PLOS Climate

ESSAY

The role of artificial intelligence in climate change scientific assessments

Alaa Al Khourdajie[1,2]*

1 Department of Chemical Engineering, Imperial College London, United Kingdom, 2 International Institute for Applied System Analysis (IIASA), Austria

* alkhourdajie@imperial.ac.uk

OPEN ACCESS

Citation: Al Khourdajie A (2025) The role of artificial intelligence in climate change scientific assessments. PLOS Clim 4(9): e0000706. https://doi.org/10.1371/journal.pclm.0000706

Editor: Taoyuan Wei, Center for International Climate and Environmental Research: CICERO, NORWAY

Published: September 12, 2025

Funding: A.A.K. was supported by the Horizon Europe R&I programme projects: IAM COMPACT (grant no. 101056306) and DIAMOND (grant no. 101081179). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Climate change scientific assessments prepared by the Intergovernmental Panel on Climate Change (IPCC) face interconnected dual challenges: the exponential growth of literature, hindering synthesis efficiency, and the increasing length of its reports, impeding accessibility. Building upon the emerging discussion of adopting artificial intelligence (AI) tools in scientific assessments, this essay develops specific operational and governance frameworks to guide the IPCC's integration of these tools. It makes three distinct contributions. First, it develops a systematic framework for AI-augmented evidence synthesis, detailing how machine learning (ML) can be integrated into each stage of the assessment workflow. Second, it provides a critical analysis of Large Language Models' (LLMs) use for reports communication through the lens of 'addressable' versus 'inherent' limitations, clarifying which risks require technical solutions versus those that demand robust governance. Finally, it proposes a novel governance structure for the IPCC based on two institutional roles, the 'producer' and the 'assessor' of AI products, to ensure scientific integrity is maintained. This essay provides a clear path for the responsible, expert-led integration of AI, ensuring it serves to augment, not replace, human expertise.

Introduction

Scientific assessments play a crucial role in identifying and communicating areas of consensus within the scientific community. They inform policymakers, stakeholders, and the public with the latest scientific findings and the confidence levels assigned to them by experts. These assessments therefore promote consensus-driven, evidence-based decision-making, and help minimise misinformation by ensuring that established scientific understanding is publicly accessible. These assessments also foster informed dialogue within the scientific community itself [1–3].

Depending on the strand of the literature, several approaches are used for undertaking scientific synthesis (In this article, I use scientific synthesis, literature synthesis, and evidence synthesis interchangeably.), each aiming to maximise rigour and minimise susceptibility to bias. These approaches include systematic literature reviews (which follow a pre-defined replicable protocol to identify and evaluate all relevant research), meta-analyses (which statistically combine results from multiple studies), weight-of-evidence approaches (which assess the relative strengths and weaknesses of different lines of evidence), and expert elicitation (structured expert judgement), collectively offering more structured and robust alternatives to traditional literature reviews [4–6]. Beyond relying on literature synthesis, climate change assessments, such as those by the IPCC, also rely on data-based approaches to drive some of their findings such as assessing observational data, paleoclimate reconstructions, and ensembles of climate change and techno-economic scenarios [7–9].

The IPCC, the leading intergovernmental science-policy interface on climate change, regularly produces comprehensive assessments of the scientific basis of climate change, the vulnerability and exposure of ecosystems and humans to climate change impacts and future risks, and options for adaptation and mitigation [10]. It combines the expertise of scientists from a wide range of disciplines and policy experts worldwide to create consensus-driven assessments. The assessments employ various evidence synthesis approaches and data-based methods applied to academic and grey literature, supported by IPCC's Uncertainty Guidance Note [11] that contain calibrated uncertainty language to guide the authors in identifying their collective confidence level in synthesised findings [12,13]. These reports are drafted and reviewed iteratively by lead authors, scientists and policy experts, and review editors. They highlight established knowledge, evolving understanding, and multiple perspectives within the literature [10,14,15].

Despite these comprehensive, consensus-driven, and uncertainty-calibrated production processes, the exponential growth and mounting complexity of scientific literature [16–20] challenge the IPCC's ability to fulfil its mandate of conducting assessments that are 'comprehensive, objective, open and transparent,' as required by its governing principles [15]. Evidence shows that the ratio of studies cited in IPCC reports relative to the total number of climate change studies indexed in the Web of Science (WoS) has declined from 60% in its first assessment report (AR1) to around 15% in AR6 [16]. While this trend may partially reflect a discerning focus by the IPCC authors on higher-quality publications, the sheer scale of the literature still presents a significant risk that critical findings might be insufficiently weighted or overlooked altogether. These concerns are echoed in other criticisms facing the IPCC, such as literature diversity gaps [8,21], underrepresentation of alternative knowledge systems [22], concerns about methodological transparency [23,24] and production speed [25], issues which are arguably intensified by the expanding scale of scientific literature.

IPCC reports are also considerably increasing in length, reflecting both the expanding body of scientific literature but also the comprehensive scope set in their initial outlines. Each assessment cycle (typically 5–7 years) produces at least three main reports and a synthesis report, often accompanied by methodological and special reports. In the previous cycle (AR6), the main reports grew substantially, with two exceeding 2,000 pages and one exceeding 3,000. Although these reports include summaries for policymakers and technical practitioners and are designed to be consulted selectively (by chapter) rather than read in full, navigating such extensive documents to locate specific evidence remains challenging. Key findings, particularly those neither highlighted in the summaries nor located in their seemingly “obvious” chapters, can be difficult to find and extract. This difficulty in accessing specific information within the reports contributes to the broader challenge of effectively communicating their extensive findings while preserving scientific nuance, complexity, and the agreed-upon levels of confidence and uncertainty.

