UNIVERSITY OF BIRMINGHAM

GSDRC

GSDRC Working Paper 2026/01

Exploring the utility of artificial intelligence (AI) research tools in knowledge for development and diplomacy

William Avis,* Kaitlin Ball, Katherine Cheeseman, James Georgalakis, Sian Herbert, Iffat Idris, Zenobia Ismail, Brian Lucas, Shreya Jose Maliakal, Jasmin Morris, Sithandiwe Mujuru, Evert-Jan Quak, Sarah Verhaeg, Mahdi Zaidan

May 2026

Governance and Social Development Resource Centre (GSDRC)

K4DD Knowledge for Development and Diplomacy

UK International Development
Partnership | Progress | Prosperity

We advance

We activate

birmingham.ac.uk

Abstract: This study presents findings from a pilot study conducted under the Knowledge for Development and Diplomacy (K4DD) programme, commissioned by the UK Foreign, Commonwealth and Development Office (FCDO). The study evaluated the effectiveness, efficiency, and ethical implications of using artificial intelligence (AI) tools in producing rapid evidence synthesis products. Three categories of AI tools (generative AI chat, research assistants, and systematic review tools) were tested across nineteen K4DD outputs. While results reveal that AI tools offer promise, particularly in early-stage research, their current utility in responding to complex, multifaceted research queries remains limited. The study concludes with recommendations for researchers, programmes, and institutions to guide ethical, effective and appropriate AI integration.

Policy relevance: This study provides policymakers with an evidence-based assessment of the role of emerging AI research tools in the delivery of rapid evidence synthesis products for time-sensitive development and diplomacy decisions. It offers practical insights to guide responsible adoption, risk management, and future integration of AI within evidence-informed policy processes.

Keywords: artificial intelligence (AI), knowledge for development, rapid evidence review, research tools, ethics, large language models

Authors: William Avis, University of Birmingham: w.r.avis@bham.ac.uk. Kaitlin Ball, (independent). Katherine Cheeseman, Institute of Development Studies. James Georgalakis, Institute of Development Studies. Sian Herbert, University of Birmingham. Iffat Idris, University of Birmingham. Zenobia Ismail, University of Birmingham. Brian Lucas, University of Birmingham. Shreya Jose Maliakal, Institute of Development Studies. Jasmin Morris, Institute of Development Studies. Sithandiwe Mujuru, University of Birmingham. Evert-Jan Quak, Institute of Development Studies. Sarah Verhaeg, University of Birmingham. Mahdi Zaidan, Institute of Development Studies.

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this study.

Acknowledgements: This research was funded by the UK Foreign, Commonwealth and Development Office (FCDO) under the Knowledge for Development and Diplomacy programme (K4DD).

DOI: https://doi.org/10.48352/uobxgsdrc.0005

1 Introduction

From the early 2020s, artificial intelligence (AI) has increasingly moved beyond its role in automation and prediction, entering the realm of what Ye et al. (2025) refer to as creative and cognitive augmentation. Ye et al. (2025) reflect that this evolution has been accompanied by a shift from workflow replication (traditional automation) to generative co-creation, where AI tools actively contribute to the creative process. This development has been catalysed by the emergence of generative AI (GenAI) technologies, particularly large language models (LLMs) like GPT, which have demonstrated capabilities in generating coherent text, summarising complex information and assisting in ideation and hypothesis generation (OECD, 2023).

The emergence of large reasoning models (LRMs) marks a further development in AI. An LRM is an advanced type of AI system designed to perform structured, multi-step logical thinking, attempting to mimic how humans break down and solve complex problems. A prominent example is OpenAI's o3 model, part of a new generation of reasoning-focused LLMs. Given this evolution, AI tools are increasingly being integrated into research workflows.

Box 1: Key concepts in modern AI systems

Machine Learning (ML): Algorithms that learn patterns from data. **Used in b**oth LLMs and LRMs for training models to understand and generate or retrieve information.

Deep Learning (DL): A subset of ML using multi-layered neural networks. **Used in** LLMs and LRMs rely on deep learning for language understanding and generation.

Natural Language Processing (NLP): Enables machines to understand, interpret, and generate human language. Core to both LLMs and LRMs for tasks like summarisation, translation, question answering, etc.

Transformer Architecture: A neural network design that handles sequential data efficiently. The backbone of LLMs (e.g., GPT, BERT) and also used in LRMs for encoding queries and documents.

Generative AI (GenAI): AI that creates new content (text, images, etc.). **Used in** LLMs generate human-like text; LRMs use generation after retrieving relevant information.

Retrieval-Augmented Generation (RAG): Combines retrieval systems with generative models. Used in LRMs to fetch relevant documents and generate grounded, accurate

Whilst the application of AI research tools has become increasingly common in certain disciplines (e.g., health research) and for particular types of research (e.g., systematic reviews), the application of AI tools in demand-responsive rapid-synthesis research is less common. This paper explores the application of AI tools within the Knowledge for Development and Diplomacy (K4DD) programme. K4DD provides demand-responsive synthesis research for the Foreign, Commonwealth and Development Office (FCDO), with an emphasis placed on rapid delivery of outputs, i.e., within one to two months.

This study is situated within a socio-technical systems perspective, which emphasises the interplay between technological tools and the social, institutional, and epistemic environments in which they are embedded (Bijker, Hughes, and Pinch 1987; Jasanoff 2004). In this paper we suggest that rather than viewing AI tools as neutral or passive instruments, it is important to recognise that their utility and impact are shaped by the contexts in which they are deployed, in this case, the rapid, policy-oriented research environment of the K4DD programme. The study also draws on the emerging literature on human-AI collaboration and collaborative intelligence, which conceptualises AI not as a replacement for human researchers but as a 'cognitive' partner that can augment human capabilities in ideation, synthesis, and decision-making (Zou et al., 2025; Shneiderman, 2020). This aligns with theories of distributed cognition (Hollan, Hutchins, and Kirsh, 2000), which argue that knowledge production is distributed across people, tools, and environments. Finally, the study engages with the concept of epistemic cultures (Knorr-Cetina, 1999), recognising that the norms, values, and practices of evidence synthesis in international development differ significantly from those in domains such as health research. This framing allows us to interrogate not only the technical performance of AI tools in the generation of rapid evidence reviews (RERs) but also their epistemological fit within the unique demands of K4DD RERs specifically for diplomacy and development.

2 Objectives and research questions

The objective of this paper was to assess the effectiveness, efficiency, and quality of AI tools in the selected stages of the generation of rapid evidence synthesis products compared to existing methods over a six-month period (October 2024 to April 2025). The paper addresses three key questions:

- Which types of AI tools and their underlying technologies, including ML and LLMs, could be of use in the production of K4DD products and how?
- What is the impact of AI tools on quality, validity, relevance, verifiability and speed in the K4DD service?
- What are the ethical and legal consequences of using AI tools in rapid reviews?

3 Literature review

The integration of AI tools into research workflows has expanded significantly in recent years, particularly in the production of literature reviews with numerous tools emerging that purport to expedite various stages of the research process. Examples of the application of AI tools include in the automation of literature search, relevance screening, summarisation and synthesis (Ye et al., 2025; OECD, 2023).

Global academic publishers, such as Springer, have also reflected more broadly on the role of AI in the generation of academic research. They state that LLMs, such as ChatGPT, do not currently satisfy their authorship criteria. They also reflect that use of an LLM should be properly documented in the Methods section of the manuscript.

There is an expanding evidence base that explores the application of AI tools in the preparation and execution of systematic literature reviews (SLRs) (Bolaños et al., 2024; Ofori-Boateng, 2024; de la Torre-López et al., 2023). SLRs have traditionally relied on manual processes for screening and coding, which are time-intensive and can be susceptible to human bias. AI is increasingly employed to support these tasks through ranking and automation. Ofori-Boateng et al (2024) reflect that AI tools have been used to automate stages of reviews — particularly leveraging natural language processing (NLP), machine learning (ML) and deep learning (DL). This includes use in the search, screening (title and abstract), data extraction, and risk of bias (RoB) assessments. Wallace et al. (2010) demonstrated how machine-learning algorithms could effectively prioritise articles for human review, improving the efficiency of the selection process. More recently, Marshall and Wallace (2019) showed how active learning techniques could drastically reduce the screening burden while maintaining high inclusion accuracy.

Further to the ranking and automation of SLRs, AI-driven summarisation tools now offer capabilities to extract key concepts, methods, and findings from academic articles (OECD, 2023). Tools like Scholarcy and GPT-based systems can generate summaries of individual papers or synthesise content across multiple documents (Smalheiser, 2017). However, concerns remain about the fidelity and interpretability of AI-generated summaries, especially when dealing with nuanced or domain-specific literature (Tang et al., 2022).

Although AI tools have been shown to have the potential to increase efficiency of SLRs, they are not without limitations. Research has shown that AI tools can inadvertently reproduce biases present in training data and miss critical papers due to limitations in algorithmic scope or lack of access to paywalled databases (Gusenbauer and Haddaway, 2020). Additionally, reliance on LLMs raises questions about transparency and reproducibility (Marcus and Davis, 2020).

More critically, recent research suggests that LRMs, an advanced form of AI, faces a 'complete accuracy collapse' when presented with highly complex problems (Shojaee et al., 2025). This study found that standard AI models outperformed LRMs in low-complexity tasks, while both types of model suffered 'complete collapse' with high-complexity tasks. The study, which tested the models' ability to solve puzzles, added that as LRMs neared performance collapse they began 'reducing their reasoning effort' (Shojaee et al., 2025).

