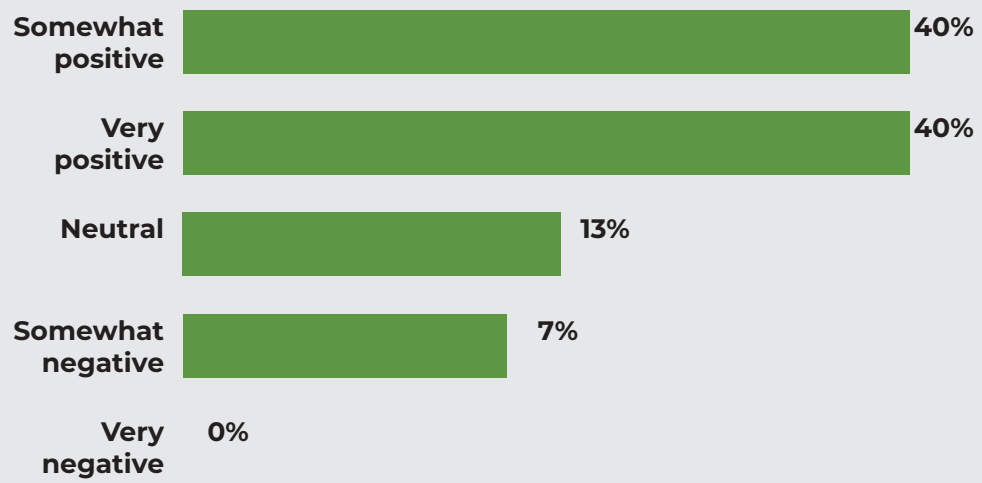
FIGURE 5: Use of AI in evaluation tasks



40. An API (Application Programming Interface) is a set of rules and protocols that allows different software applications to communicate with each other. It defines the methods and data formats that applications can use to request and exchange information, enabling them to interact in a standardized way.
Transformers are a type of neural network architecture that has revolutionized AI. BERT is a transformer-based model developed by Google that is designed to understand the context of words in a sentence by looking at both the words before and after a given word. RoBERTa is an optimized version of BERT developed by META. Social Network Analysis (SNA) is used to analyse social structures through networks and graph theory.

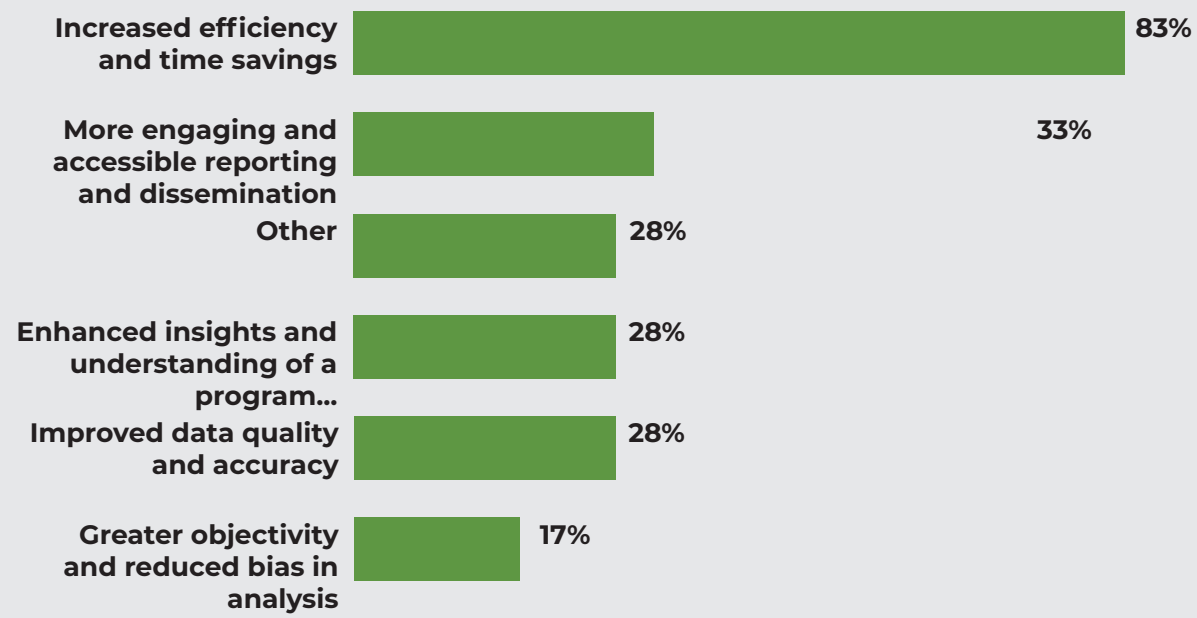
A majority of participants reported a positive outlook on the future of AI in the evaluation practice. As shown in the chart below, 40 per cent rated their experience as "very positive" and another 40 per cent as "somewhat positive". However, 13 per cent of respondents remain neutral, while a minority (7 per cent) reported a "somewhat negative" experience. Notably, no respondents rated their experience as "very negative".

FIGURE 6: Survey participants' experience with AI



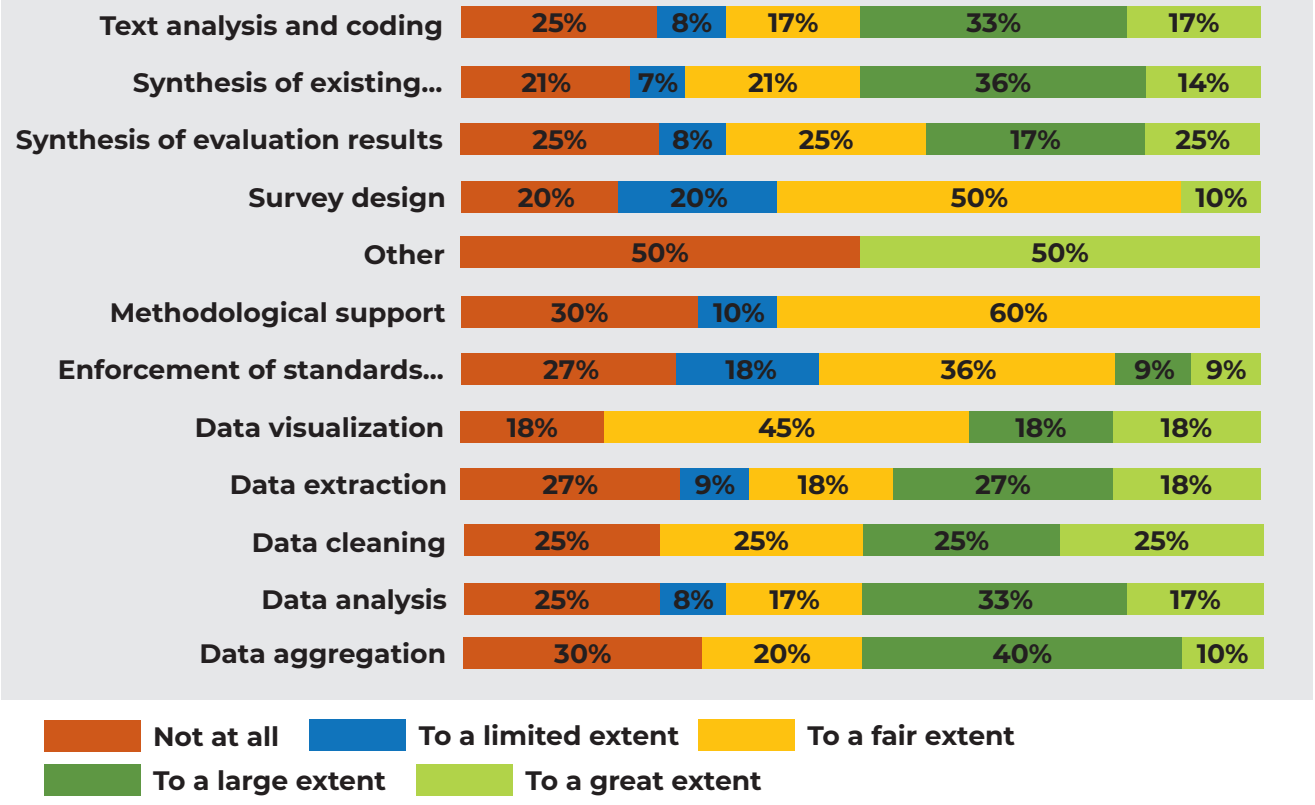
The figure below shows that 83 per cent of respondents identified the "increased efficiency and the saving of time" as the most significant benefit of AI. Other key benefits include "more engaging and accessible reporting and dissemination of results" (33 per cent), "improved data quality and accuracy" (28 per cent), and "enhanced insights and understanding of program outcomes and impacts" (28 per cent). A smaller proportion of respondents (17 per cent) identified greater objectivity and reduced bias in analysis as key benefits of AI. About 28 per cent of the respondents (under the category Others) identified other benefits, such as the ability to perform analyses which would otherwise be impossible due to cost and time constraints, or AI's potential to save time in data analysis. However, several respondents within the latter groups also expressed concerns about AI's limitations, particularly in understanding the local contexts and cultural nuances.