These interconnected challenges, managing exponential literature growth and communicating extensive findings, have prompted exploration of artificial intelligence as a potential solution. With the IPCC seventh assessment cycle (AR7) now underway, the challenge of assessing the increasing volume of literature is being actively considered, with planning underway for a workshop titled ‘Methods of Assessment’ [26] that will focus on the use of AI and ML methods for IPCC assessments. One of the workshop’s proposed aims is to ‘evaluate whether AI could be integrated into IPCC processes, and if so, under what conditions and with what safeguards’ [27]. Meanwhile, the emergence of LLMs, sparked by making ChatGPT publicly available late 2022, has brought the use of AI in scientific processes to the forefront of the wider scientific community [28,29] including science communication [30]. For instance, the International Energy Agency recently launched a chatbot tool that enables users to interrogate their flagship World Energy Outlook 2024 report through natural language queries [31], complementing other emerging climate-focused chatbots based, in parts, on IPCC reports, such as ChatClimate [32] and ClimateQA [33].

To that end, in this essay I explore how, and to what extent, existing and emerging AI tools can mitigate these interconnected challenges of managing the exponential growth of literature for robust evidence synthesis and consensus building as well as effectively communicating extensive and nuanced findings. Recent developments and increasing adoption of AI tools raise important questions: How can AI tools augment human expertise in synthesising and assessing the scientific literature? What would an ‘expert-in-the-loop’ AI-augmented processes for climate change assessments look like? How can such processes leverage the strengths of both technology and expert judgement, and mitigate against potential weaknesses and arising biases? Moreover, what are the potential benefits and limitations of using AI-assisted tools in climate science communication?

Addressing these questions requires examining several interconnected dimensions. Recent commentaries have effectively outlined the broad opportunities in this area [28], while new methodological guidance urges a shift from appraising primary literature to formal ‘knowledge syntheses’ to manage the information deluge [16]. This essay bridges these conversations by providing a specific operational framework for creating and governing the AI-augmented syntheses that this new assessment landscape demands.

To do so, it makes three distinct contributions. First, it develops a systematic framework for AI-augmented evidence synthesis, detailing how specific ML practices can be integrated into each distinct stage of the assessment workflow and codified in a detailed taxonomy (Table 1) (Section 2). Second, it provides a forward-looking critique of LLMs' use for science communication, analysing their risks through the lens of ‘addressable’ versus ‘inherent’ limitations (Section 3). Finally, it proposes potential hybrid, expert-judgement-led approaches grounded in two alternative options: one where the IPCC acts as a ‘producer’ by developing AI *tools* internally, and another where it acts as a critical ‘assessor’ of externally published AI-driven *products* - i.e. analyses (Section 4). This final section, therefore, explores the overarching governance required to maintain scientific integrity for both the synthesis process and for science communication, balancing innovation with expert judgement in an ‘expert-in-the-loop’ system.

AI tools for literature assessment: capabilities and limitations

The exponential growth and mounting complexity of scientific literature challenges the efficiency and comprehensiveness of scientific assessments. Continuing to follow traditional evidence synthesis approaches is becoming untenable for processing this expanding knowledge base, potentially undermining the robustness of consensus-building processes.

This trend has driven increasing interest in using AI tools to augment literature synthesis workflows [16–20]. Consistent with the IPCC's mandate to assess existing literature, the potential integration of AI can be viewed through two primary lenses, explored further in Section 4: firstly, the internal application of validated AI *tools* to generate specific products (like evidence maps or screened literature lists) that support the assessment process; and secondly, the critical assessment of already AI-generated *products* published in the scientific literature itself. The systematic framework for AI-augmented evidence synthesis presented in this section, and codified in a detailed taxonomy in Table 1, is foundational to both options, providing an operational guide for the IPCC whether it acts as an internal 'producer' of AI tools or a critical 'assessor' of externally published analyses.

While specific implementations vary across applications, the ML-augmented literature synthesis workflow generally mirrors the stages of traditional approaches. However, it strategically integrates ML tools to enhance efficiency, scalability, and broaden the scope at each stage. These stages typically encompass question formulation, literature query and collection, documents screening and selection, knowledge mapping and analysis, and reporting and visualisation. This workflow represents a practical application of the broader knowledge synthesis methodologies that are becoming essential for climate change assessments [16]. Moreover, the specifics of ML tool integration may differ depending on relevant research considerations, such as a particular level of spatial or temporal resolution necessitating different techniques. While *question formulation* remains primarily expert-led, augmenting subsequent stages in the workflow with ML tools can inform researchers about literature breadth and coverage, thereby enabling the identification of knowledge gaps that warrant further research.