Given these concerns, Zou et al. (2025) urge the use of AI as a collaborator rather than a replacement. The emphasis thus shifts toward human-AI partnerships where the machine augments human insight rather than replacing it.

More broadly, there exist ethical concerns regarding the use of AI tools. Indeed, the ethical use of AI in literature reviews requires careful oversight to avoid misrepresentation or omission of critical studies. Whilst AI technology matures, some, (e.g., Booth, Sutton, and Papaioannou, 2021) reflect that researchers should be advised to treat AI outputs as assistive rather than authoritative, ensuring that the researcher remains central. The increasing use of AI tools also necessitates new guidelines for citation and disclosure of automated assistance in research (Lipworth et al., 2023).

Beyond the application of AI tools in SLRs, a more limited literature has explored application of AI tools in other types of synthesis products.

In the context of this paper, we are focused on what has been referred to variously as rapid evidence reviews (RERs) or ultra rapid literature reviews (URLRs). Although there is no single rigid approach to implementing a RER[1] or what they can deliver, they are broadly considered to serve the following functions:

- RERs provide evidence in a timely and cost-effective manner (Tricco, Langlois, and Straus, 2017).
- RERs provide a tool for getting on top of the available research evidence on a policy issue, as comprehensively as possible, within the constraints of a given timetable (Government Social Research Service, 2014).

Our study found limited evidence regarding the application of AI tools in the preparation of RERs or URLRs. That is, those reviews that are executed over days rather than weeks, months or years. Such reviews are emerging as a distinct methodological response to time-sensitive decision-making needs in domains like public health, clinical practice, and policy development. These reviews aim to deliver robust evidence synthesis in days or weeks rather than months, necessitating streamlined workflows.

The evidence that does exist highlights that AI tools have been utilised in the preparation of URLRs by automating literature searches, deduplication, relevance ranking, and even preliminary data extraction. Authors such as Tricco et al. (2017) suggest that AI can significantly reduce turnaround times without wholly sacrificing quality. In a COVID-19 context, for example, AI-assisted URLRs were used to rapidly synthesise evidence on emerging treatments and public health strategies (Haby et al., 2016;).

Gusenbauer and Haddaway (2020) caution that the speed of URLRs comes with trade-offs. AI-generated outputs may bypass nuanced human appraisal or introduce selection bias, especially when using LLMs trained on general corpora (Gusenbauer and Haddaway, 2020).

[1] Rapid evidence reviews are also referred to as Rapid Evidence Assessment or Rapid Reviews

Transparency, auditability, and documentation of AI's role in such reviews is essential for maintaining research integrity.

More broadly, it is also important to note that AI tools are evolving rapidly. For example, in 2023, Haman and Školník (2024) tested the utility of ChatGPT in conducting a literature review, and found that two-thirds of the papers that ChatGPT claimed to have found did not in exist. They concluded that 'we firmly recommend not using ChatGPT in the research process' (Haman and Školník, 2024: 1245). Two years later, the researchers repeated their experiment using ChatGPT in conjunction with its Deep Research tool. This time, the researchers found that ChatGPT with Deep Research was very effective in identifying real and influential research papers. Haman and Školník (2025: 3) concluded that 'in direct contrast to the conclusion of our previous publication, we must now take the opposite stance: we strongly recommend the use of ChatGPT with Deep Research in the research process'.

Our paper acknowledges that AI tools are rapidly transforming how literature reviews are conducted, offering potential gains in efficiency, coverage, and scalability. However, their use must be balanced with critical human oversight to ensure the validity, transparency, and ethical rigor of outputs.

Recent research emphasises that the integration of AI tools into research workflows should be understood through a socio-technical lens. Technologies such as LLMs and reasoning models (LRMs) do not operate in isolation; rather, their utility and impact are shaped by the institutional, epistemic, and ethical contexts in which they are deployed (Jasanoff, 2004; Bijker et al., 1987). In the context of rapid evidence synthesis, AI tools must be evaluated not only for their technical capabilities but also for their alignment with the norms and expectations of policy-oriented research. The K4DD programme, with its emphasis on speed, relevance, and responsiveness to complex policy questions, presents a unique socio-technical environment that challenges the assumptions embedded in many AI tools designed for more structured or academic domains.

The concept of collaborative intelligence offers a useful framework for understanding the evolving role of AI tools in research. Rather than replacing human researchers, AI tools are increasingly seen as cognitive partners that augment human capabilities in ideation, synthesis, and decision-making (Shneiderman, 2020; Zou et al., 2025). This perspective aligns with theories of distributed cognition, which posit that knowledge production is distributed across people, tools, and environments (Hollan et al., 2000). However, the effectiveness of such collaboration depends on the epistemic culture in which it occurs. As Knorr-Cetina (1999) argues, different fields of inquiry have distinct ways of producing and validating knowledge. The K4DD programme operates within an epistemic culture that values rapid, policy-relevant synthesis over exhaustive academic rigor. This creates tensions when AI tools, often trained on general or biomedical corpora, are applied to complex, under-researched, or politically sensitive topics. Understanding these tensions is essential for designing and deploying AI tools that support human researchers in diverse knowledge production contexts.

3.1 Research gap and contribution

While existing research has explored the application of AI tools in SLRs, particularly within biomedical and health domains (Marshall and Wallace, 2019; Bolaños et al., 2024), there remains a notable gap in understanding how these tools perform in more dynamic, policy-oriented research environments. The epistemic culture of rapid evidence synthesis, characterised by time-sensitive, multifaceted, and often politically nuanced queries, poses distinct challenges that are not adequately addressed by AI tools designed for structured academic reviews. This study contributes to the literature by empirically examining the utility of AI tools within the K4DD programme — a setting that is reflective of the sociotechnical complexity of demand-driven policy research. By applying a sociotechnical systems lens (Jasanoff, 2004) and engaging with theories of collaborative intelligence (Zou et al., 2025) and epistemic cultures (Knorr-Cetina, 1999), this study not only evaluates tool performance but also interrogates the conditions under which AI can meaningfully augment human reasoning in rapid, applied research contexts. In doing so, it offers a nuanced understanding of the limitations and potential of AI tool integration beyond traditional academic domains.

4 Methodology

This study employed a mixed-methods approach to evaluate the utility of AI research tools within the K4DD programme. A rapid review of available AI tools was first conducted to identify suitable candidates based on three criteria: tool type (generative AI chat, research assistant, systematic review), institutional risk (as determined by terms and conditions), and cost-effectiveness. Three tools were selected and applied across 19 K4DD outputs, including evidence summaries, rapid evidence reviews, annotated bibliographies, and emerging issues reports. Data collection involved three complementary methods: (1) structured pro formas completed by researchers to capture tool usage and reflections across five performance dimensions: quality, validity, relevance, verifiability, and speed; (2) three rounds of researcher polling to explore researcher positionality in relation to and track evolving perceptions of AI tools; and (3) facilitated workshops to gather qualitative insights and foster collective reflection. This methodology was grounded in a sociotechnical systems perspective, recognising that the effectiveness of AI tools is shaped not only by their technical capabilities but also by the institutional, epistemic, and ethical contexts in which they are deployed.

Three tools were selected, informed by three key considerations: Type of tool: To ensure coverage of the different types of tools including generative AI chat, research assistance and systematic review. Terms and conditions: Tools were included that were considered to pose least institutional risk (see Box 2). Cost: To ensure value for money, the pilot included tools for which existing subscriptions were held, where access would be provided free of charge and where costs were considered to be acceptable.

The first tool selected was an AI-powered generative chat tool designed to enhance productivity. It integrates with other programmes, offering contextual assistance to users.

The tool can help draft documents, analyse data, create presentations, summarise emails, and manage projects, making it a versatile tool. The technology behind the tool involves LLMs; it leverages AI models, such as GPT-4, to understand and generate human-like text. These models are fine-tuned using supervised and reinforcement learning techniques.

The second tool selected was an AI-powered research assistant designed to streamline and enhance the research process. It leverages advanced language models to automate tasks such as literature reviews, data extraction, and summarisation. This makes it useful for conducting reviews and synthesising information from extensive academic databases. The technology behind the tool enables users to analyse research papers, extract data from large volumes of text, and generate insights. It supports researchers by providing tools for search, discovery, and source verification.

The third tool selected was a web-based software designed to support SLRs and other types of literature reviews. The tool is particularly useful for managing references, storing PDFs, and conducting both qualitative and quantitative analyses. The technology behind the tool includes features such as text-mining; this streamlines a SLR by identifying relevant studies more efficiently. It connects with resources like OpenAlex and Zotero for reference management and data import/export.

To explore the application of AI research tools across the work of K4DD, four types of K4DD knowledge product were included in the pilot. These were included to examine the application of tools over different time frames (i.e., research products compiled for which one research day was allocated to those that were more substantial, that is products involving an allocation of twenty research days). A description of products included is provided below (the figure in brackets reflects number of these products included in the pilot):

Evidence summaries (1) are based on one day of desk-based research. The rapid bi-weekly search for peer-reviewed literature is carried out through a keyword search, restricted to articles published in English in the previous two weeks. **Rapid evidence reviews (15)** are based on six days of desk-based research. They provide summaries of current research, evidence, and lessons learned. They are intended to provide an overview of the most important evidence related to a research question. **K4DD Rapid annotated bibliographies (1):** Annotated bibliographies are based on six days of desk-based research. They provide brief summaries of current research, evidence, and lessons learned. They are intended to provide an overview of the most important evidence related to a research question. **K4DD Emerging issues reports (2):** Emerging issues reports are based on 20 days of desk-based research and five days of expert input and highlight research and emerging evidence for policymakers. K4DD works with thematic experts and FCDO to identify where new or emerging research can inform and influence policy.