FIGURE 7: Benefits of AI



Survey respondents reported that AI tools have the potential to save time and effort in various evaluation functions. As can be seen from the figure below, this is particularly the case with regards to text analysis and coding, synthesis of existing evidence and knowledge, data cleaning, data extraction, data analysis, data aggregation, and synthesis of evaluation results. However, the impact of AI clearly varies across functions and respondents, with some experiencing significant time savings while others report limited or no benefits. While AI is perceived as a technology that offers time-saving advantages in several evaluation areas, the effectiveness varies, and in many cases, respondents remain neutral or only moderately satisfied with its impact compared to manual methods. This suggests that while AI has potential, its implementation and integration into evaluation processes may still require refinement and adaptation to specific needs.

FIGURE 8: Use of AI in evaluation tasks

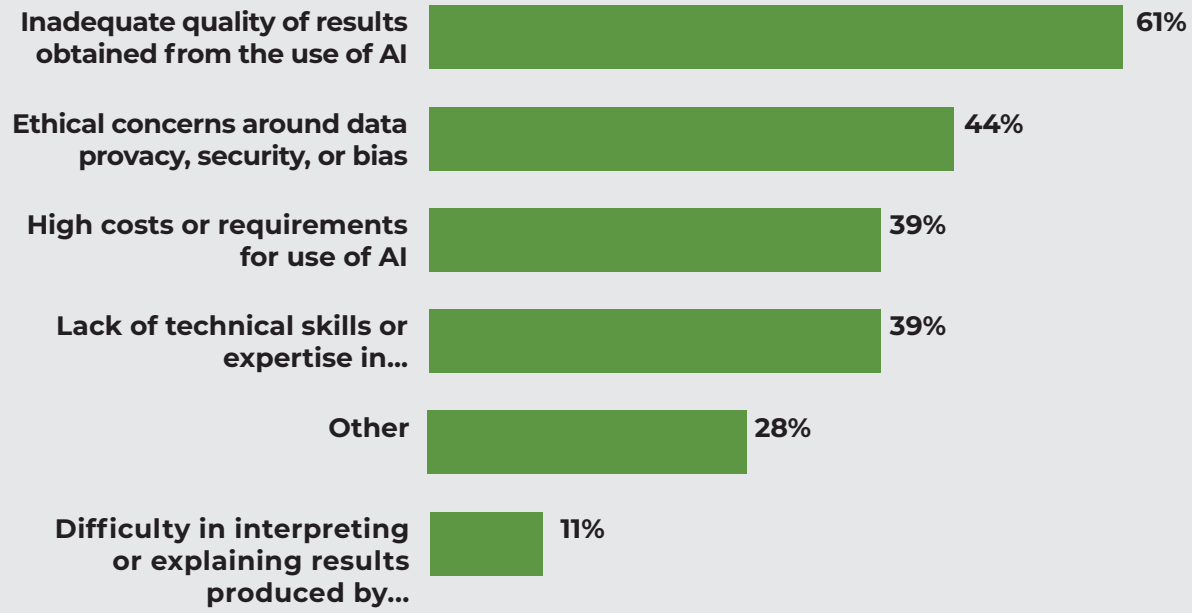


When asked to identify the primary advantages of using AI in the evaluation of climate projects compared to other areas, the respondents singled out the increased efficiency and time savings, enhanced data analysis capabilities, and improved clarity in reporting. AI is particularly valued by the respondents for its ability to process large, complex datasets, provide real-time data and predictive analytics, and help standardize evaluation processes. However, some respondents believe that these benefits are not unique to climate change evaluations and apply to other thematic areas. They also stressed that the effective use of AI depends on the evaluator's technical knowledge, with some respondents noting limitations if such expertise is lacking.

The figure below shows the main challenges reported by survey respondents when using AI in evaluations. They include the challenge of inadequate quality of results (61 per cent) and ethical concerns around data privacy, security, or bias (44 per cent). The survey thus reconfirms the main challenges identified in the literature review. Additional challenges include the lack of technical skills or expertise in understanding/using AI (39 per cent), the high costs or requirements for use of AI (39 per cent), and difficulties in interpreting or explaining the results produced by AI (11 per cent). Some respondents (category "Other" with 28 per cent) mentioned other challenges such as prejudice from colleagues and the need for extensive validation and workarounds

due to AI's propensity to commit errors or hallucinations. These challenges indicate a need for improved AI tools, better training, and greater awareness on the ethical implications of the use of AI in evaluations.

FIGURE 9: Challenges of AI

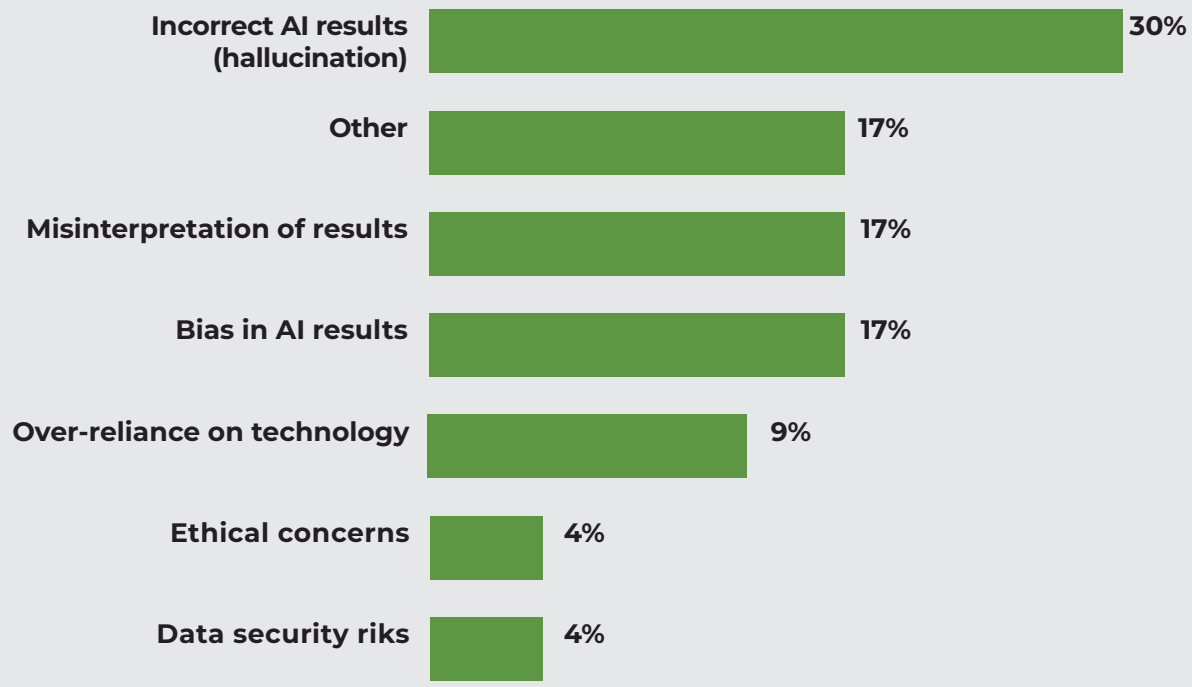


The survey also highlighted some key opportunities, as perceived by the participating practitioners. Several respondents noted that AI can improve the efficiency of the evaluation process significantly by automating certain routine tasks, like data cleaning, data collation, and preliminary analysis. Liberated from routine tasks, the evaluators will be able to focus on more complex and critical aspects of the evaluation process, where the premium on human capabilities is higher. The ability of AI to analyse large and complex datasets – especially, data from satellite images and climate models – can help practitioners make evaluations more accurate and comprehensive. Survey respondents also indicated that AI has significant potential in the area of predictive analytics. This a function that is crucial for forecasting climate trends and informing adaptation strategies. AI's potential is also seen in helping with decision-making by allowing policymakers identify patterns and insights that may not be immediately apparent. This in turn leads to better-targeted interventions. However, several respondents also expressed caution. They noted several challenges, in line with those identified in the literature review, and the need for human validation of the insights generated with AI. A small minority of survey participants said they see

more challenges and issues rather than opportunities in the use of AI, which suggests that there may be widely varying levels of perception of AI among evaluators.