The initial stage of *literature query and collection* often still relies heavily on conventional search methods. Keywords, Boolean logic operators (i.e., 'AND', 'OR'), and wildcards are typically used to query established bibliographic databases like Web of Science or OpenAlex [19,34,35]. While providing a replicable baseline, these methods can struggle with diverse or non-standard terminology, creating a risk of overlooking relevant literature. To enhance comprehensiveness, more sophisticated, expert-driven query strategies were tested. Examples include iterative multi-step searching protocols that use keyword analysis and synonyms from key papers to refine subsequent searches [36], or approaches combining top-down searches (guided by assessment outlines) with subject-specific expert queries to broaden retrieval [37]. However, such iterative refinements, while valuable for increasing sensitivity, may sometimes retrieve a high proportion of irrelevant documents, requiring careful screening [38]. Distinct from refinements of conventional methods, recent advances in AI offer alternative approaches. Semantic search, leveraging natural language processing (NLP), aims to overcome keyword limitations by discerning the meaning of search terms, potentially identifying relevant papers missed by traditional queries. Examples such as Elicit or SciSpace platforms utilise capabilities like semantic search expanding search scope beyond initial keywords [39].

The *screening and selection stage* is crucial for determining which articles are relevant to the research question, based on clearly defined inclusion and exclusion criteria. Increasingly, this stage incorporates ML text relevance classifiers to support and streamline decision-making. A foundational approach involves blending expert judgement with ML, as illustrated by Callaghan and colleagues [19] in climate impact attribution studies. In this example, the experts manually label a sample of documents, allowing an ML algorithm to learn relevant patterns. Additionally, continuous cross-validation helps ensuring robust performance on unseen data. Active learning techniques can also be employed, as demonstrated by [34,40] for climate change and health literature. In this approach, the classifier iteratively refines its own predictions based on newly screened data, progressively improving its ability to prioritise papers for manual review. ML applications in screening also extend beyond simple relevance classification. For instance, Lamb and colleagues [35] used the Geonames database to classify geographical contextual information, specifically assigning urban mitigation case studies to typologies, thereby revealing geographic research gaps. Further advancements are demonstrated in recent studies by Sietsma and colleagues [41] and Callaghan and colleagues [42] who utilise powerful transformer-based language models like ClimateBERT (ClimateBERT is a transformer-based language model specifically pre-trained on climate-related texts to enhance performance in understanding and classifying climate discourse.) and SciNCL (SciNCL (Neighbourhood Contrastive Learning) is a transformer-based model that integrates citation network data via contrastive learning, improving its ability to capture the nuanced context of scientific literature. In the context of literature assessment, such transformer-based models (ClimateBERT and SciNCL) are primarily employed for discriminative tasks like classification or representation learning, rather than generative tasks such as creating new text.). Because these models are pre-trained on extensive climate-related literature, they offer enhanced capabilities in detecting relevant textual nuances and improving the classification of documents.

The stage of *knowledge mapping and analysis* is increasingly augmented by advanced ML methods used to extract latent structures from large literature collections. One key technique is topic modelling, an unsupervised ML approach that identifies clusters of co-occurring words, thereby revealing the underlying semantic structure of a literature corpus. Complementary to topic modelling, network analysis is deployed to identify citation and co-authorship networks, thereby pinpointing influential publications and clustering research communities [37]. These computational techniques operate on different dimensions of the literature: topic models extract thematic content, while network analyses reveal relational structures between papers and authors. Both approaches help researchers navigate complexity and identify key themes or relationships within the evidence base, facilitating expert exploration rather than replacing expert interpretation.

The *reporting and visualisation* stage translates these analytical outputs into intuitive, accessible formats that guide further investigation. The products of ML-driven analysis include interactive topic maps, citation networks, and geospatial visualisations that display the structure of literature discovered through computational techniques. For example, Callaghan and colleagues [38] applied topic modelling to produce a topographic map of climate change literature, highlighting thematic clusters and disciplinary trends within IPCC reports. Similarly, Lamb and colleagues [35] used these approaches on approximately 4,000 urban climate case studies, generating visualisations that revealed thematic clusters by sector and exposed geographic research gaps. Creutzig and colleagues [37] coupled narrative analysis with visual mapping to illustrate thematic contours of demand-side mitigation literature. More dynamically, Callaghan and colleagues [42] developed a ‘living systematic map’ of climate policy research that continuously updates visual representations of policy instruments and sectoral trends. These visualisation products serve as exploratory tools within the assessment workflow, aiding researchers in navigating complex literature landscapes prior to expert synthesis.

Beyond the tools integrated into formal synthesis workflows, a range of other computational tools, primarily commercial offerings emerging since the advent of ChatGPT, now facilitate broader literature exploration and summarisation by leveraging recent advances in NLP. These tools can be broadly divided into three categories. First, academic search and visualisation platforms (e.g., LitMaps website) identify networks of interconnected papers based on authorship and citation relationships, drawing on data from bibliographic databases like arXiv and Web of Science. Second, query-based summarisation tools (e.g., SciSpace, Elicit platforms) deliver summaries of research findings tailored to user queries, often operating on abstracts. Third, publisher-integrated solutions (e.g., Scopus AI, Semantic Scholar’s TLDR – Too Long Didn’t Read – feature) embed AI features directly into their databases. More recently, conversational generative AI platforms like Perplexity, ChatGPT and Gemini have introduced ‘Deep Research’ functionalities that attempt to synthesise information from multiple online sources, often providing citations to overcome the common limitation where many other tools operate primarily on metadata or abstracts.