The testing phase of the AI pilot involved applying the selected AI tools to the production of 19 K4DD outputs (see Table 1).

Table 1: Researcher and K4DD output allocation

Tool 1: generative AI chat tool	Tool 2: research assistant	Tool 3: systematic review tool
Seven RERs One EIR	Eight RERs One EIR	Evidence summary
One RER		

Source: authors' illustration.

Box 2: The importance of reviewing terms and conditions of AI tools

When selecting a tool it is crucial to review the terms and conditions which form the legal agreement between the supplier and the user (user here may also mean the organisation that the individual user works for). These terms should specify who can access the service, the costs involved, any usage restrictions, the supplier's and users' legal obligations, and the allocation of responsibilities and legal costs in case of issues.

Content and intellectual property

Upload: Many AI tools require users to upload documents for summarisation, reformatting, or the tool extracts information to present it in a new and useful way. It is imperative that users understand the permissions granted to the tool provider for the uploaded content, as only the copyright owner can authorise the reuse of their work. Some AI tools have broad terms that allow providers to reuse the content beyond the intended service. Users cannot grant rights to works if they are not the copyright owner, making these tools impractical if users adhere strictly to the terms as it would restrict them to only being able to upload their own works that they hold the copyright in. AI tool providers must be aware of this paradox, but still offer the service, and protect themselves via warranty and indemnification clauses.

Warranties and indemnification clauses: To mitigate copyright issues, terms and conditions often require users to upload only items they own the copyright for and provide a warranty for this. A number of AI tool providers have clauses where the end user is indemnifying the supplier of the tool. If a legal case for infringement is brought the end user will be liable for all legal costs including those of the supplier and the party suing. These costs are not normally capped and would often be held under the US legal system, meaning that costs incurred could be quite significant. Traditional content providers, such as journal publishers, typically warrant that their content has the necessary permissions and indemnify the end user against legal costs.

Reuse and commercialisation: Terms and conditions usually specify how outputs can be reused. Some AI tools allow only snippets of outputs, while others restrict reuse to non-commercial purposes. This can significantly impact researchers who wish to use the materials for commercial purposes, such as inclusion in a book, developing further AI tools for use in industry, or using outputs to form products and/ or services.

Accessibility

In the UK, Public Sector Bodies must ensure that all web services meet the WCAG 2.2 AA accessibility standard and have accessibility statements. Many new AI tools lack both despite WCAG standards being a legal requirement in many jurisdictions.

4.1 Data collection methods

To fully understand the impact of AI research tools on the generation of K4DD knowledge products, a mixed-methods approach was adopted to capture both qualitative and quantitative data pertaining to research experiences. Data was collected through three methods (pro forma, working groups, and polling),

4.2 Pro formas

For each product included in the AI pilot, researchers completed a pro forma to capture experiences. This collected a mix of qualitative (e.g., perceptions regarding the efficacy of AI tools) and quantitative (e.g., how researchers used the tools) information (see Annex B). The pro forma was structured as follows:

1. Tool use across different elements of the research process.
2. Overarching reflections on the impact of AI tool use on the research process (positive and negative).
3. Assessment of AI tool against a set of criteria.
    - Quality: i.e., the extent to which the output of the AI tool meets high standards in terms of accuracy, completeness, clarity, and consistency.
    - Validity: i.e., the degree to which the AI tool produces results that are credible and align with established or accepted findings. Validity focuses on whether the tool's results are scientifically or logically sound and based on reliable evidence.
    - Relevance: i.e., the appropriateness of the AI tool's output to the specific goals or questions of the evidence review. Relevance assesses whether the information provided is directly applicable and helpful to the topic under investigation.
    - Verifiability: i.e., the ability to trace and confirm the accuracy and sources of the AI-generated information. Verifiability ensures that the AI tool's outputs are transparent and can be cross-checked against primary sources or evidence.
    - Speed: i.e., the efficiency and timeliness with which the AI tool processes data and delivers results. Speed measures how quickly the tool can complete tasks like evidence gathering, synthesis, and reporting compared to traditional methods.

4.3 Workshop group discussions

To complement the above, three workshops involving participating researchers and invited experts were organised. These were convened as semi-structured group discussions on the use of tools and to share experiences and collate feedback. The intended purpose of these sessions was to generate a more collective understanding of the use of tools across different

types of outputs and for these discussions to compliment the output specific reflections captured in the pro formas.

4.4 Polling of researchers

Three polls were circulated to participating researchers to capture experiences of using AI tools. Poll 1 was executed at the start of the pilot, poll 2 at midpoint, and poll 3 upon completion of the testing phase. These polls were designed to capture, if at all, changing perceptions regarding tool use. Researchers were asked to provide their perceptions of AI tools based on their individual experience, scoring responses to the following questions across a Likert scale of 1–5 (1 = strongly disagree, to 5 = strongly agree). Questions related to the following themes, quality, validity, relevance, verifiability and speed.

Polls also asked researchers to categorise themselves according to a series of profiles identified by Oxford University Press research (2024: 10):

> *Researcher profile: Based on your experimentation with AI research tolls to date, how would you categorise your research profile in relation to attitudes towards, and perceptions of AI research tools.*

- Challengers: 'I'm completely against AI'
- Sceptics: 'I don't trust AI; I don't need AI'
- Concerned pessimists: 'AI negatively impacts research, and my role'
- Wary observers: 'AI negatively impacts research, but it doesn't affect me'
- Neutrals: 'I'm still weighing up the positive and negative effects'
- Cautious time savers: 'I don't trust AI, but it will save time'
- Careful optimists: 'I'm excited, but concerned'
- Pioneers: 'I'm fully embracing AI'

5 Results and analysis

5.1 Pro formas

Tool use

To better understand how research tools were applied across the research process, researchers were asked to record how they utilised the tools. Findings are presented below (see Table 2) with green indicating most and red indicating least use. The findings are tool-specific.

Table 2: Most common (green) to least common (red) recorded usage of tool

AI research tool use										

Source: authors' illustration.

Data collated from pro formas highlighted that different tools perform different functions depending on the research output commissioned and the researcher using the tool. Broad reflections include the following (see Table 3 for an illustrative overview):

Generative AI chat tools (i.e., tool 1) were principally utilised by researchers to support initial immersion in a topic with the majority of researchers identifying preliminary research (eight mentions) and query scope clarification (nine mentions) as the main application of the tool.

Research assistant tools (i.e., tool 2) had a more mixed application across a number of stages, from preliminary research (seven mentions) and query scope clarification (six mentions) to literature search (seven mentions) and screening (seven mentions).

Systematic review tools (i.e., tool 3) had less utility in the preliminary research and query scope clarification stages but leant itself to literature search (one mention), screening (one mention) and ranking stages (one mention).

Table 3: Most common to least common recorded usage of tools

	Initial negotiation and defining the question		**Research process**				**Writing**	**Quality assurance**
	Preliminary research	Query and scope clarification	Literature search	Screening	Ranking	Quality assessment	Writing	Clarity and concision
Tool 1	8	9	6	4	3	4	3	2
Tool 2	7	6	7	7	3	5	4	4
Tool 3	0	0	1	1	1	0	0	0

Source: authors' illustration.

Across all tools researchers highlighted that they were less useful in terms of assessing the quality of sources, writing reviews or enhancing the clarity and conciseness of reviews.

Sentiment analysis

The research team synthesised findings across the 19 outputs included in the pilot, coding statements according to their dominant sentiment i.e. positive, negative and mixed:

Positive: Positive sentiment was recorded when the statement was considered to express constructive, optimistic, or supportive views of the tools utility in responding to the research question, e.g.:

Meets high standards in terms of accuracy, completeness, clarity, and consistency; quality improves with access to full papers. (tool 2)

Negative: Negative sentiment was recorded when the statement was considered to convey criticism, denial, or a less favourable perspective of the tools' utility in answering a research question, e.g.:

Lack of transparency and limited sources made it difficult to integrate key findings. (tool 1)

Mixed: Mixed sentiment was recorded when the statement was considered to convey both positive and negative elements, e.g.:

Acceptable quality; directs the researcher to relevant organizations but lacks prioritization of academic sources. (tool 1)

Statements were coded according to sentiment across thematic areas, tool and across all tools (see Table 4 for an overview of statements coded).