The survey also assessed the sense of perceived risks and concerns among evaluation practitioners. As can be seen from the figure below, the most commonly cited concern is the potential for incorrect results or hallucinations, with 30 per cent of respondents highlighting this problem. Concerns about bias in the results of AI and the misinterpretation of those results were each noted by 17 per cent of respondents, suggesting concerns about the accuracy and fairness of AI-produced results. Over-reliance on technology was a concern for 9 per cent of respondents, while data security risks and ethical concerns were each mentioned by 4 per cent of the respondents. Additionally, some respondents (under the "Other" category with 17 per cent) indicated several other concerns, such as the need for human oversight and the challenge of integrating AI into evaluation processes. Again, these insights confirm the findings of the literature review. They suggest that while AI offers significant potential benefits, there are also substantial perceived risks that need to be carefully monitored and managed.

FIGURE 10: AI Risks

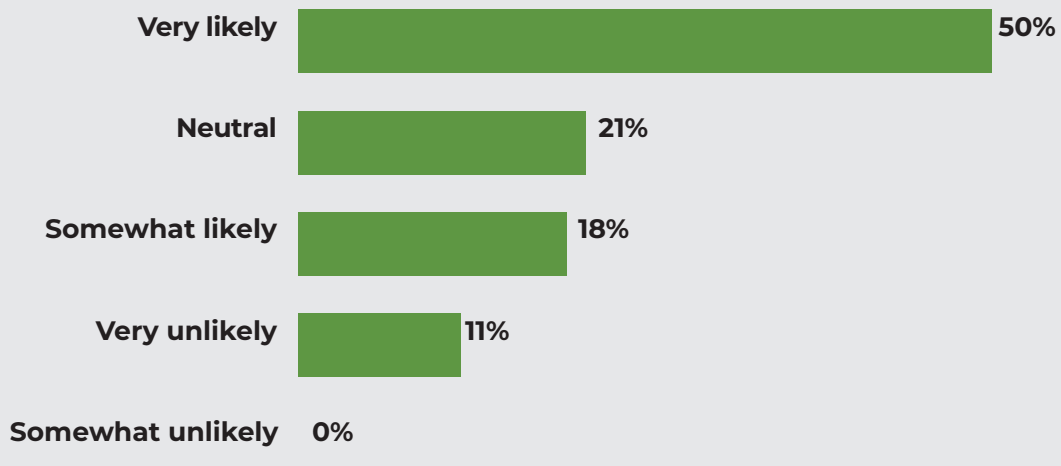


The survey also asked practitioners about potential risk mitigation strategies or approaches that could be deployed to address the risks of AI. At the policy level, survey respondents proposed the development of AI risk management frameworks that ensure the transparency and explainability of AI processes. They also suggested the conduct of regular audits and the establishment of strong data governance policies. Another key suggestion was the organization of training on AI and digital literacy for evaluators and staff. Another key mitigation factor stressed by the respondents is oversight and human supervision. They are seen as crucial for ensuring that AI results are triangulated and validated by the human factor with traditional research methods.

Overall, a general insight from the survey on the aspect of mitigation is that evaluators need to develop systematic approaches for validating AI results and should maintain a critical eye on the outputs generated by AI systems. Additionally, fostering interdisciplinary collaboration and making AI tools more open and collaborative will be crucial for improving the quality of context-specific assessments. Finally, participants stressed the importance of continuous learning and adaptation of AI systems, an insight largely in line with the main findings of the literature review.

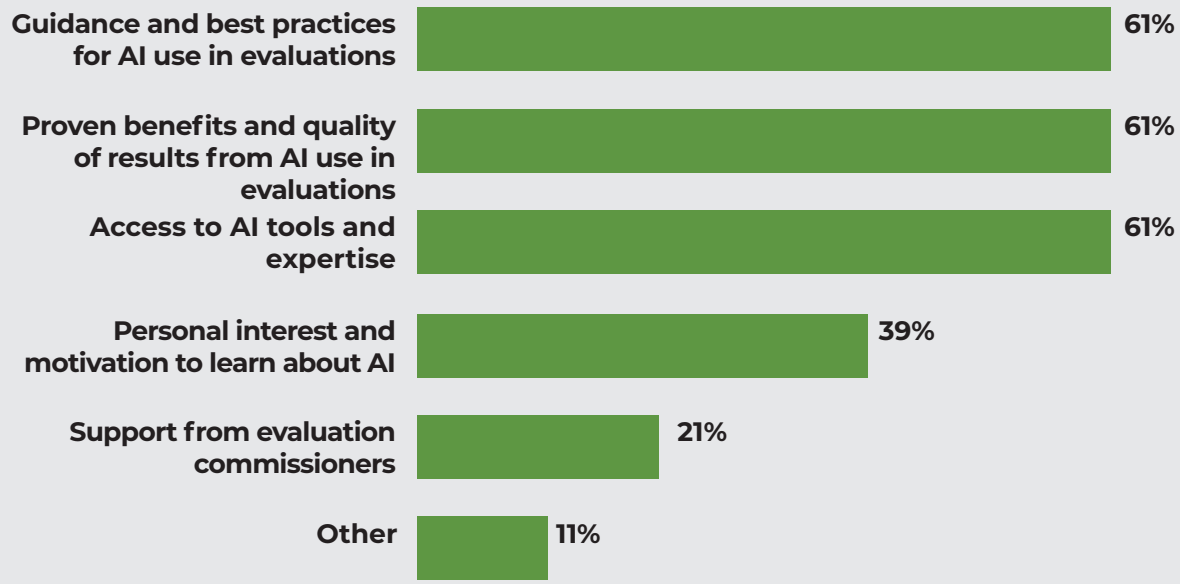
The survey indicates that a majority of respondents are inclined to use AI in future evaluations, with 50 per cent stating they are “very likely” and 18 per cent “somewhat likely” to do so. A notable portion, 21 per cent, remains neutral. A smaller group, 11 per cent, are “very unlikely” to incorporate AI into their future evaluations. These results suggest a generally positive outlook toward the future use of AI in evaluations, though some reservations remain among a minority of respondents.

FIGURE 11: Likelihood of using AI in the future



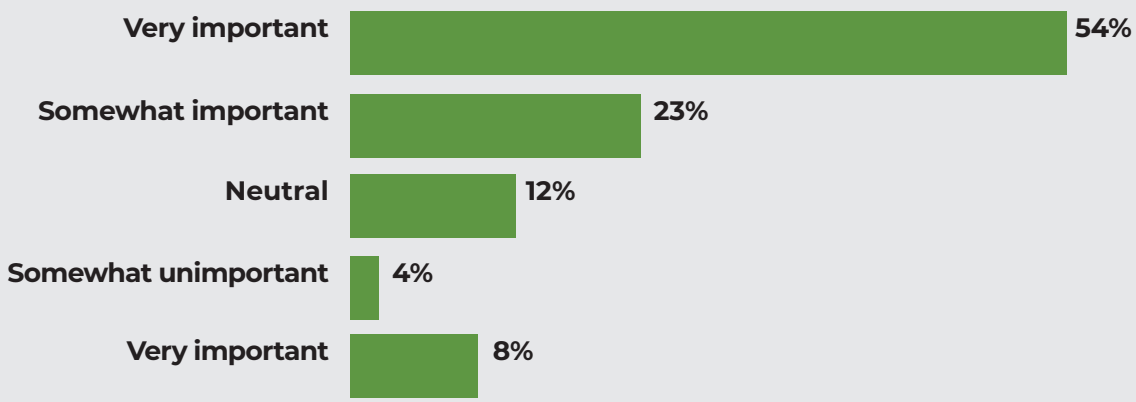
The survey also asked participants about the factors that are most likely to influence their decision on whether to use AI tools in future evaluations. The results are shown in the figure below. The following are the three top factors: access to AI tools and expertise; proven benefits and quality of results from AI use; and guidance and best practices for AI implementation (each identified by 61 per cent of the respondents). Further down the list of factors comes personal interest and motivation to learn about AI, identified by 39 per cent of respondents. Support from evaluation commissioners is a less common, but still relevant, factor, reported by 21 per cent of respondents. Some participants (under the "Other" category with 11 per cent of respondents) noted the importance of AI's ability to understand and incorporate country-specific contexts, cultural nuances, and the subtleties of personal stories and local values into its algorithms as a factor for future use of AI tools. These findings underscore the need for practical resources, especially guidance, proven effectiveness through best practices and use cases, and contextual sensitivity.