Despite these computational tools’ potential for reducing manual workload and expanding search scope beyond initial keywords, they share fundamental limitations. They currently lack the capacity to critically appraise study quality or methodological rigour, nor can they adequately interpret or weigh conflicting findings within the broader scientific context—steps crucial to the expert judgement process mandated by the IPCC Uncertainty Guidance Note. Outputs from such tools require careful expert validation to ensure accuracy and comprehensiveness [39], as summaries may oversimplify complex findings or omit crucial context. Even tools providing citations can sometimes generate inaccuracies or misrepresent sources [43]—a phenomenon often referred to as ‘hallucination’ (Check Text Box 1 for further discussion). Consequently, whilst these tools may streamline literature navigation and initial topic investigation, thorough engagement with primary literature by researchers remains indispensable for robust understanding and synthesis [44].

AI tools offer considerable advantages in the context of scientific assessments. Notably, they can markedly improve efficiency by reducing the time and resources required for literature searching, screening, and visualisation, and potentially enabling continuous monitoring approaches [19,35]. By processing significantly larger volumes of literature than traditional methods, these tools can expand the scope of evidence synthesis and decrease the likelihood of overlooking relevant studies. Furthermore, the use of consistent, algorithm-driven criteria during screening may help reduce certain types of selection bias (e.g., potentially reducing variability introduced by different human screeners applying criteria slightly differently), while the clear documentation of computational methods can enhance transparency and replicability across assessments.

However, these benefits are counterbalanced by significant limitations and risks. AI tools are highly data-dependent; the quality and representativeness of the training data (i.e. WoS, OpenAlex) fundamentally determine their performance. Consequently, biases inherent in the underlying datasets – such as the underrepresentation of certain regions, languages, or research areas – can be inadvertently amplified [37], potentially worsening existing literature diversity gaps. Moreover, while modern NLP techniques excel in pattern recognition, they often lack the nuanced contextual understanding and critical appraisal skills of human experts, particularly when assessing methodological rigour or synthesising conflicting findings [45]. Equity issues also arise, as access to advanced AI tools and the required computational infrastructure is not universal, potentially disadvantaging researchers in resource-constrained settings and exacerbating existing structural inequities within the global scientific community [12]. Additionally, many current systems predominantly process text and struggle with non-textual data such as figures. It is also crucial to recognise that AI tools do not inherently address broader systemic issues such as gender imbalances within research fields or assessment teams [46].

These considerations underline the necessity of a hybrid approach, where AI *augments rather than replaces* expert judgement. The optimal model likely lies on a spectrum: while fully automated systems might maximise coverage and speed for certain tasks, expert-guided systems – where domain specialists define parameters, validate outputs, critically interpret results, and perform the core synthesis – are essential for ensuring analytical depth, contextual accuracy, and overall assessment integrity. Potential frameworks for implementing such hybrid approaches within the IPCC assessment process are explored further in Section 4 below. Table 1 below provides a detailed taxonomy of the AI approaches

Table 1. AI augmentation in evidence synthesis workflows.

Workflow Stage	AI Augmentation Examples	Key Opportunities for Assessment Teams	Key Challenges	Expert Role	Illustrative References
Query and collection	Semantic search; citation network analysis	Broader literature discovery; Identifying influential research clusters	Potential bias, missing niche terms	Define scope, refine strategy, validate relevance	[36,37,39]
Screening and selection	ML text classification; active learning	Increased efficiency; Handling large volumes; Prioritisation assistance	Dependency on training data, nuance loss, potential bias	Define criteria, validate classifications, handle ambiguity	[19,34,35,41]
Knowledge mapping and analysis	Topic modelling; network analysis	Revealing latent themes/trends; Mapping research landscapes/communities	Requires interpretation, cannot replace critical appraisal/synthesis	Guide analysis, interpret patterns, perform synthesis	[37,38]
Reporting and visualisation	Interactive evidence maps; living maps and databases	Enhanced exploration; Dynamic updates; Aiding interpretation	Potential for misinterpretation	Select methods, interpret visuals, ensure clear communication within assessment team	[35,42]
(Cross-cutting)	Exploratory tools ("summarisers", platforms)	Quick overviews; Initial navigation	Superficiality, inaccuracy/hallucination, requires validation	Critical evaluation, use as a starting point only and rely on primary sources	Examples include: LitMaps; Elicit; SciSpace; and 'deep research' using generative AI platforms (e.g., Perplexity, Gemini)

https://doi.org/10.1371/journal.pclm.0000706.t001

discussed thus far in this section, their potential applications within the workflow stages, key limitations, and the critical role of expert judgement.

Large language models in science communication: opportunities and risks

The advent of large language models (LLMs), demonstrated by widely accessible chatbots, has generated significant interest in their potential application for communicating complex scientific knowledge, including the extensive findings within IPCC assessment reports [18,28]. However, their deployment against such authoritative science requires a rigorous critique, particularly from the perspective of the non-expert end-user, which is the focus of this section. To provide a clear analytical framework for this critique, a distinction is made between 'addressable' limitations (technical flaws that may be mitigated with better engineering) and 'inherent' limitations (problems fundamental to current architectures that demand robust governance and user literacy). Understanding this framework is essential for evaluating three core risks users likely to face when interpreting IPCC content via LLMs, including the erosion of calibrated uncertainty, the propagation of factual inaccuracies, and the amplification of systemic biases. The section concludes by considering emerging advancements through this same framework.