Table 4: Sentiment analysis total outputs

	Total	**Tool 1**	**Tool 2**	**Tool 3**
Quality	19	9	9	1
Validity	19	9	9	1
Relevance	19	9	9	1
Verifiability	19	9	9	1
Speed	19	9	9	1
TOTAL	**95**	**45**	**45**	**5**

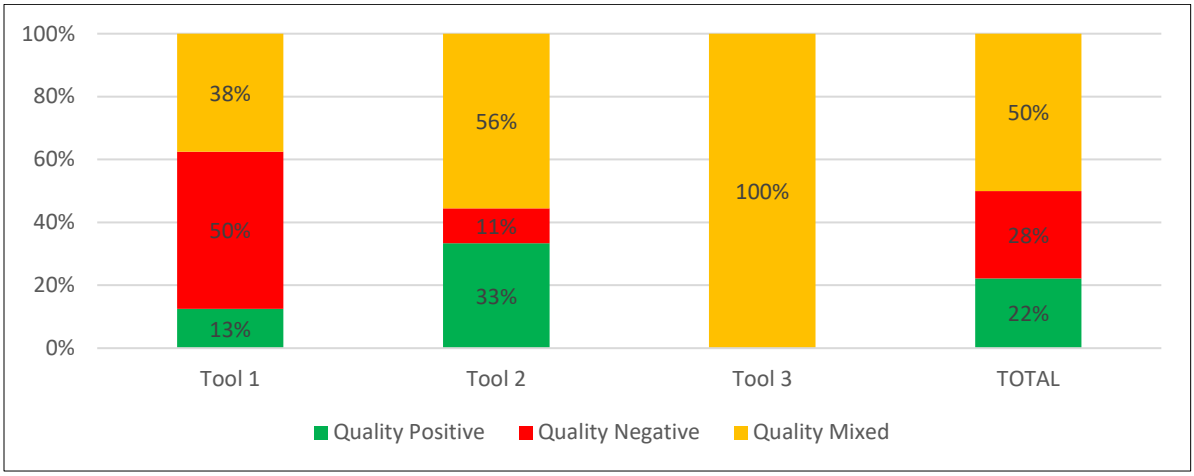
Source: authors' illustration.

Across all outputs, sentiments were coded and presented per theme, per tool and across all AI tools (see Figures 1-6). Whilst tool 3 (systematic review tool) is included in the figures below, it was excluded from the individual tool analysis as only one pro forma was coded. Findings are, however, included in the combined tool assessment.

Quality

Researchers were asked to reflect on whether the output of the AI tool met high standards in terms of accuracy, completeness, clarity and consistency. The quality of AI tool output is crucial for ensuring accuracy, reliability and researcher trust.

Figure 1: Sentiment analysis (quality)



Source: authors' illustration.

In terms of the coding of statements regarding quality, findings suggest that the tool performance was mixed (50% of statements) with more negative (28%) than positive (22%) statements reported across all tools. Tool 2 (research assistant) was viewed more positively (33%) than tool 1 (Generative AI chat tool) (13%), however, the majority of researcher statements were mixed (56%) regarding tool 2.

Validity

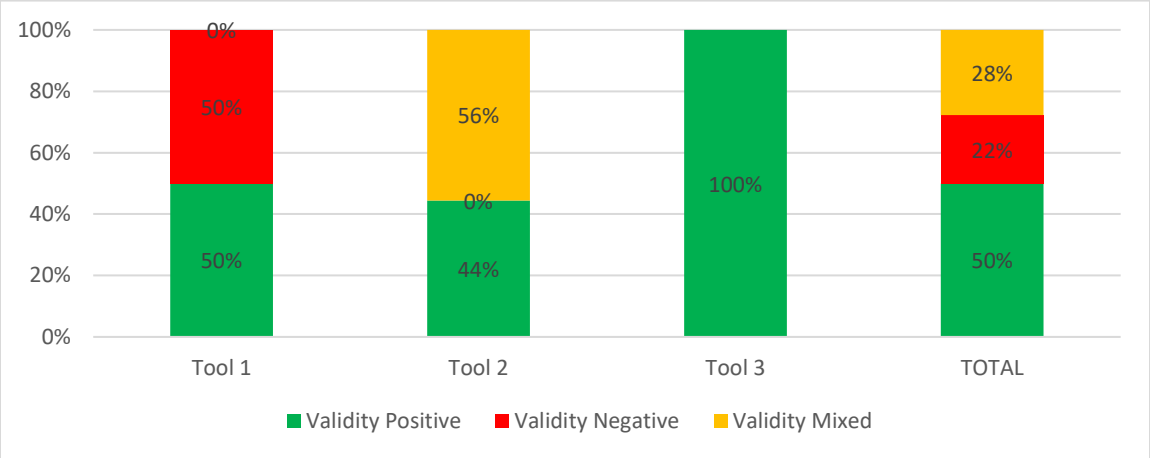
Researchers were asked to reflect whether the AI tool produced results that were credible and aligned with established or accepted findings. Validity focused on whether the tool's results are scientifically or logically sound and based on reliable evidence. Validity in AI outputs matters because it ensures the results are reflective of what they are intended to represent.

In terms of validity the tools performed relatively well with more positive (50%) than negative (22%) statements reported across all tools. In terms of tools, tool 1 (Generative AI chat) performed better (50% positive statements) compared to tool 2 (research assistant) (44% positive statements).

Relevance

The researchers were asked to reflect on whether the AI tool's output was appropriate to the specific goal or questions of the evidence review. Relevance assesses whether the information provided is directly applicable and helpful to the topic under investigation. Relevance in AI outputs is essential because it ensures that the information provided is meaningful and contextually appropriate. Relevant outputs help users make informed decisions.

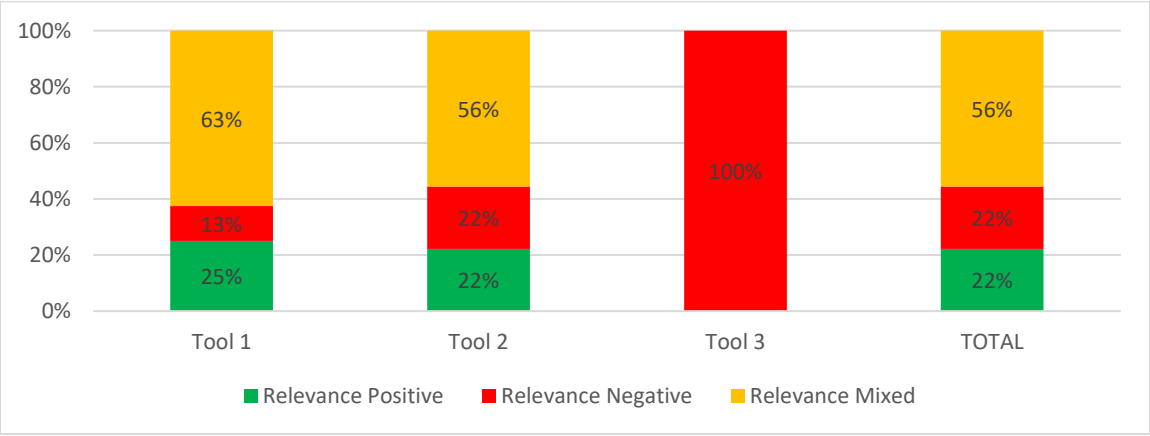
Figure 2: Sentiment analysis (validity)



Source: authors' illustration.

In terms of relevance tool performance was relatively mixed (56% of statements) with negative (22%) and positive (22%) statements reported across all tools. In terms of tools, tool 1 (Generative AI chat) (25% positive statements) performed marginally better than tool 2 (research assistant) (22% positive statements), though the majority of statements expressed a mixed sentiment (63% for tool 1 and 56% for tool 2).

Figure 3: Sentiment analysis (relevance)

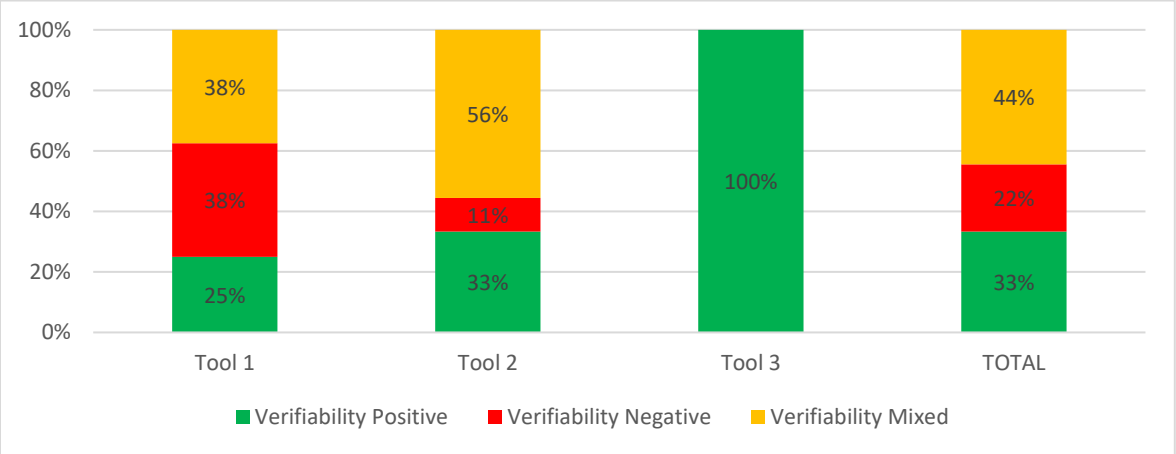


Source: authors' illustration.

Verifiability

Researchers were asked to reflect whether they were able to trace and confirm the accuracy and sources of the AI generated information. Verifiability ensures that the AI tool outputs are transparent and can be cross-checked against primary sources or evidence. Verifiability in AI outputs is crucial because it ensures that the information generated can be independently confirmed.

Figure 4: Sentiment analysis (verifiability)



Source: authors' illustration.