FIGURE 12: Factors likely to influence use of AI in the future



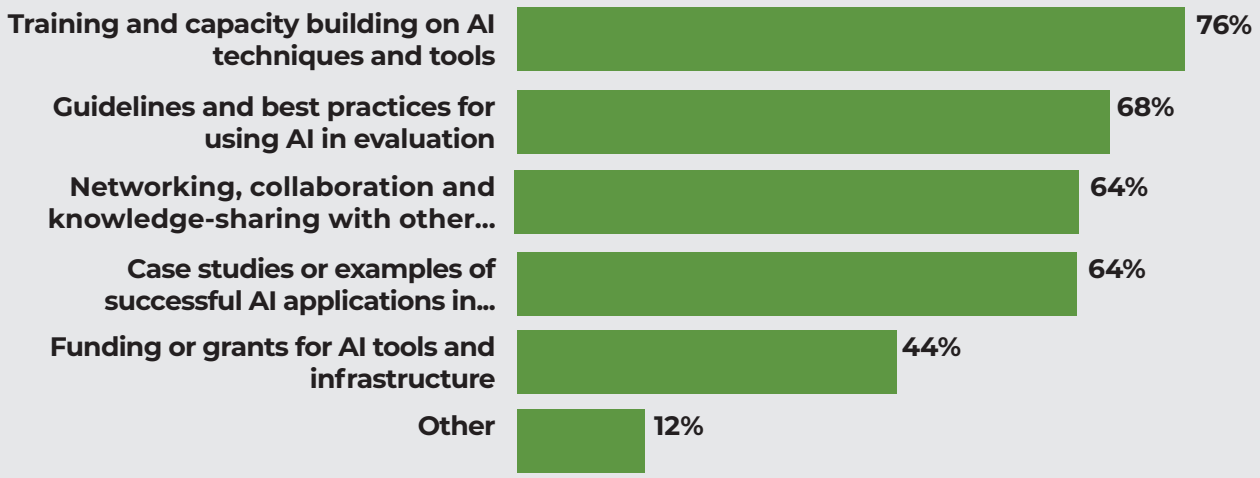
The survey also reveals that a majority of respondents consider it important for evaluators to develop skills in the use of AI. Specifically, 54 per cent believe that this matter is "very important," and 23 per cent think it is "somewhat important". A smaller group of about 12 per cent remain neutral on this issue. On the other hand, 4 per cent of respondents view it as "somewhat unimportant", whereas 8 per cent consider it "very unimportant". These results suggest the need for capacity development among evaluators. This might also have an impact on the minority, which remains skeptical or sees less value in developing these skills.

FIGURE 13: Importance of developing AI skills



The survey also shows that the most useful types of support expected by evaluators include training and capacity building on AI techniques and tools (which is selected by 76 per cent of respondents). About 68 per cent of respondents identified guidelines and best practices for AI use as important resources. Case studies or examples of successful AI applications and opportunities for networking, collaboration, and knowledge-sharing were each noted by 64 per cent of respondents. About 44 per cent of respondents identified funding or grants for AI tools and infrastructure as important for skill development. A small portion of respondents (about 12 per cent) emphasized the need for assurance that concerns about AI, such as accuracy and ethical considerations, are addressed, and that evaluators are equipped to integrate these considerations into AI tools effectively. These findings indicate a strong demand for practical resources, training, and collaborative opportunities.

FIGURE 14: Type of support needed



Survey participants identified some factors which, in their opinion, are likely to influence the adoption of AI in evaluations in the coming years. The main ones include stronger proof of the benefits of AI, access to training and education on AI technologies, and the development of standards and best practices for the use of AI. Respondents also highlighted the need to address ethical concerns, particularly those related to data privacy. They also pointed out the need to ensure that AI systems are capable of generating accurate results before using them more widely. The availability of time and financial resources to institutionalize AI, advancements in AI technology, and the necessity of high-quality data were also noted as critical factors. Additional factors mentioned by the respondents were the importance of stakeholder acceptance, the need for continuous testing and innovation, and the need for support from organizations with training and resources for the development of AI capacities.

The survey included a specific question on those AI applications or techniques that the respondents thought have the greatest potential for improving evaluations in the climate sector. The respondents identified some specific AI methods that could be effective in climate evaluations, e.g., Random Forests for predictive modelling and climate simulations, Convolutional Neural Networks for analysing satellite imagery, and Bayesian Networks for uncertainty modelling and probabilistic climate risk assessment. Additionally, the participants singled out machine learning algorithms for their ability to analyse vast datasets and identify patterns. They also noted that NLP tools could process and analyse textual data from reports and social media to gauge public sentiment and gather qualitative data. Remote sensing technologies powered by AI were also noted for their ability to monitor environmental changes in real-time, offering precise data on key metrics such as deforestation and land use changes. Tools like ChatGPT were recognized for their ease of use, particularly for simple queries. Some respondents also expressed uncertainty or a lack of familiarity with specific AI applications, which further reinforces the idea that further education, training, and knowledge sharing are critically needed within organizations and the evaluation community.

IV. Conclusions and way forward

The literature review, interviews and survey conducted in the framework of this scoping study yielded a wealth of information. The following are the key take-aways from this research organized in two sections – the current state of the use of AI in evaluations (specially climate evaluations) and the way forward.

Current situation

Current Use of AI in evaluations:

- AI is being increasingly used in evaluations. Yet, adoption remains low and has been slower than in other fields such as medical research, business analysis, software coding, etc.

- The potential benefits of AI are not evenly spread. AI holds particular promise in the analysis of qualitative data – such as summarizing documents, extracting themes, generating ideas, producing first drafts, and coding.

- AI can be helpful for external validity (detecting patterns across large datasets, which can help determine if findings can be extended to other situations). However, it struggles with internal validity (capturing causal relationships) because of the lack of contextual understanding.

- AI may also be used to classify and predict socio-economic outcomes from large data (e.g., satellite imagery, mobile phone data, and behavioural indicators). This enables more cost-effective and scalable data collection, especially valuable in resource-constrained environments.

- For all the potential benefits, human judgement is crucial for contextual interpretation and will remain indispensable in the evaluation practice. Also, the modelling of causality will remain crucial for valid evaluation results.

- In climate evaluations, the use of AI has been very limited, despite its potential. Most of the existing use of AI has focused on broader climate research and modelling.

- In climate research AI is used in data collection and analysis (satellite imagery analysis, data quality enhancement, automation), climate modelling (increased resolution, data assimilation, parameterization), and prediction accuracy

(pattern recognition, model calibration, dynamic ensembles).

- Prompts are crucial for the effective use of AI; they drive the quality of results. Evaluators must craft well-structured and clear prompts that align with evaluation frameworks and guide AI to behave like an evaluator.

- Most international organizations are in early stages of AI adoption, experimenting with various tools for specific evaluation tasks. Common applications include document synthesis, text analysis, intervention targeting, qualitative data analysis, and evidence extraction. The organizations and evaluators using AI should clearly and transparently acknowledge and note the use of AI in their work.

- International organizations are divided in two camps when it comes to the issue of internal versus external AI systems and tools. In the face of AI risks, some organizations are exploring the development of secure in-house AI models. Others think that the costs of developing these models, and the level of success provided by commercial models, do not justify investments in internal systems. There is also a middle path, which involves the use of available models, but with the data sitting in-house.

Opportunities:

- AI has the potential to enhance significantly the efficiency and quality of evaluations. This potential seems to be well recognized within the evaluation community. If used responsibly, AI can increase productivity, responsiveness to citizen needs, and accountability in the public sector.

- Tailored AI applications built on pre-trained language models, using specialized prompt engineering and fine-tuning, are expected to have the greatest impact in evaluation.

- AI offers significant advantages in the automation of repetitive tasks, processing of larger evidence bases, and the conduct of more comprehensive analyses. It has the potential to displace many human tasks in the evaluation profession (e.g., transcription, translation, document screening, thematic analysis, sentiment analysis, coding), while augmenting others (e.g., research design, establishing criteria).