A significant concern is the erosion of nuance and calibrated uncertainty. IPCC assessments meticulously employ calibrated language to articulate confidence levels and likelihood statements, forming an integral part of the scientific finding [11]. LLMs, operating fundamentally as probabilistic text generators optimising for linguistic coherence based on training data patterns, may fail to preserve this essential metainformation. Standard LLM outputs may present simplified or paraphrased statements stripped of the crucial qualifiers regarding scientific certainty, hence resulting in potential misinterpretations of the assessment's basis and rigour [47,48]. This risk is not merely theoretical; recent experiments have shown that even specialised chatbots struggle to correctly interpret nuanced scientific concepts like 'acceleration' directly from IPCC texts [28]. This difficulty in preserving scientific meaning represents a core, inherent limitation of current generative architectures.

Equally concerning are factual inaccuracies and 'hallucinations.' LLMs construct responses by predicting statistically probable 'token' (The basic unit of text processing in LLMs, typically representing parts of words, or whole words. For example, the word "unprecedented" might be split into multiple tokens. LLMs process text by converting it into these discrete tokens which are then represented as numerical vectors.) sequences, not by querying a verified knowledge base or engaging in logical reasoning. This operational paradigm can result in outputs containing factual errors or 'hallucinations' – statements that appear plausible and coherent but are incorrect or lack any factual basis [49,50]. When prompted on specifics within IPCC reports, LLMs might generate inaccurate claims, potentially misleading users who assume fidelity to the source material (i.e. IPCC reports). The inherent probabilistic and pattern-completion nature of current generative models makes eliminating such errors entirely a significant technical challenge (see Text Box 1). This is where the distinction between limitations is key: while simple factual errors may be an addressable limitation that can be mitigated by grounding techniques like Retrieval-Augmented Generation (RAG) that ground LLM responses in specific, provided source documents (e.g., IPCC report PDFs), the fundamental tendency to generate plausible falsehoods is an inherent limitation of systems that lack true factual reasoning.

A third critical issue involves the propagation and amplification of biases. LLMs are trained on vast amounts of text data, inevitably inheriting societal biases related to geography, language, gender, or perspective embedded within that data. While specific impacts on interpreting IPCC reports require further investigation, these inherent biases could subtly influence the selection, framing, or emphasis of information presented to users, potentially leading to skewed understanding or reinforcing existing inequities [51]. Transparency regarding training data and methods for bias detection and mitigation remains crucial. While technical mitigation strategies can make this an addressable issue to some extent, the complete elimination of bias is likely impossible, making it a persistent and inherent challenge that requires non-technical solutions like expert oversight and critical evaluation of outputs [28].

While these risks are significant, strategies are being deployed to mitigate them. One example is RAG architectures that shift the reliance from the model's general parametric knowledge to source documents improving factual consistency [52]. Implementations like ChatClimate.ai demonstrate this approach, using IPCC AR6 as the primary knowledge source and attempting to preserve source references and confidence levels [32]. Such systems typically involve sophisticated prompt engineering (The practice of carefully crafting input instructions (prompts) to guide LLM outputs toward desired formats, styles, or content constraints. This involves techniques like specifying roles, providing examples, establishing constraints, or including specific instructions about how to handle certain types of information.) to constrain the LLM's output behaviour and may incorporate automated fact-checking routines. Nonetheless, even with RAG and careful prompt engineering, the inherent limitations of the underlying generative models persist [48].

Although the field is rapidly advancing, each technological frontier introduces specific and complex challenges. Reasoning-focused models, for instance, are designed to perform explicit stepwise analysis, which could theoretically improve the handling of calibrated uncertainty. However, research reveals this stated reasoning can be 'unfaithful': the model may construct a plausible post-hoc rationalisation for a correct answer that it actually reached via a hidden, unreliable cue. This lack of a verifiable reasoning process means the model cannot be trusted to be correct on subsequent, slightly different questions [53,54]. Similarly, the development of multimodal models represents a significant capability extension, offering the potential to interpret figures and tables. Yet, their capacity for deep semantic interpretation of complex scientific visuals is limited [55]. Even advanced retrieval architectures like GraphRAG, which use knowledge graphs to address 'global' sense-making queries across an entire corpus where standard RAG fails, are critically dependent on the quality of that underlying knowledge graph (As described by [56]), a knowledge graph in this context is a structure built by an LLM where the extracted entities from the source text become the 'nodes' and the identified relationships between them become the 'edges' of the graph.) [56]. Therefore, while these advancements demonstrate progress on seemingly addressable limitations, they also reinforce the inherent challenge of substituting algorithmic processes for genuine expert comprehension.

Consequently, navigating IPCC information via LLM tools demands a high degree of critical literacy from the user. Verification against original IPCC source documents, particularly the Summaries for Policymakers and Technical Summaries is crucial. Users should exercise scepticism towards outputs lacking precise citations or calibrated uncertainty language. Developers constructing tools based on IPCC content should prioritise verbatim extraction for key findings, preserve associated uncertainty qualifiers, provide clear source attribution, and explicitly communicate the tool's limitations [32].

Text Box 1: Understanding large language model limitations

To understand the root causes of the risks discussed in this section, it is necessary to examine the fundamental operational principles of LLMs. This text box breaks down the key technical limitations that give rise to these user-facing challenges.