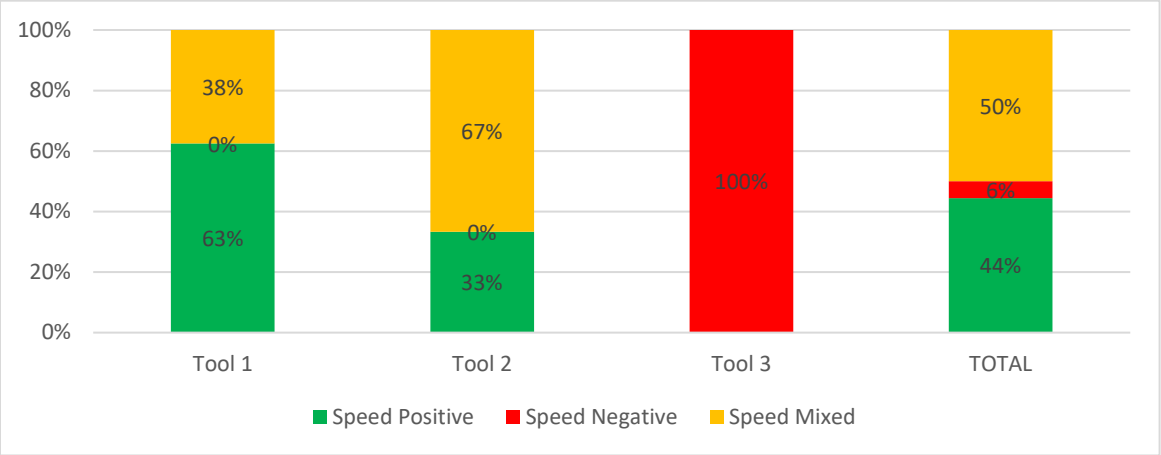
In terms of verifiability, tool performance was considered mixed (44% of statements) with negative (22%) and positive (33%) statements reported across all tools. In terms of tools, tool 2 (research assistant) (33% positive statements) performed better than tool 1 (Generative AI chat) (25% positive statements).

Speed

Researchers were asked to reflect on whether the AI tool processed data and delivered results in an efficient and timely manner. Speed measures how quickly the tool can complete tasks — such as evidence gathering, synthesis, and reporting — compared to 'traditional' methods. Speed in AI outputs is vital because it directly impacts user experience, efficiency, and the practical application of AI tools.

In terms of speed, tool performance was relatively mixed (50% statements) with less negative (6%) than positive (44%) statements reported across the tools. In terms of tools, tool 1 (Generative AI chat) (63% positive statements) performed better than tool 2 (research assistant) (33% positive statements).

Figure 5: Sentiment analysis (speed)



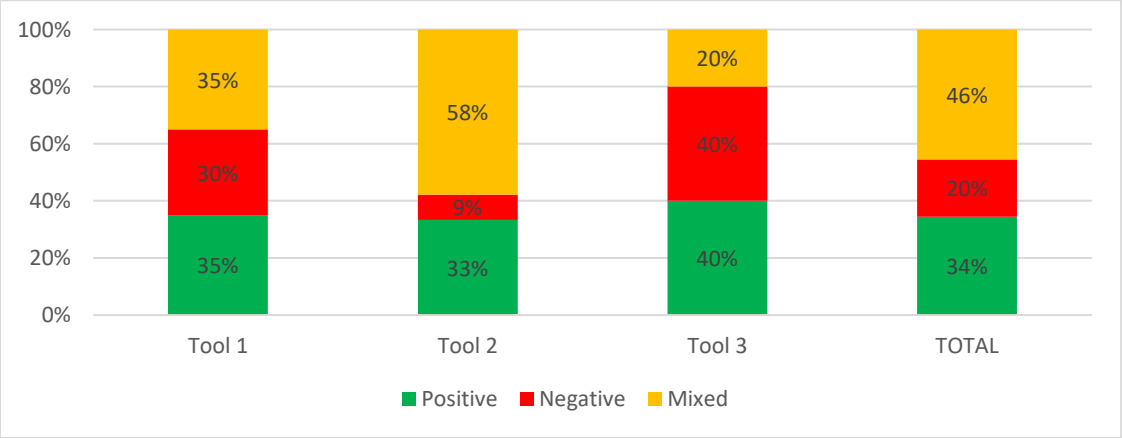
Source: authors' illustration.

Collated feedback

To provide an overview of sentiment pertaining to each AI tool and of all responses across the AI pilot, responses were collated. Findings suggest that perceptions regarding the tools is broadly mixed (46% of statements). Mixed sentiment was recorded when the statement was considered to convey both positive and negative elements. There were, however, more positive (34% of statements) than negative (20% of statements) sentiment recorded across all tools.

In terms of tools overall, tool 1 (Generative AI chat) (35% positive statements) performed marginally better than tool 2 (research assistant) (33% positive statements). However, researchers recorded more negative statements regarding tool 1 (30% of statements) compared to tool 2 (9% of statements).

Figure 6: Sentiment analysis consolidated



Source: authors' illustration.

Researcher polling

During the pilot, researchers were asked to complete a poll that would be repeated three times during the project. The intention of this poll was to capture changes, if any, in perceptions regarding AI research tools over time. A Likert scale (see Table 5) was utilised to help researchers quantify subjective opinions and analyse trends over time. The ten researchers involved in the pilot participated in each poll.

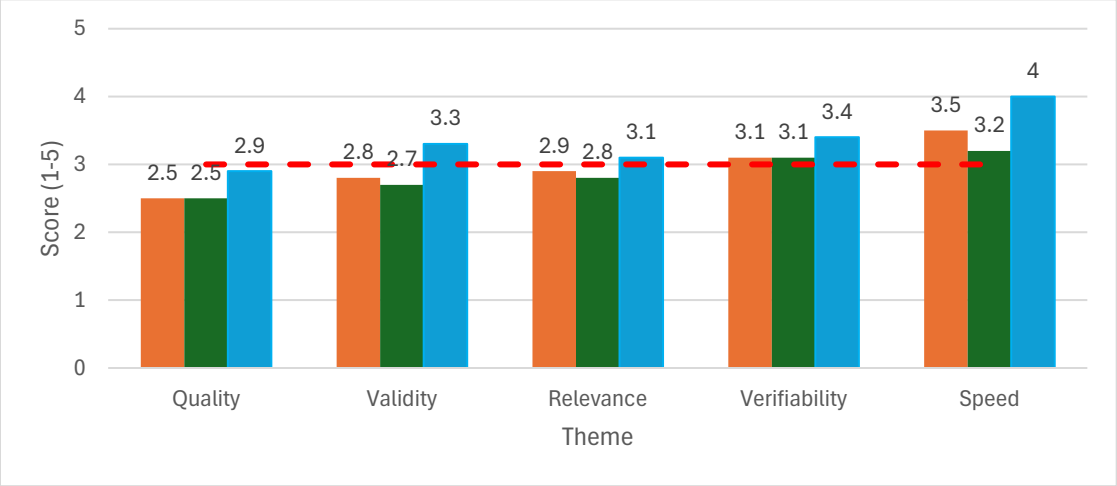
Table 5: Likert scale utilised in the poll

Completely disagree	Disagree	Neutral	Agree	Completely agree
1	2	3	4	5

Source: authors' illustration.

Findings were collated to provide an overview of changes in sentiment over time across all researchers. Whilst there were subtle shifts during the pilot, findings suggest that perceptions were broadly neutral in response to statements regarding quality, validity, relevance, verifiability and positive regarding speed (see Figure 7).

Figure 7: Consolidated polling (average across all researchers per poll)



Source: authors' illustration.

Researchers were also asked to respond to the following question to explore whether their positionality changed on relation to the application of AI tools in their research see Figure 8).

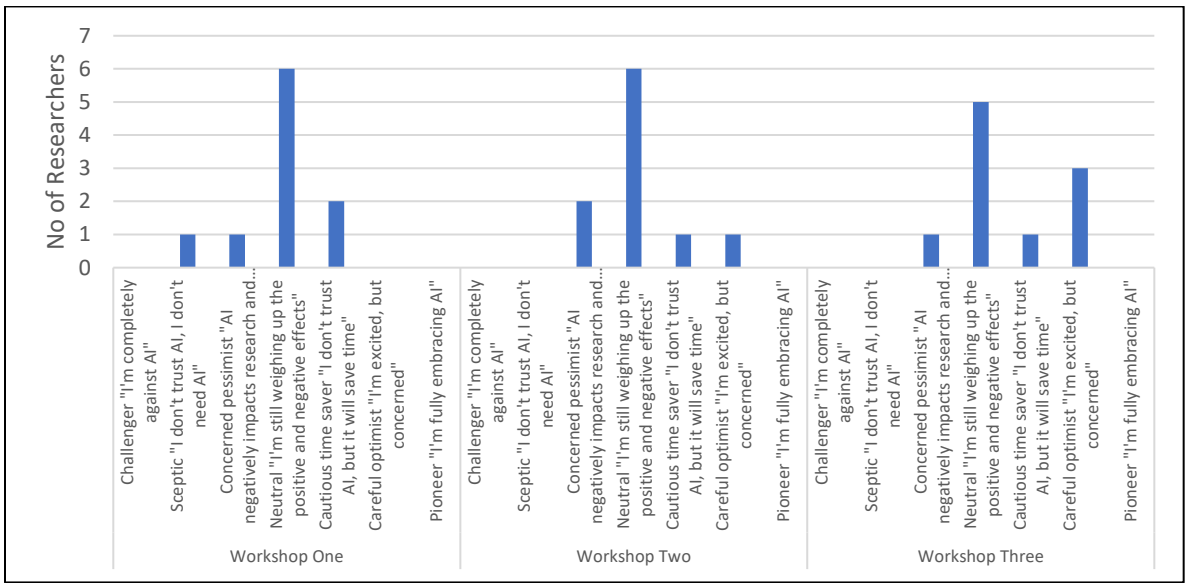
Based on your experimentation with AI research tolls, how would you categorise your research profile in relation to attitudes towards, and perceptions of AI research tools.