- The main benefits of AI include reduced costs/time, larger samples, granular analysis, and evaluating complex programs. About 80 per cent of surveyed evaluators and experts reported a positive experience with AI, citing increased

efficiency and time savings as the most significant benefit. However, theory-driven evaluations will remain important.

- Complexity is a critical aspect of large-scale programs, especially in climate change. In this area, AI offers considerable advantages – it can help collect and analyse the dynamic and interacting data needed to understand complex systems. Nevertheless, few evaluators are currently using these techniques in the climate area.

- Building AI systems is a resource-intensive enterprise. It also requires extensive data resources and data systems. Organizations with small budgets and resources have the opportunity to leverage existing large language models (e.g., Claude, GPT) open to the public, rather than building their own from scratch. Larger organizations will need to balance their concerns about information privacy and the benefits of rapidly evolving commercial systems in the decision-making about the feasibility of internal systems.

Challenges and risks:

- The results of AI depend on the quality and representativeness of the data used for its training. AI tools may hallucinate false information and reflect biases present in Internet data they were trained on.

- AI poses interpretability challenges. It is not fully reliable in the interpretation of facts and internal validity (causality) checking. Also, the replicability of results is a challenge. Even with the same prompt and same tool, generating the same result is unlikely. This inherently raises concerns about reliability and scalability.

- Also, given that AI models are trained primarily on data from larger, wealthier countries, they may not adequately capture the unique challenges and solutions relevant to smaller or poorer nations.

- Another key AI risk is misalignment with human values. AI may perpetuate existing and dominant forms of bias, racism and prejudice present in the large-scale Internet data.

- There is a risk that in certain situations the efficiency gained by using AI gets lost or diluted due to extensive validation and workarounds. Therefore, a cost-benefit assessment on a case-by-case basis is necessary – at least in the current condition of AI.

- For most international organizations the limited availability of well-structured

and machine-readable evaluation data limits the potential benefits of AI use. Efforts are needed to standardize data formats and develop data governance frameworks.

- Data security and privacy are major concerns, particularly with using external AI platforms for sensitive information. International organizations, including the climate funds, often deal with sensitive data about project proposals, financial information, and stakeholder details. Using external AI platforms could potentially expose this information to unauthorized access or breaches.

- The growing computational infrastructure for AI has environmental impacts, primarily through greenhouse gas emissions and water usage.

- There is also a challenge with the distribution of the benefits of AI. First, some of these tools are not available in all countries or regions. Further, nations with small populations are inherently disfavoured in the AI game by virtue of the small written base of their language. There are also requirements for investments in computational infrastructure which are not available to many small and poor countries. Given that the effectiveness of AI rises exponentially with the volume of available data, language-driven disparities in AI benefits will represent an ethical issue for the international community. The unequal access to AI tools across countries and regions could exacerbate existing technological divides. Also, the concentration of AI expertise in certain regions may exacerbate the brain drain from smaller or poorer countries, further hampering their ability to develop locally relevant AI solutions for just transitions.

Mitigating strategies:

- The evaluation community should engage with AI critically, assess it for bias and validity of results, and ensure that all stakeholders and affected communities have a voice in its design and accountability.

- Evaluators and organizations should engage with AI through a continuous process of verification, improvement, and incorporation of best practices.
- Validation is key. Rigorous validation of AI outputs against reliable sources and expert judgement is essential. Evaluators should employ techniques like spot checks, expert review, and cross-referencing to assess the accuracy and completeness of AI results.

- Evaluators should thoroughly assess the evaluation context and needs before choosing an AI tool.

- Collaboration between AI and human judgement is crucial. Key to responsible use of AI will be combining AI's pattern recognition and analytical power with human skepticism, contextual understanding, and critical perspective. A human-in-the-loop process, involving subject matter experts and data scientists, will be essential for validating AI results, fine-tuning AI models, and ensuring AI's alignment with the evaluation framework.

- The rapid pace of change in AI technologies makes it difficult for research and testing to keep pace. Flexible and iterative approaches to AI testing and implementation are needed. Ongoing testing, research, and experimentation are needed to understand and realize AI's benefits, while at the same time identifying and mitigating risks.

- Organizations need to develop ethical frameworks and guidelines and protocols for responsible AI use in evaluation, including human oversight and validation.

- Capacity building is critical. There is a need for comprehensive capacity development (training and education) for evaluators and organizations on the use and limitations of AI. Organizations need to develop AI literacy and hands-on experience for their staff. At the same time, their IT/AI experts should understand evaluation needs and contexts. Collaboration between the two is essential.

- To address data privacy concerns, organizations can access AI models behind a firewall, ensuring that sensitive data never leaves their secure environment. Interested organizations can explore the experience of the World Bank with the securing of its private data through an agreement with Microsoft.

- There is also an opportunity to advocate for organizations that own AI models to train them using diverse, representative data to reduce bias and the risk of exclusion.

Way forward

- AI is expected to get better and at an accelerating pace. This presents incredible opportunities for all research fields, the evaluation field included.
- Evaluators and organizations will have to inevitably embrace the use of AI in the evaluation process. However, this should be done in a gradual, critical, and responsible manner.

- The adoption of AI in evaluations should be seen as a journey, not a destination.

Organizations should start with simple applications of AI and gradually progress to more complex tasks, continuously building on their experiences and learning from others.

- Customization is important. Off-the-shelf AI tools may not fully meet evaluation needs. Organizations will need to develop tailored applications or fine-tune existing models to their specific requirements.

- Managing expectations is crucial. AI is not a panacea; it requires human oversight, validation, and collaboration to ensure accurate and contextually relevant results.

- It is also important to strike a balance between AI-driven analysis and human creativity in the generation of new knowledge and innovative solutions.

- AI applications have to be carefully tailored to specific evaluation contexts, questions, and data landscapes. This requires close collaboration between evaluators, data scientists, and stakeholders to design AI solutions that are fit-for-purpose and aligned with organizational needs and values.

- Organizations should foster a culture of experimentation and innovation in the use of AI in evaluations. This should include the creation of spaces for piloting new approaches, sharing lessons learned, and constantly improving AI models and processes based on feedback and validation. It is crucial to recognize that waiting for AI solutions to fully mature before implementation can significantly hinder progress and learning opportunities. To mitigate potential downsides, organizations should implement risk management strategies, start with small-scale pilots, and maintain transparency about the experimental nature of these initiatives.

- The responsible use of AI requires collaboration, peer learning, and knowledge sharing, both within and across organizations. Organizations should establish platforms and dialogue for exchanging best practices, case studies, and lessons learned.

- Organizations will need to invest in AI skills and capacity. This includes training on AI techniques, tools, and ethical considerations, as well as the creation of a culture of continuous learning and adaptation.

- Organizations will need to address ethical concerns around data privacy, security, and bias. They should establish robust data governance frameworks

and ensure that the use of AI adheres to ethical principles.

- Organizations have to build a secure physical data architecture that ensures data privacy and confidentiality, particularly when dealing with sensitive information.

- Resource constraints will shape the adoption of AI in many organizations. Those with limited budgets and staff will need to be more strategic in their AI investments, focusing on high-impact, external, and portable solutions.

- Evaluators and organizations will need to continuously assess the impact and value-added of AI by closely monitoring performance, efficiency gains, and unintended consequences.

- It will be essential for all organizations to develop guidelines, frameworks, and best practices for the responsible use of AI in evaluations. These should cover aspects such as data governance, bias mitigation, transparency, and risk management, and be regularly updated in line with updates of AI capabilities.

The final chapter discusses recommendations for the four climate funds to align their strategies and learning with the rapid evolution of AI.

V. Key recommendations for the four funds

The scoping study has laid the ground for further systematic and joint work by the four climate funds (Adaptation Fund, Climate Investment Funds, Global Environment Facility, and Green Climate Fund) on the use of AI in the evaluation process. It provides the starting point for several pathways that the four climate funds could follow individually and jointly, depending on the political realities of cooperation.

The following are a set of practical and actionable recommendations on how the four funds could responsibly and gradually integrate AI into their evaluation practice. The recommendations are organized in two categories: short-term and long-term recommendations.