- Probabilistic pattern generation: A primary driver of both factual inaccuracies and the erosion of scientific nuance, this principle means that LLMs function by predicting the most statistically likely sequence of tokens (words or sub-words) given an input prompt and the preceding generated tokens. They leverage complex neural network architectures, typically Transformers [57], to model intricate statistical patterns of language learned from massive training datasets. Their goal is linguistic coherence based on these patterns, not factual verification or logical deduction. For example, when summarising a scientific debate with a majority and minority view, an LLM might incorrectly present the majority view as a unanimous consensus because it optimises for the most probable linguistic pattern.
- Stochasticity and variability: A fundamental challenge to scientific reproducibility, the token prediction process often involves sampling from a probability distribution over possible next token. This inherent stochasticity, controllable via parameters like 'temperature', means identical prompts can yield different outputs across separate queries [58],

challenging reproducibility. This poses a significant challenge for scientific work; for instance, two policymakers asking the same question about the IPCC's findings on permafrost thaw could receive slightly different summaries, leading to inconsistent conclusions.

- Parametric knowledge limitations: A key source of both outdated information and embedded biases, an LLM's 'knowledge' is implicitly encoded within its model parameters (weights) derived from its training data. This knowledge is static (unless retrained), potentially outdated, and may contain biases or inaccuracies present in the underlying sources. It cannot actively query external sources or update its knowledge post-training unless specifically designed with mechanisms like RAG.
- Hallucinations as artefacts: The most direct cause of plausible-sounding misinformation, the generation of plausible but false or nonsensical information arise naturally from the generative process. When faced with ambiguity, insufficient relevant patterns in its training data, or prompts requiring information beyond its scope, an LLM may generate text that maintains linguistic flow but deviates from factual accuracy [49]. This is not intentional deception but an artefact of optimising for probable sequences. A common failure mode is the invention of sources; an LLM asked for evidence on a niche topic might generate a reference to a non-existent but plausible-sounding publication.

Addressable versus inherent limitations

Certain addressable limitations can be mitigated: Retrieval-Augmented Generation (RAG) grounds responses in specific documents [52]; more advanced architectures like GraphRAG aim to improve this with better contextual understanding; Reinforcement Learning from Human Feedback (RLHF) aligns outputs with human preferences for helpfulness and harmlessness [59]; careful prompt engineering guides LLMs output. Furthermore, the field is developing a sophisticated evaluation ecosystem to systematically benchmark LLM performance. This includes holistic multi-dimensional frameworks like HELM, contamination-resistant tests (Benchmarks designed to minimise the risk that a model has already been trained on the test questions, a phenomenon known as 'data contamination' which can artificially inflate performance scores.) such as LiveBench, and large-scale human preference platforms like Chatbot Arena [60–62]. A prominent technique is the use of another powerful LLM as an automated evaluator, the 'LLM-as-a-judge' approach, to score complex outputs. Although efficient, this method is susceptible to significant biases, such as a preference for verbosity or answer position, and it raises broader concerns about reliability and leaderboard stability, reinforcing the need for transparent application and critical human oversight [63,64]. However, the core inherent limitation – the lack of genuine comprehension, reasoning, and causal understanding – persist even in advanced models. LLMs manipulate linguistic form based on statistical correlations, they do not understand semantic meaning or scientific principles in a human sense. Critical evaluation by human users remains crucial.

Expert-led integration: Responsible Options for AI in Scientific Assessments

In the preceding sections, I have established a basis for the responsible integration of AI into IPCC assessments. Section 2 presented a systematic framework for using AI to manage the evidence base for robust synthesis, and Section 3 provided a critical analysis of the communication risks posed by LLMs, structured around the distinction between 'addressable' and 'inherent' limitations. Building upon these insights, and mindful that the IPCC's core role is to assess the state of scientific knowledge rather than generate new primary research, this concluding section proposes a governance framework based on practical options for responsible and effective integration of AI capabilities into current and future assessment cycles, such as the IPCC's AR7. The key question is not whether AI can contribute to the assessment process, but rather how it can best augment human expertise while maintaining scientific integrity.

Two complementary options emerge for how the IPCC could leverage AI. These can be understood as two distinct institutional roles: the IPCC acting as a 'producer' by employing validated AI tools to process and analyse the assessed literature internally (Option 1), or acting as a critical 'assessor' of AI-generated products, such as topic maps, that have already been published within the peer-reviewed literature (Option 2). The production and evaluation of such AI-driven products align with the call by Ford et al. [16] for assessment bodies to focus more on appraising knowledge syntheses rather than primary literature. Their recommendations offer valuable guidance for both options: for the internal development of robust tools (option 1) and for author teams needing to judge the quality of externally published AI-driven analyses (option 2). Integrating these rapidly evolving technologies effectively, however, requires more than just technical capability; it demands understanding their strengths and limitations and, crucially, careful consideration of how they align with IPCC principles mandating assessments be 'comprehensive, objective, open and transparent' [15].