In workshop one, 10% of researchers considered themselves sceptics, 10% concerned pessimists 60% neutrals and 20% cautious time savers.

In workshop three there had been a shift in opinions. Of the researchers involved 10% considered themselves concerned pessimists 50% neutrals 10% cautious time savers and 30% careful optimists. That is sceptics had become less sceptical and cautious timesavers had become careful optimists.

Despite these shifts and reflecting findings across other elements of the pilot, 50% of researchers (5 of 10) categorised themselves as neutrals — 'I'm still weighing up the positive and negative effects'. An interesting finding from this poll is that whilst no researchers categorised themselves as 'Careful optimists — I'm excited, but concerned' in the first poll, 30% of respondents (3 of 10 researchers) recorded this in the final poll.

Figure 8: Researcher profile relation to attitudes towards, and perceptions of AI research tools.



Source: authors' illustration.

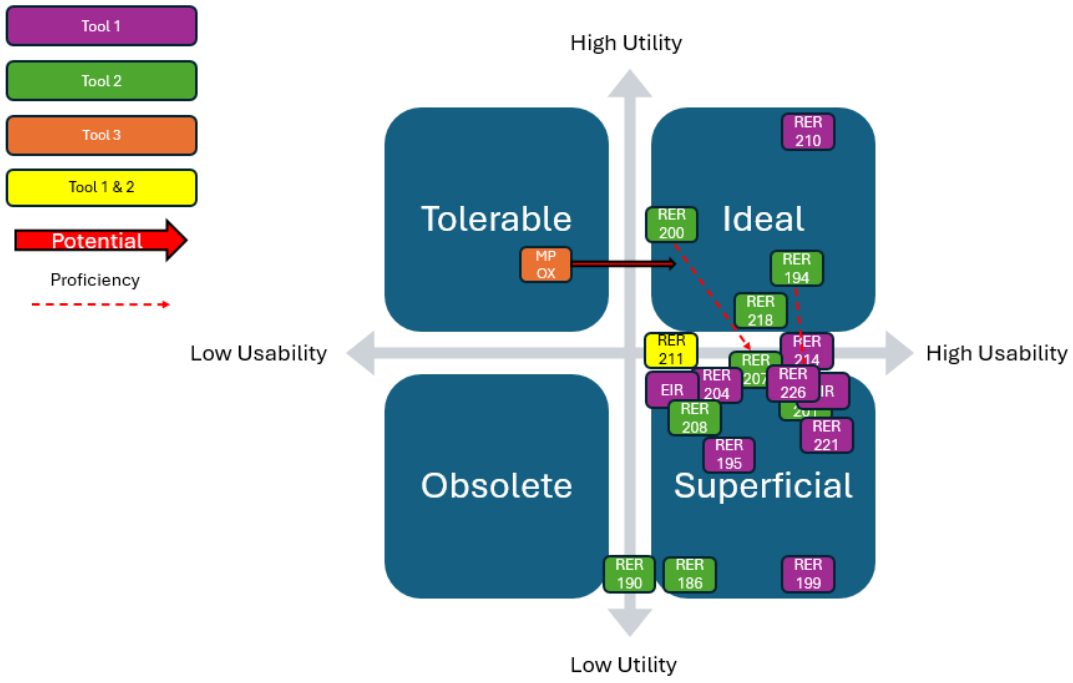
5.2 AI research tools utility and usability

Beyond broad reflections on the use of AI tools, the research team also provided a visual mapping of different products and the useability and utility of the tools piloted. To better understand researcher perspectives of AI tools, researchers were asked to reflect on two criteria:

- Useability related to how useable the tool was; i.e., ease of use.
- Utility, in the context of our pilot, this related to how useful the tool was in answering the question posed.

Each of the products involved in the pilot was mapped onto a matrix to visualise these two elements and reflect, overall, how useful particular tools were for specific products (see Figure 9).

Figure 9: Useability and Utility of AI Tools



Source: authors' illustration.

Findings suggest that:

- The AI tools were considered to be 'ideal' in responding to 21% (4 of 19) queries balancing higher levels of useability and or utility.
- The AI tools were considered to be 'tolerable' in responding to 5% (1 of 19) queries with low levels of useability but higher levels of utility.
- The AI tools were considered to be 'superficial' in responding to 58% (11 of 19) queries with lower levels of utility and or useability.
- The AI tools were considered to be 'no use' in responding to 16% (3 of 19) queries with particularly low levels of utility despite higher levels of useability.

6 Discussion

The findings of this pilot echoes the sociotechnical systems perspective articulated by Jasanoff (2004) and Bijker et al. (1987), which posits that technologies do not operate in isolation but are co-constructed with the institutional and epistemic environments in which they are embedded. For example, the limited utility of AI tools in later-stage synthesis tasks, despite their technical sophistication, highlights the importance of aligning tool use with the specific epistemic demands of policy-oriented research. This suggests that AI integration must be not only technically feasible but also epistemologically congruent.

Further to this, the mixed performance of AI tools across K4DD outputs underscores Knorr-Cetina's (1999) argument that different domains of knowledge production operate within distinct epistemic cultures. While AI tools have shown promise in health-related fields characterised by structured methodologies and standardised evidence hierarchies, their application in development and diplomacy, where evidence is often fragmented, contested and reveals a misalignment between tool assumptions and domain-specific knowledge practices.

The pilot's findings support the shift from automation to augmentation, as proposed by Shneiderman (2020) and Zou et al. (2025). Rather than replacing human researchers, AI tools function more effectively as cognitive aids, particularly in early-stage ideation and scoping. This aligns with the concept of collaborative intelligence, where human or researcher judgment remains central, and AI serves to extend rather than supplant human reasoning.

The challenges encountered when retrofitting AI tools into existing workflows suggest that effective integration requires a reconfiguration of research practices themselves. For example, the challenge of retrospectively applying an AI tool to the generation of evidence summaries that, whilst demonstrating high levels of potential, was unsuccessful given existing practice. Future research workflows may need to be co-designed with AI capabilities in mind, enabling more functional human-machine collaboration.

These findings suggest a need for AI tool developers, and proponents of their use, to engage more deeply with the epistemic and institutional logics of the domains they aim to serve. Tools designed for academic rigor may falter in policy contexts unless they are adapted to accommodate ambiguity, urgency, and contextual nuance.

The findings of this pilot study underscore the importance of evaluating AI tools not solely on their technical capabilities but within the broader sociotechnical systems in which they are deployed. While AI tools demonstrated utility in early-stage research tasks, particularly in preliminary literature searches and topic familiarisation, their performance was markedly less effective when applied to complex, multifaceted queries typical of the K4DD programme. This reflects the limitations of current AI models in navigating the epistemic demands of policy-oriented rapid evidence synthesis, where nuance, contextual sensitivity, and interdisciplinary reasoning are essential. The mixed sentiment expressed by researchers across dimensions such as quality, relevance, and verifiability suggests that AI tools are not yet fully aligned with the epistemic culture of rapid, policy-responsive research (Knorr-Cetina, 1999). These findings reinforce the argument that AI integration must be context-sensitive and guided by human researcher oversight, as emphasised in collaborative intelligence frameworks (Zou et al., 2025; Shneiderman, 2020).

Moreover, the pilot highlights the need for a shift in how AI tools are conceptualized within research workflows. Rather than standalone solutions, they should be understood as

components of a distributed cognitive system (Hollan et al., 2000), where human judgment, institutional norms, and technological affordances co-produce knowledge. The limited success of tool 3, for instance, when retrofitted into existing workflows, suggests that effective AI integration may require rethinking research design from the outset. As AI tools continue to evolve, future research should explore how workflows can be redesigned to harness their strengths while mitigating risks related to bias, opacity, and epistemic misalignment. Ultimately, approaches should not replace human researchers but, rather, foster meaningful human–AI partnerships that enhance the speed, scope, and quality of evidence synthesis in development and diplomacy contexts.

7 Conclusions, limitations and recommendations

The AI pilot set out to address three key questions. This was approached through the collection and collation of both qualitative and quantitative data via a number of methods.

Which AI tools and technologies could be of use in the production of K4DD products and how?

AI tools have potential application in the K4DD project and across the knowledge products that it develops. Our appraisal of three tools that leverage a range of technologies (principally LLMs) reached the following conclusions:

Across all tools (and their underlying technologies) the conclusion of this pilot is that whilst tools have some utility, in their current iteration, their contribution is relatively limited. The tools were ill-suited to addressing complex or multifaceted queries. There were also concerns regarding the quality of outputs and the time it took to engage with tools (these points are assessed below).

Where tools were most useful was in the initial ideation stage; i.e., to support subject immersion and familiarisation. It is also important to recognise that no single AI tool assessed could effectively cover all requirements of K4DD knowledge products.

A final consideration here, is that K4DD knowledge products have been refined over a 20-year period and have adopted an approach that meets the needs of FCDO. Whilst AI tools may currently not meet our specific needs they are rapidly evolving and will likely improve significantly in the future.