Short-term recommendations

1. Maintain and further institutionalize the AI Joint Working Group established for the scoping study:

- Maintain the Joint Working Group with representatives from the four funds that oversaw the conduct of the scoping study and further institutionalize it by formalizing some aspects of its operations, such as regular meetings, a ToR for the functioning of the group, roles, and responsibilities within the group, etc.
- Task this group with overseeing initial AI testing and preparatory work.

2. Initiate small-scale AI testing pilots:

- Select 2 to 3 low-risk, high-potential use cases (e.g., synthesis of evidence, thematic analysis, tailored reporting).
- Identify publicly available AI tools (e.g., ChatGPT, Claude) for initial testing.
- Set clear objectives and tracking metrics for each pilot.
- Document lessons learned and best practices.

3. Establish a basic validation protocol:

- Develop a simple framework for validating AI-generated outputs.
- Identify steps for human review and cross-checking with traditional methods.
- Test the protocol during the testing pilots.

4. Develop basic AI guidelines:

- Create initial guidelines (preferably joint) for the responsible use of AI in evaluations.

- Focus on key areas: data privacy, transparency, human oversight, and just transition.

5. Establish an AI experience log:

- Set up a shared document or platform to record experiences with AI.
- Encourage all pilot participants to document insights, challenges, and lessons learned.
- Use this log to inform future AI approach/strategy development.

6. Organize AI awareness (light training) sessions:

- Organize introductory workshops on AI for evaluation staff, but also other departments in the organizations.
- Cover basic concepts, potential applications, and ethical considerations.

7. Develop an approach (or even better a strategy) for the systematic use of AI in evaluation:

- Develop a 1- to 2-year roadmap (joint or individual, depending on political realities of cooperation) for the gradual integration of AI in evaluation processes.
- Define clear objectives, milestones, and success metrics for the integration process.
- Align the strategy with each fund's broader digital transformation objectives.

Long-term recommendations

8. Establish a joint AI governance framework:

- Establish a joint AI steering committee with representatives from evaluation, IT, and ethics departments, which has more strategic powers and oversight responsibilities than the preliminary AI working group.
- Maintain the current Joint Working Group across the four funds as a platform to share AI experiences and best practices and to inform the work and decision-making of the steering committee.
- Develop joint guidelines for the ethical use of AI, addressing data privacy, bias mitigation, and transparency.

9. Establish a validation and risk assessment framework:

- Building on the basic validation protocol (above), further develop the protocols for validating AI-generated insights and results.
- Establish a risk assessment protocol for the use of AI in evaluations.
- Develop guidelines for human oversight and intervention in AI-assisted

evaluations.
- Create a system for documenting and learning from AI errors or biases.

10. Invest in AI capacity building:
- Organize basic and advanced training content for relevant staff within each fund.
- Provide AI literacy training to all staff (within and outside the evaluation units).
- Engage to the extent possible with AI/data science experts from leading AI companies.

11. Strengthen data infrastructure:
- Assess current data management practices and identify areas for improvement.
- Implement standardized data collection and storage protocols to ensure AI-readiness.
- Explore secure cloud-based solutions for data storage and AI model deployment.

12. Develop AI-assisted evaluation toolkits:
- Identify a suite of AI tools suitable/tailored for climate-related evaluations (e.g., satellite image analysis, climate data processing).
- Adapt AI tools for processing and analysing complex climate datasets.
- Ensure these tools are user-friendly and accessible to evaluators with varying levels of technical expertise.
- Regularly update these toolkits based on user feedback and technological advancements.

13. Promote collaboration and knowledge sharing:
- Continue to collaborate and share information with other international organizations on best practices in the use of AI in evaluations.
- Organize annual workshops or conferences on AI in climate evaluations.
- Develop a shared repository of AI use cases and lessons learned (the AI experience log recommended mentioned above could be converted into the shared repository).

14. Engage with external stakeholders:
- Conduct regular consultations with other organizations, evaluation teams, project implementers, and beneficiaries on AI use.
- Ensure transparency about AI use in evaluation reports and communications.
- Address concerns and incorporate feedback into AI strategies.

15. Develop AI-specific evaluation criteria:

- Create a framework for evaluating the effectiveness and impact of AI in evaluation processes.
- Regularly assess the cost-benefit ratio of AI implementations
- Monitor the unintended consequences of the use of AI.

16. Establish partnerships:

- Collaborate with academic institutions, research institutes, and tech companies for cutting-edge AI research and development.
- Partner with other international organizations to share resources and learnings on AI in evaluation.

17. Prioritize environmental considerations:

- Assess the environmental impact of AI implementations (e.g., energy consumption of data centres).
- Explore and prioritize energy-efficient AI solutions.
- Offset the carbon footprint of AI operations as part of the funds' climate commitments.

18. Develop an AI communication strategy:

- Craft clear messaging about the funds' approach to AI in evaluations.
- Develop clear guidelines for communicating the use of AI in evaluation reports.
- Prepare FAQs and talking points for addressing stakeholder concerns about AI.

19. Establish an AI ethics review process:

- Form an ethics review board for the use of AI.
- Develop a checklist for ethical considerations in AI-assisted evaluations.
- Audit on a regular basis AI systems for potential biases or ethical issues.

20. Establish a feedback loop:

- Implement a system for evaluators to provide feedback on AI tools and processes.
- Regularly survey stakeholders on their perceptions and experiences with AI in evaluations.

It is advisable to use this feedback as a continuous learning and adaptation loop to refine and enhance your AI strategies.

Annex I: Interview participants

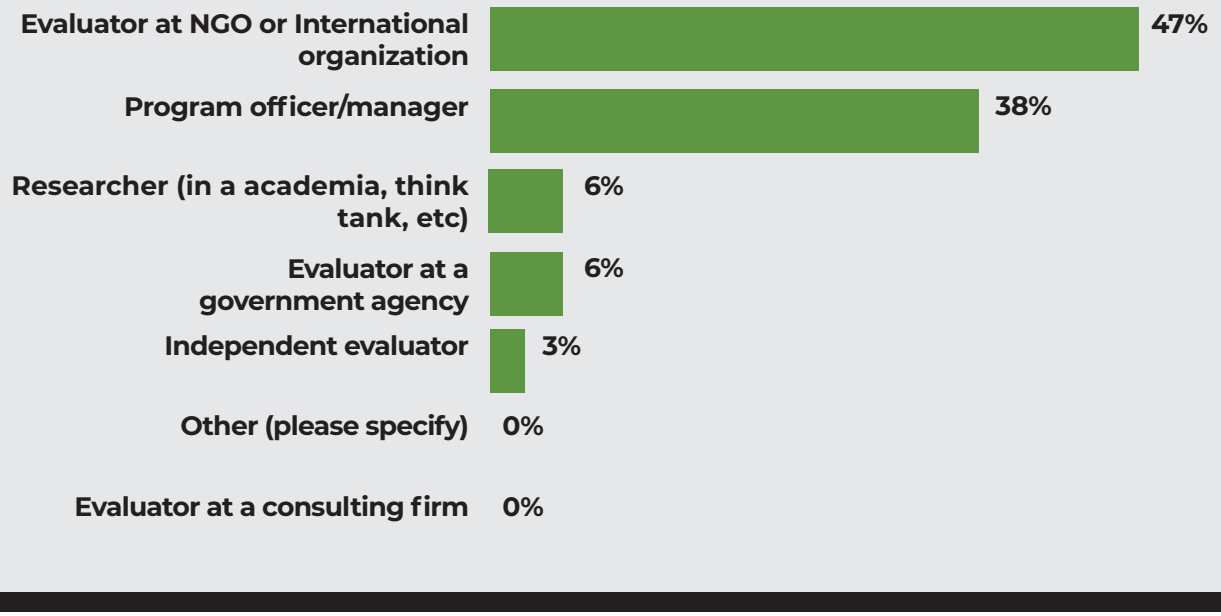
1. Anish Pradhan, UNDP Independent Evaluation Office
2. Anoop Sharma, IFAD Evaluation Office
3. Dr. Anupam Anand, GEF Independent Evaluation Office
4. Brittney Dana Melloy, Climate Investment Funds
5. Devi Prasad Ayyannamahanty, Climate Investment Funds
6. Dr. Bianca Montrosse-Moorhead, University of Connecticut
7. Dr. Nanthikesan Suppiramaniam, IFAD Evaluation Office
8. Gagneet Kaur, IFAD Evaluation Office
9. Gonzalo Gomez Melazzini, UNDP Independent Evaluation Office
10. Hannah den Boer, IFAD Evaluation Office
11. Harsh Anuj, World Bank Independent Evaluation Group
12. Kai Rompczyk, Deval (private company)
13. Michael Bamberger, 3ie (non-profit)
14. Michael Ward, Climate Investment Funds
15. Neha Karkara, UNFPA Evaluation Office (and co-convener of the United Nations Evaluation Group/working group on data and AI)
16. Srilata Rao, Chief of Section, Inspection and Evaluation Division, Office of Internal Oversight Services, United Nations
17. Sven Harten, Deval (research firm)
18. Uyen Kim Huynh, UNICEF
19. Virginia Ziulu, World Bank Independent Evaluation Group