The first option positions the IPCC as a 'producer' (potentially via its Technical Support Units or specific author teams; [28]), directly employing AI *tools* that have been previously validated and documented in the scientific literature. For instance, validated ML classifiers could assist in literature screening, or established topic modelling algorithms could generate evidence maps. This approach would require the operationalisation of the systematic workflow framework detailed in Section 2 (Table 1), supported by clear internal protocols for tool selection (ensuring they are state-of-the-art, validated, and appropriate for the task), operation (defining inputs, parameters, and execution), and the evaluation of their products (e.g., screened lists, maps). Opportunities exist to embed such AI-generated products at various stages, creating a hybrid, expert-guided workflow (visualised conceptually in Fig 1). For example, during scoping, AI-generated evidence maps [19,35] could offer rapid literature overviews. Topic modelling outputs could assist Lead Authors in structuring drafting assignments. Within drafting, interactive evidence map products [42] or filtered literature lists (generated using validated screening tools, with human oversight) could aid exploration and analysis, always stopping short of automated synthesis.

The second option positions the IPCC as an 'assessor' of AI-generated *products* (e.g., large-scale literature syntheses, complex data visualisations, model-output analyses produced using AI) that are already published in the peer-reviewed literature, treating them like any other scientific publication. This aligns directly with the IPCC's mandate but requires authors to possess the necessary expertise to critically evaluate the AI methodologies employed in those publications, including their assumptions, limitations, potential biases, and the robustness of their findings. Guidance for author teams on how to conduct such critical appraisals of knowledge syntheses is now beginning to emerge [16]. The assessment would need to consider the transparency and reproducibility of the published AI methods alongside their findings.

Regardless of the option chosen – whether employing validated *tools* internally or assessing published AI *products* – careful consideration within the IPCC's procedural and assessment frameworks is essential, particularly concerning

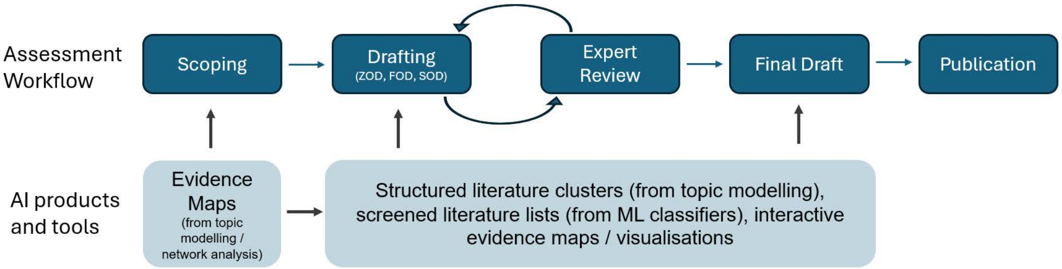
Assessment Workflow
Scoping
Drafting
(ZOD, FOD, SOD)
Expert Review
Final Draft
Publication
AI products and tools
Evidence Maps
(from topic modelling / network analysis)
Structured literature clusters (from topic modelling), screened literature lists (from ML classifiers), interactive evidence maps / visualisations

Fig 1. Conceptual integration of AI-generated products into the IPCC assessment workflow. Fig 1 illustrates potential points of interface where AI-generated products (bottom row) could support the established IPCC assessment workflow (top row). The top row depicts key stages from Scoping through Drafting to Publication. The bottom row shows examples of AI products relevant to specific stages, such as evidence maps supporting Scoping and Drafting, or screened literature lists aiding early drafting. These products could be generated either through the internal application of validated AI tools by the IPCC (Option 1) or potentially adapted from published, peer-reviewed AI analyses (Option 2), always serving as inputs for expert judgement rather than replacing it.

https://doi.org/10.1371/journal.pclm.0000706.g001

the Uncertainty Guidance Note [11]. The Guidance Note bases confidence levels on evaluations of evidence (type, amount, quality, consistency) and the degree of agreement. The key questions remain pertinent: How should author teams evaluate the 'comprehensiveness' or potential biases introduced by AI tools or reflected in published AI products when assessing the overall 'amount' and 'consistency' of evidence? How does the methodology and validation behind an AI tool (Option 1) or a published AI product (Option 2) influence expert judgement regarding the 'quality' of the evidence base? How might AI-driven structuring of information (e.g., topic maps) influence the dynamics of achieving expert 'agreement'? Addressing these questions ensures that AI integration upholds the integrity of IPCC confidence assessments.

Successfully implementing either option requires robust governance frameworks and safeguards. Furthermore, the urgency for such frameworks is heightened by the likelihood that many scientists are already using publicly available AI tools in an ad-hoc manner. For Option 1 (the IPCC as 'producer'), clear protocols defining appropriate use-cases, tool validation/selection criteria, operational transparency, product evaluation standards, and documentation are essential for transparency, like any line of evidence used in IPCC assessments [23]. For Option 2 (the IPCC as 'assessor'), guidance may be needed for authors on critically appraising AI methodologies within the literature. For both, boundaries between automated support and *indispensable* expert judgement – particularly in appraisal, synthesis, and confidence assessment – must be explicit. A precautionary, adaptable approach is also warranted, given how rapidly these technologies are evolving. Critically, equity considerations remain central. Ensuring equitable access to necessary AI tools and the skills to use / evaluate them (Option 1) or the expertise to critically assess published AI products (Option 2), alongside data and computational resources, is vital to avoid exacerbating disparities [12].