What is the impact of AI tools on quality, validity, relevance, verifiability and speed in the K4DD service?

The pilot assessed the impact of AI tools on quality, validity, relevance, verifiability and speed in the K4DD service as relatively mixed. Whilst there were examples of when AI tools were exceptionally useful, these were limited and can be considered outliers. The impact of AI tools on these areas, on balance, was not transformative.

Across all tools and across all themes the data collected from the pilot highlights a mixed experience with some positive and negative outliers. This is likely indicative of the unique nature of K4DD research which produces rapid responses to complex queries posed by FCDO. These are often on emerging, under-researched or highly specific topics. These test the bounds of what AI tools can accomplish and their ability to generate high quality, valid, relevant, verifiable information at speed.

What are the ethical and legal consequences of using AI tools in rapid reviews?

This pilot identified a range of legal and ethical challenges specific to the K4DD research project but also reflective of broader concerns regarding AI. These include, though are not limited to:

Bias and fairness: AI tools may perpetuate or amplify biases. For example, tool 2 privileged academic papers that were well cited and typically from academics based in the Global North. Tool 1 drew on information in the public domain and as such not on information that is not publicly accessible.

Research integrity: There is a risk that researchers become over reliant (and led) by AI-generated content. Researchers must critically evaluate the outputs of AI tools and ensure that researchers maintain intellectual lead.

Copyright and intellectual property: There exist risks that AI tool use (particularly the uploading of documents) infringes on copyright and intellectual property rights. This is an important consideration given indemnity clauses that are common in the terms and conditions of tools.

Regulatory uncertainty: AI tool development has outpaced the development of regulations, and this has led to uncertainty regarding the appropriate and efficient use of tools.

Limitations of the pilot: The conclusions of this paper should be considered in light of the following limitations. These include, though not limited to:

Tool selection: Whilst the pilot sought to select a representative list of AI tools to assess based on a set of criteria, these choices inevitably impacted on our results. Pragmatic considerations informed our selection of tools and different choices may have led to different findings.

Researcher differences: Whilst we attempted to standardise the process of undertaking the review, individual differences (and indeed inclinations) inevitably influenced findings. There were inevitable differences in subjective judgement when completing pro formas or providing feedback on tools.

The unique nature of K4DD: The K4DD project is a demand-responsive project that generates rapid research and our reviews typically draw on a range of evidence (academic and grey). Topics are often complex or contested which makes the use of AI tools challenging.

Scope: The pilot was executed over a short time period (six months) drew on a small sample (19 reviews) and involved a limited number of researchers (10 researchers).

Bias: The interpretation of findings is inevitably subject to the biases of researchers.

8 Recommendations

Whilst the K4DD pilot reached a relatively mixed conclusion regarding the impact of AI tools on our work, it is evident that AI tools will continue to evolve rapidly and as such researchers, research programmes and institutions will have to prepare for adoption.

Recommendations outlined below are structured across multiple levels (from the individual to the institution) and should be considered forward looking i.e., considering a likely future where AI tool use is increasingly the norm.

8.1 Recommendations for researchers

Researchers need to invest time in familiarising themselves with AI Tools. This will include an investment of time in understanding the functionalities and limitations of AI tools.

Explainer: There exists an array of AI tools that can be used to support research. These range from Generative AI chat tools such as tool 1 that are relatively accessible (though questions remain as to how to maximise their application) to complex tools such as tool 3 (systematic review tool) that have more diverse applications but can be challenging to use. For all tools, researchers will need to familiarise themselves and test application and identify those tools which best suit the research task and the preference of researchers.

Researchers need to develop new approaches to research involving a combination of AI informed and non-AI informed methods; i.e., using AI tools in conjunction with existing research methods to ensure comprehensive and accurate results.

Explainer: Research involves multiple stages from ideation and subject immersion to structured literature searches, quality assessment, writing and revision. AI tools can support different elements of this process but will need to be integrated into workflows. Similarly, existing approaches may require revision. For example, in the AI pilot we attempted to apply tool 3 to an existing approach to producing Mpox evidence summaries. This proved challenging and was ultimately unsuccessful. Retrospectively trying to add a new tool to an existing process was not possible, however designing a new approach that integrated tool 3 from the outset may have been more successful. Linked to the above, researchers will need to

consider how they reference the use of AI tools in their research, and for which stages its application is appropriate.

Researchers need to be supported in identifying and mitigating ethical and legal risks associated with using AI tools. In particular, issues pertaining to data privacy and copyright as well as epistemological primacy will need to be considered.

Explainer: When researchers apply AI tools for a specific task they must consider both the legal (copyright, data protection, confidentiality) and ethical (surveillance, discrimination, privacy) issues. These must be considered on a case-by-case basis and will require a process of critical reflection, discussion and debate. In particular, it will be important to interrogate how AI tools operate, the risks they pose and what forms of knowledge are privileged and which are ignored.

Researchers need to adopt new or adapt existing approaches to documenting the research process to record how and why AI tools have been used. This is essential to support transparency, accountability and replication of research methods.

Explainer: Researchers will need to reflect on how best to record and reference the approaches used when adopting AI tools. Within our pilot we asked researchers to catalogue application and NOT to use AI to generate text for reviews. Establishing clear processes for the application of AI tools and frameworks to support and defend decision making will be essential. This may entail an adaptation of existing approaches but also the adoption of new ones.

Researchers need to critically evaluate the outputs of AI tools across a number of areas including quality, validity, relevance, verifiability and speed.

Explainer: Within our pilot we asked researchers to critically reflect on the outputs of AI tools and report back (this was aided by the provision of additional time to engage in critical reflection). Researchers will need to apply such critical reflection when using AI tools to ensure that tools are fit for purpose. For example, it is important that the research process is led by the researcher and not by the outputs of AI tools that can be misleading or factually incorrect.

Recommendations for research programmes. Research programmes need to consider an interdisciplinary approach to recruitment, involving experts from different fields and with a variety of skills to maximise the benefits of AI tools.

Explainer: Research programmes will need to consider the capabilities and capacities of teams i.e., combining 'traditional' research skills alongside a wider range of expertise to maximise the use of AI tools. The role of data scientists, knowledge curators, coders etc. may become increasingly important in research programmes. The balance of team composition will be an important consideration.

Research programmes need to develop and adhere to guidelines for the use of AI tools, ensuring compliance with legal, ethical and institutional standards.

Explainer: A recurrent theme within the pilot was requests for guidance from researchers regarding both the efficient use of tools (i.e., training) and how to use them appropriately (i.e., guidance). As noted elsewhere, the emergence of AI tools will necessitate a review of existing programme specific and institutional guidelines. For example, data management, research integrity and ethics, data protection policies etc. and their application.

Research programmes will need to allocate sufficient resources for the acquisition and maintenance of AI tools and provision of related training. This can be costly and will need to be approached in conjunction with home institutions.

Explainer: There exists a plethora of AI tools that can be adopted (both free to use and subscription based), research programmes will need to assess which tools are most appropriate and how this will affect overall budgets (both in terms of licence cost and training needs). For example, the cost of one tool (tool 2) for 10 researchers for one year would equate to £4,500 (individual licences). Research programmes will need to assess how many licences, for what length of time and for how many tools.

Research programmes need to develop robust data management plans that include guidelines for using AI tools in data collection, analysis, and storage.

Explainer: The emergence and increasing importance of AI tools poses a number of risks for research programmes regarding potential improper or inappropriate use. There remain concerns regarding how data uploaded to tools is used to train foundational models and the extent to which uploading may problematise copyright or data use principles. Data management plans will need to be developed that factor in how AI tools will be used in relation to the collection, management, storing and sharing of data and any potential risks.

Research programmes need to establish robust quality-assurance protocols to monitor and evaluate the performance of AI tools (and of researchers using these) in research projects to ensure that use is appropriate, efficient and that legal/ethical standards are adhered to.

Explainer: To ensure the appropriate use of AI tools, research programmes will need to establish clear protocols regarding the adoption of specific tools and their use by researchers within the research project. This will have to be developed on a project-to-project basis but referencing institutional principles, guidelines, policies and frameworks. For projects involving multiple partners, discussions will need to be had during project development cycles about alignment and consistency.

Recommendations for institutions. Institutions need to develop policies that support the legal, ethical and effective use of AI tools in research.

Explainer: Institutions, in dialogue with research projects and researchers, will need to adapt existing and develop new policies to govern the use of AI in a range of tasks. These policies

should be structured around guidance and advice (decision making frameworks) that encourage adoption and experimentation but also ensure critical reflection and decision making regarding legal and ethical issues. This will imply discussion across different institutional bodies (e.g., ethics committees, data protection, legal services, procurement departments and libraries). Such policies, guidelines and decision-making frameworks will need continuous review.

Institutions need to establish and adequately finance support services, such as helpdesks, research skills training or advisory committees to assist the effective use of AI tools.

Explainer: The effective and appropriate use of tools requires support and investment. Institutions must consider how best to support research programmes and researchers to understand both the potential of AI tools but also the risks associated with their use.

Institutions need to allocate resources for the continuous evaluation, assessment and improvement of AI tools used in research.

Explainer: As noted elsewhere, institutions will need to manage budgets whilst ensuring that research programmes and researchers are supported in accessing the tools they need to fulfil their function. This will imply discussion and debate across a range of users to balance competing demands for different tools (e.g., for 100 researchers the cost of a subscription to tool 2 would be £45,000 per year, excluding training).