1. Footnote xxxx

Annex II: Survey participants

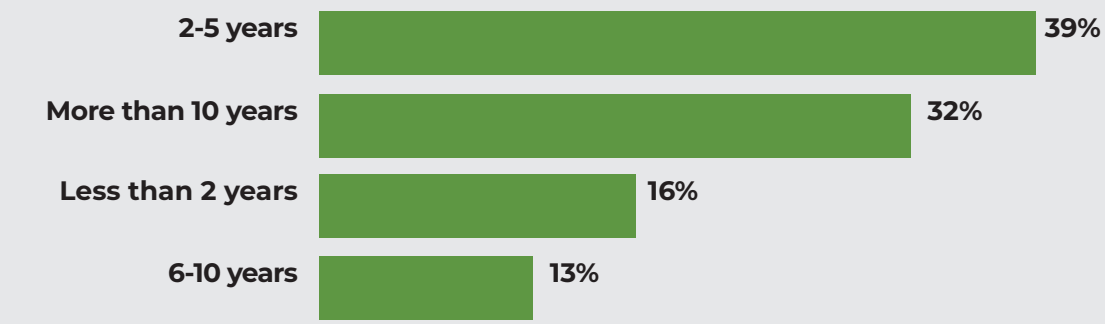
This annex provides a brief description of the profile and background of the participants of the survey that was organized for this study. In total 32 evaluators and experts responded to the survey. The majority of survey respondents reporting to being evaluators at NGOs or international organizations (47 per cent), followed by program officers or managers (38 per cent). There was minimal representation of independent evaluators (3 per cent), government agency evaluators (6 per cent), and researchers (6 per cent), and no respondents from consulting firms.

FIGURE 15: Categories of survey respondents



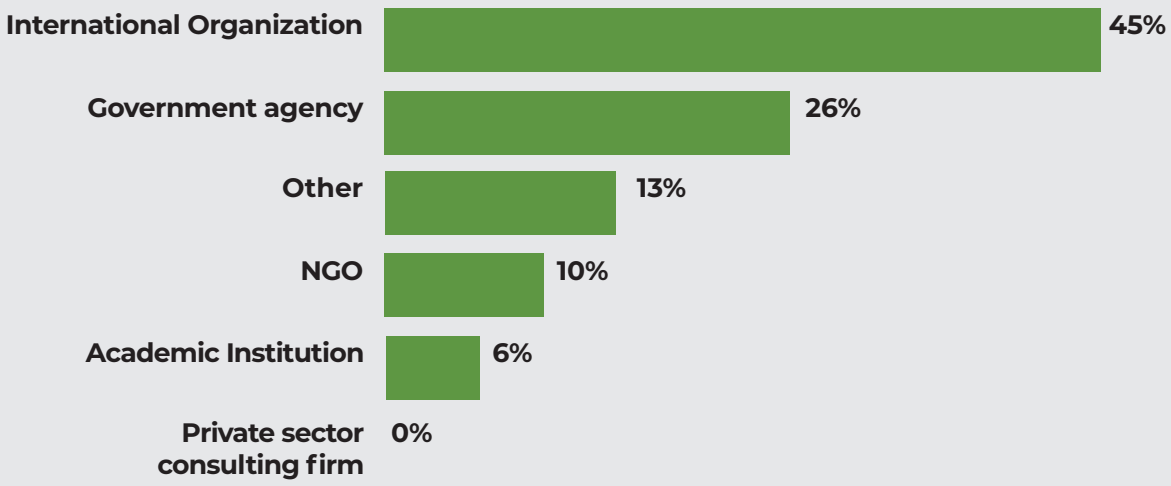
The majority of respondents had substantial experience with evaluations. Specifically, 39 per cent had two to five years of experience, while 32 per cent had more than 10 years of experience. This suggests a fairly experienced cohort overall, with a significant portion (over 70 per cent) having more than two years in the field. A smaller group of respondents, 16 per cent, were relatively new to program evaluation with less than two years of experience, and 13 per cent had between six to 10 years of experience.

FIGURE 16: Years of experience of survey respondents



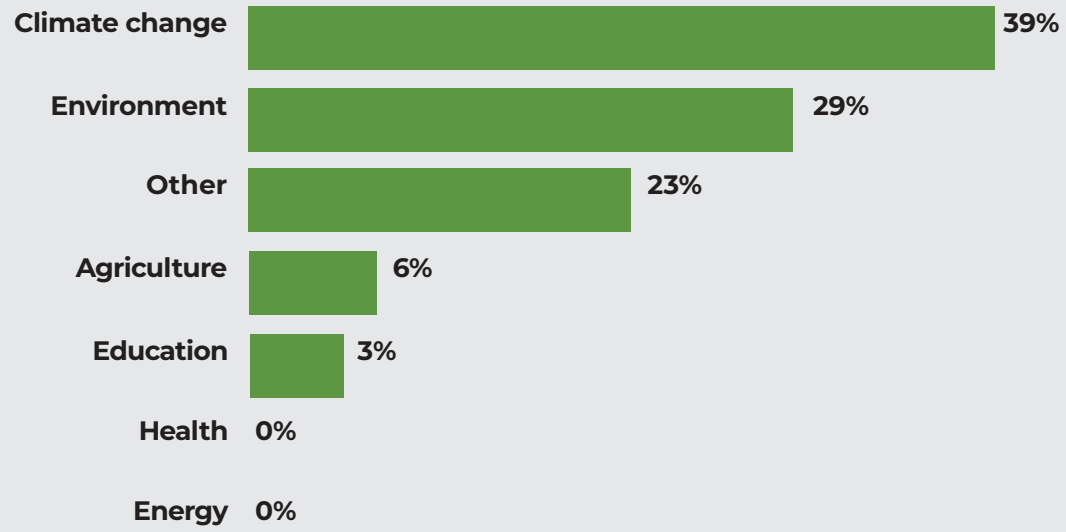
The majority of respondents (45 per cent) reported to be working for international organizations, indicating a strong representation from this sector in the field of program evaluation. Government agencies also had significant representation, with 26 per cent of respondents working in this sector. NGOs accounted for 10 per cent of the respondents, while academic institutions represent 6 per cent. Interestingly, there were no respondents from private sector consulting firms, which could suggest a lower engagement or representation of this group in the survey. The others (13 per cent of the respondents) worked for parastatal or intragovernmental organizations.

FIGURE 17: Organizational affiliation of survey respondents



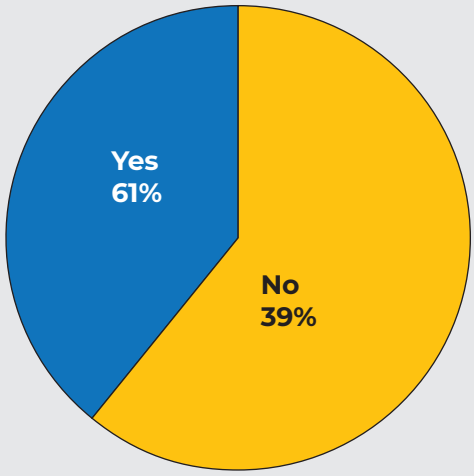
The majority of respondents had conducted evaluations in the areas of climate change (39 per cent) and environment (29 per cent). Some respondents were involved in other environmental programs, like biodiversity, and waste management (under the "Other" category (with 3 per cent). There was limited representation in other sectors, with agriculture accounting for 6 per cent and education for 3 per cent, while areas such as energy and health had no representation.

FIGURE 18: Areas of experience of survey respondents



A majority of respondents (61 per cent) had conducted evaluations of initiatives or projects funded by key climate finance mechanisms.

FIGURE 19: Experience of survey respondents in climate evaluations



Annex III: Reviewed working documents

Rao, S. (2024, June). OIOS-IED experiences with Artificial Intelligence (including Natural Language Processing) [PowerPoint slides]. Office of Internal Oversight Services, United Nations.