Institutions need to implement risk management strategies to address potential issues related to data security and intellectual property when using AI tools.

Explainer: AI tools, whilst having great potential, also pose a range of risks. Institutions will need to critically evaluate these, particularly how to manage large numbers of staff using different AI tools for different purposes and in different ways. For example, whilst our AI pilot strictly prohibited the uploading of documents to tools for a range of legal and confidentiality reasons, this was easier to manage on a small pilot. Risks increase with the number of users, particularly in contexts where confidential documents may be shared.

Institutions need to provide ongoing professional development opportunities for staff to enhance their skills in using AI tools effectively and appropriately.

Explainer: Continuous professional development will need to be supported for both research and non-research focussed staff to ensure that AI literacy is developed but also to critically empower users to assess legal and ethical risks. This pilot worked closely with experts across the University of Birmingham, IDS and University College London to provide guidance and reflect on a range of key questions regarding AI tool adoption and use.

References

Bijker, W. E., Hughes, T. P., and Pinch, T. (1987). The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology. Cambridge MA: MIT Press.

Bolaños, F. Salatino, A. Osborne, F. and Motta, E. (2024). Artificial intelligence for literature reviews: opportunities and challenges. *Artificial Intelligence Review*, 57: art. 259. https://link.springer.com/article/10.1007/s10462-024-10902-3

Booth, A., Sutton, A. and Papaioannou, D. (2021). *Systematic approaches to a successful literature review.* 2nd ed. London: Sage.

de la Torre-López, T. Ramírez, A. and Raúl Romero, J. (2024). Artificial intelligence to automate the systematic review of scientific literature. *Computing*, 105: 2171-2194 https://link.springer.com/article/10.1007/s00607-023-01181-x

Jasanoff, S. (2004). States of Knowledge: The Co-production of Science and the Social Order. London: Routledge.

Government Social Research Service. (2014). Rapid Evidence Assessment Toolkit index. Civil Service. HM Government. https://webarchive.nationalarchives.gov.uk/ukgwa/20140402164155/http:/www.civilservice.gov.uk/networks/gsr/resources-and-guidance/rapid-evidence-assessment

Gusenbauer, M. and Haddaway, N.R., (2020). What every researcher should know about searching — Clarifying the fundamentals of effective literature search. *PLOS ONE*, 15(3): p.e0230495.

Haby, M.M., Chapman, E., Clark, R., Barreto, J., Reveiz, L. and Lavis, J.N. (2016). What are the best methodologies for rapid reviews of the research evidence for evidence-informed decision making in health policy and practice: a rapid review. *Health Research Policy and Systems*, 14(1): 83.

Haman, M., and M. Školník (2024). 'Using ChatGPT to Conduct a Literature Review.' *Accountability in Research*, 31 (8): 1244–1246. https://doi.org/10.1080/08989621.2023.2185514.

Haman, M., and Školník, M. (2025). Fake no more: The redemption of ChatGPT in literature reviews. *Accountability in Research*, 1–3. https://doi.org/10.1080/08989621.2025.2465619

Hollan, J., Hutchins, E., and Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, 7(2): 174–196.

Jasanoff, S. (2004). States of Knowledge: The co-Production of Science and Social Order. New York: Routledge.

Knorr-Cetina, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge.* Cambridge MA: Harvard University Press.

Lipworth, W., Munsie, M., Kerridge, I. and Selgelid, M.J., 2023. Ethical implications of artificial intelligence in academic writing and peer review. *Accountability in Research*, 30(1): 1–18.

Marcus, G. and Davis, E., (2020). Rebooting AI: Building artificial intelligence we can trust. New York: Vintage.

Marshall, I.J. and Wallace, B.C. (2019). Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8(1): 1–10.

OECD (2023). *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research.* OECD. https://www.oecd.org/en/publications/artificial-intelligence-in-science_a8d820bd-en/full-report/elicit-language-models-as-research-tools_fec8a6ab.html#top

Ofori-Boateng, R. Aceves-Martins, M. Wiratunga, N. Francisco Moreno-Garcia, C. (2024). Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review. *Artificial Intelligence Review.* https://link.springer.com/article/10.1007/s10462-024-10844-w

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe and trustworthy. *International Journal of Human–Computer Interaction*, 36(6): 495–504. https://doi.org/10.1080/10447318.2020.1741118

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S. and Farajtabar, M., 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Smalheiser, N.R., 2017. Rediscovering Don Swanson: the past, present and future of literature-based discovery. *Journal of Data and Information Science*, 2(4): 43–64.

Tang, R., Nogueira, R., Zhang, E., and Lin, J., 2022. What makes a good summary? Reconsidering the evaluation of automatic summarization in the era of large language models. *arXiv preprint* arXiv:2209.12356.

Tricco, A.C., Langlois, E.V. and Straus, S.E. (2017). *Rapid reviews to strengthen health policy and systems: A practical guide*. Geneva: World Health Organization.

Wallace, B.C., Trikalinos, T.A., Lau, J., Brodley, C. and Schmid, C.H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, 11(1): 55.

Ye., R. Varone, M. Huang, O. Lee, P. Liut, M. and Nobre, C. (2025). The Design Space of Recent AI-assisted Research Tools for Ideation, Sensemaking, and Scientific Creativity. Unpublished preprint.

Zou, H. Huang, W. Wu, Y. Miao, C. Li, D. Liu, A. Zhou, Y. Chen, Y. Zhang, W. Li, Y. Fang, L. Jiang, R. Philip Yu, S. (2025). A Call for Collaborative Intelligence: Why Human-Agent Systems Should Precede AI Autonomy. Unpublished preprint.

Appendix A:

AI tool licence review tool (University of Birmingham

Name of Tool	Include the name of the tool.
What type of tool is it?	AI tools can have a range of different features. Detail here some information about the purpose of this tool and what its main capabilities are. Is the tool something you access via the web or something you need to download. If the item is software that you need to download you should contact your IT department for guidance as there may be security concerns involved with downloading the service.
What does it do with inputs uploaded?	Some tools allows a user to upload their own items, detail here any information about how they store those files or use the data from them.
IP warranty notes	Provide information here about any intellectual property warranty that you provide the tool/vendor. If you warrant that users will only upload items that they have the intellectual property rights to but then materials are uploaded that do not meet this requirement the vendor may be able to make a claim for damages.
Do you need to own the copyright / IP in the input?	Does the tool state that the users needs to be the owner of the copyright/intellectual property of any content that is loaded into it or used as an input? If yes, this would restrict what a user could include as an input.
Do you give a licence to the inputs?	Does the licence state that you provide them with a licence to any content that you put into the tool?
Licence to inputs notes	Put the details of any information that details the user providing licence rights to any content that they upload into the tool. A user will not have the legal rights to provide a licence to the AI tool for content where they are not the owner of the intellectual property in that content. You also need to consider data protection issues. How inputting and providing a licence may impact on commercialisation of the research.
How are outputs allowed to be shared?	Does the tool give any restrictions on how any outputs that are generated can be shared or used? Detail that information here. Will your users be able to use the tool for the purposes that they want to use it for with these restrictions?
Does it meet current WCAG standard?	The current UK Public Bodies WCAG standard is WCAG 2.2 AA. This is a legal accessibility requirement.

Who indemnifies whom?	We need to avoid licences where we, as the licensee, indemnify the licensor (the provider of the resource). By agreeing to uncapped indemnification clauses where you indemnify the licensor, you are opening up your organisation to potentially uncapped legal costs. Most of the time for normal eResources licences the licensor will indemnify us. Licences where you indemnify the provider of the resource can normally only be agreed to and signed by a limited number of senior ranking staff in an organisation. Contact your licensing or procurement teams before agreeing to licence terms where you indemnify the licensor.
Data protection	Does the licence say that it meets GDPR or UK Data Protection laws? If the answer is NO or the licence says that it will share use data with third parties, or it will be reused, or resold without permission then further advice from a data protection officer should be sought.
Final notes	List all of the key findings from your review of the terms and conditions e.g. are the following issues okay or not: Accessibility statement Indemnification clause Data Protection / GDPR Copyright / IP issues

Appendix B:

AI Pilot FCDO-K4DD Pro Forma

Name: TO BE ADDED

Output Type/Title: TO BE ADDED

On balance, was the tool useful for this review: YES/NO

	Initial negotiation and defining the question		Research process				Writing	Quality Assurance
	Preliminary Research	Query and scope clarification	Literature search	Screening	Ranking	Quality assessment	Writing	Clarity and concision
Tool #								
Tool #								
Tool #								
Reflections (Positive)								
Reflections (Negative)								
Assessment of tool #								
Quality: The extent to which the output of the AI tool meets high standards in terms of accuracy, completeness, clarity, and consistency.								
Validity: The degree to which the AI tool								

produces results that are credible and align with established or accepted findings. Validity focuses on whether the tool's results are scientifically or logically sound and based on reliable evidence.	
Relevance: The appropriateness of the AI tool's output to the specific goals or questions of the evidence review. Relevance assesses whether the information provided is directly applicable and helpful to the topic under investigation.	
Verifiability: The ability to trace and confirm the accuracy and sources of the AI-generated information. Verifiability ensures that the AI tool's outputs are transparent and can be cross-checked against primary sources or evidence.	
Speed: The efficiency and timeliness with which the AI tool processes data and delivers results. Speed measures how quickly the tool can complete tasks like evidence gathering, synthesis, and reporting compared to traditional methods.	