United Nations Secretariat. (2023, July 5). Guidance on the Responsible Use of Publicly Available Generative Artificial Intelligence (AI) Tools. Internal UN Broadcast Message.

Presentation “GenAI-powered evaluation function at UNFPA” by Neha Karkara, UNFPA Evaluation Office.

World Bank. (2024, April). AI Framework 2.0. Artificial Intelligence at WBG, Unpublished Presentation.

World Bank Group. (2024, May). Responsible Use of Generative Artificial Intelligence at the World Bank Group, Unpublished Document.

Annex IV: Bibliography

Mason, S. (2023). Finding a safe zone in the highlands: Exploring evaluator competencies in the world of AI. New Directions for Evaluation, 2023; 11–22.
Thornton, I. (2023). A special delivery by a fork: Where does artificial intelligence come from? New Directions for Evaluation, 2023; 23–32.

Head, C. B., Jasper, P., McConnachie, M., Raftree, L., & Higdon, G. (2023). Large language model applications for evaluation: Opportunities and ethical implications. New Directions for Evaluation, 2023; 33–46.

Nielsen, S. B. (2023). Disrupting evaluation? Emerging technologies and their implications for the evaluation industry. New Directions for Evaluation, 2023; 47–57.

Sabarre, N. R., Beckmann, B., Bhaskara, S., & Doll, K. (2023). Using AI to disrupt business as usual in small evaluation firms. New Directions for Evaluation, 2023; 59–71.

Ferretti, S. (2023). Hacking by the prompt: Innovative ways to utilize ChatGPT for evaluators. New Directions for Evaluation, 2023; 73–84.

Azzam, T. (2023). Artificial intelligence and validity. New Directions for Evaluation, 2023; 85–95.

Tilton, Z., LaVelle, J. M., Ford, T., & Montenegro, M. (2023). Artificial intelligence and the future of evaluation education: Possibilities and prototypes. New Directions for Evaluation, 2023; 97–109.

Reid, A. M. (2023). Vision for an equitable AI world: The role of evaluation and evaluators to incite change. New Directions for Evaluation, 2023; 111–121.
Montrosse-Moorhead, B. (2023). Evaluation criteria for artificial intelligence. New Directions for Evaluation, 2023; 123–134.

Shakil, H., Mahi, A. M., Nguyen, P., Ortiz, Z., Mardini, M. T. (2024a). Evaluating Text Summaries Generated by Large Language Models Using OpenAI's GPT. (retrieved from https://arxiv.org/pdf/2405.04053 on 11 August 2024).

Shakil, H., Ortiz, Z., & Forbes, G. C. (2024b). Utilizing GPT to Enhance Text Summarization: A Strategy to Minimize Hallucinations. (retrieved from https://arxiv.org/abs/2405.04039 on 11 August 2024).

Balvir, S., Tembhurney, S., Moharil, M., Anjarkar, S., Dhawale, B., Kamble, S., & Dogra, K. (2024). Text Summarization in The Age of Deep Learning: A Comprehensive Analysis. International Journal of Innovative Research in Technology, 10(11), 1379-1388.

Kates, A. W., & Wilson, K. (2023). AI for evaluators: Opportunities and risks. Journal of MultiDisciplinary Evaluation, 19(45), 99-104.

Hitch, D. (2024). Artificial intelligence augmented qualitative analysis: The way of the future? Qualitative Health Research, 34(7), 595-606.

Christou, P. A. (2024). Thematic Analysis through Artificial Intelligence (AI). The Qualitative Report, 29(2), 560-576.

Leeuw, F. & Bamberger, M. (2025). Artificial Intelligence and Big Data: Lessons from evaluations of Rule of Law and Development (Scheduled for publication in 2025: Edward Elgar Publishers).

Morgan L. D. (2023). Exploring the Use of Artificial Intelligence for Qualitative Data Analysis: The Case of ChatGPT. International Journal of Qualitative Methods, Volume 22: 1–10.

Raftree, L., & Tilton, Z. (2023). What's next for Emerging AI in Evaluation? Takeaways from the 2023 AEA Conference. MERL Tech.

Leo, B., Pattni, S., Winn, C., Lewis, Q., Paton, C. & Persaud, M. (2020). Using Machine Learning for Climate Related Impact Evaluations. eVALUation Matters, Second Quarter 2020, 23-32.

Ubaldi, B. & Zapata, R. (2024). Governing with Artificial Intelligence: Are governments ready?. OECD.

Burke, M. & Lobell, D.B. (2017). Satellite-based assessment of yield variation and its determinants in smallholder African systems. Proceedings of the National Academy of Sciences, 114(9), 2189-2194.

Richards, C. (2023). What has the World Bank's Internal Evaluation Group learned from experimenting with NLP?. World Bank Blogs. (three blogs)

Raimondo, E., Anuj, H., & Ziulu, V. (2023). Setting up Experiments to Test GPT for Evaluation. World Bank Blogs.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790-794.

Goldblatt, R., You, W., Hanson, G., & Khandelwal, A. K. (2016). Detecting the boundaries of urban areas in India: A dataset for pixel-based image classification in Google Earth Engine. Remote Sensing, 8(8), 634.

Samii, C., Paler, L., & Daly, S.Z. (2016). Retrospective causal inference with machine learning ensembles: an application to anti-recidivism policies in Colombia. Political Analysis, 24(4), 434-456.

Farley, S. (2017). Tracing Hurricane Maria's wake. Mapbox. (https://blog.mapbox.com/tracing-hurricane-marias-wake-7217a7d08380)

Xu, Y., & González, M. C. (2017). Collective benefits in traffic during mega events via the use of information technologies. Journal of the Royal Society Interface, 14(129), 20161041.

Parthasarathy, R., Rao, V., & Palaniswamy, N. (2017). Deliberative Inequality: A Text-As-Data Study of Tamil Nadu's Village Assemblies. World Bank Policy Research Working Paper, 8119.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076.

Björkegren, Daniel, and Darrell Grissen. 2020. "Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment." The World Bank Economic Review, 34(3): 618–634.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications, 11(1), 233.

Kim, K., & Boulanin, V. (2023). Artificial Intelligence for Climate Security: Possibilities and Challenges. Stockholm International Peace Research Institute (SIPRI) Policy Report.

Artificial Intelligence Forum of New Zealand. (2022). Artificial Intelligence for the Environment in Aotearoa New Zealand. https://aiforum.org.nz/wp-content/uploads/2022/11/Artificial-Intelligence-for-the-Environment-in-Aotearoa-New-Zealand.pdf

OECD (2022), “Measuring the environmental impacts of AI compute and applications: the AI footprint”, OECD Digital Economy Papers, No. 341, OECD Publishing, Paris.

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. AI & SOCIETY.

Lucas, M. (2024). The Role of AI in Climate Change Mitigation and Environmental Monitoring. Department of CS, UAF.

Dannouni, A., Deutscher, S. A., Dezzaz, G., Elman, A., Gawel, A., Hanna, M., Hyland, A., Kharij, A., Maher, H., Patterson, D., Jones, E. R., Rothenberg, J., Tber, H., Texier, M., & Ziat, A. (2023). Accelerating Climate Action with AI. Boston Consulting Group (commissioned by Google).

Kaack, L. H., Donti, P. L., Strubell, E., & Rolnick, D. (2020). Artificial intelligence and climate change: Opportunities, considerations, and policy levers to align AI with climate change goals. Heinrich-Böll-Stiftung.

Taghikhah, F., Erfani, E., Bakhshayeshi, I., Tayari, S., Karatopouzis, A., & Hanna, B. (2022). Artificial intelligence and sustainability: Solutions to social and environmental challenges. In N. Chakpitak, H. Abdel-Basset, A. Qishta (Eds.), Smart technologies for sustainable development: Artificial intelligence and data analytics approaches. Elsevier.

Chen, L., Chen, Z., Zhang, Y., Liu, Y., Osman, A. I., Farghali, M., Hua, J., Al-Fatesh, A., Ihara, I., Rooney, D. W., & Yap, P.-S. (2023). Artificial intelligence-based solutions for climate change: a review. Environmental Chemistry Letters, 21, 2525–2557.
Changlani, K., & Thakore, R. (2023). Artificial Intelligence for Climate Action. (retrieved from https://www.researchgate.net/publication/374899918 on 14 August 2023).

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters, 14(12), 124007.