Beyond the potential integration of AI tools within the evidence synthesis workflow, the advent of LLMs introduces distinct considerations for how assessment findings are communicated and interpreted externally. As elaborated in Section 3, the capabilities of LLMs to process and generate text present opportunities for enhancing the accessibility of complex reports, they also carry substantial risks, from omitting critical nuances to generating factual inaccuracies. A constructive path forward, therefore, requires a dual approach that maps directly onto the distinction between addressable and inherent limitations. The responsibility for mitigating addressable limitations lies primarily with developers. Establishing clear best-practice guidelines for third-party developers, a recommendation emphasised by Muccione and colleagues, [28], is crucial. Such guidelines should advocate for transparency regarding model limitations, the use of robust RAG architectures strictly grounded in IPCC source texts, and clear source attribution. Conversely, navigating the persistent inherent limitations requires empowering users. A crucial element involves promoting critical literacy among diverse user groups regarding the fundamental nature of these technologies (See Text Box 1). Users must be encouraged to approach LLM-generated summaries with caution, to cross-reference information with authoritative IPCC source documents, and to critically assess the fidelity of outputs. This dual focus on developer accountability for the solvable problems and user empowerment for the fundamental ones highlights the irreplaceable role of critical human interpretation.

In conclusion, integrating artificial intelligence into the demanding processes of the IPCC presents a significant opportunity but also a substantial challenge that focusing on specific, operational frameworks. This essay has argued that enhancing evidence synthesis requires a systematic, workflow-based approach to integrating AI tools. Simultaneously, the responsible use of generative AI for science communication necessitates a governance model built on a clear understanding of the technology's 'addressable' versus 'inherent' limitations. To implement these insights, this essay proposes a governance structure based on two institutional roles, the IPCC as an internal 'producer' or as a critical 'assessor' of AI-driven products. Successfully navigating this evolving landscape demands a precautionary yet adaptable strategy, consistently prioritising scientific integrity, objectivity, and transparency. Ultimately, the goal is not to replace indispensable expert judgement but to augment human capabilities responsibly, ensuring that AI serves to strengthen the credibility and impact of scientific assessments in guiding informed decision-making.

References

1. van der Hel S, Biermann F. The authority of science in sustainability governance: A structured comparison of six science institutions engaged with the Sustainable Development Goals. Environmental Science & Policy. 2017;77:211–20. https://doi.org/10.1016/j.envsci.2017.03.008
2. van der Linden S, Leiserowitz A, Rosenthal S, Maibach E. Inoculating the Public against Misinformation about Climate Change. Glob Chall. 2017;1(2):1600008. https://doi.org/10.1002/gch2.201600008 PMID: 31565263
3. Kowarsch M, Garard J, Riousset P, Lenzi D, Dorsch MJ, Knopf B, et al. Scientific assessments to facilitate deliberative policy learning. Palgrave Commun. 2016;2(1). https://doi.org/10.1057/palcomms.2016.92
4. Haddaway NR, Bethel A, Dicks LV, Koricheva J, Macura B, Petrokofsky G, et al. Eight problems with literature reviews and how to fix them. Nat Ecol Evol. 2020a;4(12):1582–9. https://doi.org/10.1038/s41559-020-01295-x PMID: 33046871
5. Suter G, Nichols J, Lavoie E, Cormier S. Systematic Review and Weight of Evidence Are Integral to Ecological and Human Health Assessments: They Need an Integrated Framework. Integr Environ Assess Manag. 2020;16(5):718–28. https://doi.org/10.1002/ieam.4271 PMID: 32196925
6. Mach KJ, Mastrandrea MD, Bilir TE, Field CB. Understanding and responding to danger from climate change: the role of key risks in the IPCC AR5. Climatic Change. 2016;136(3–4):427–44. https://doi.org/10.1007/s10584-016-1645-x
7. Pirani A, Fuglestvedt JS, Byers E, O'Neill B, Riahi K, Lee J-Y, et al. Scenarios in IPCC assessments: lessons from AR6 and opportunities for AR7. npj Clim Action. 2024;3(1). https://doi.org/10.1038/s44168-023-00082-1
8. Peters GP, Al Khourdajie A, Sognnaes I, Sanderson BM. AR6 scenarios database: an assessment of current practices and future recommendations. npj Clim Action. 2023;2(1). https://doi.org/10.1038/s44168-023-00050-9
9. Intergovernmental Panel on Climate Change (IPCC). Climate Change 2021 – The Physical Science Basis. In: Masson-Delmotte V, Zhai P, Pirani A, Connors SL, Péan C, Berger S. et al. Editors, Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press. 2023. https://doi.org/10.1017/9781009157896
10. IPCC. About the IPCC. 2024a. [cited 2024 February]. https://www.ipcc.ch/about/
11. Mastrandrea MD, Field CB, Stocker TF, Edenhofer O, Ebi KL, Frame DJ, et al. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties. IPCC Cross-Working Group Meeting on Consistent Treatment of Uncertainties, Jasper Ridge, CA, USA, 6-7 July 2010.
12. Vardy M, Oppenheimer M, Dubash NK, O'Reilly J, Jamieson D. The Intergovernmental Panel on Climate Change: Challenges and Opportunities. Annu Rev Environ Resour. 2017;42(1):55–75. https://doi.org/10.1146/annurev-environ-102016-061053
13. Maas TY, Pauwelussen A, Turnhout E. Co-producing the science–policy interface: towards common but differentiated responsibilities. Humanit Soc Sci Commun. 2022;9(1). https://doi.org/10.1057/s41599-022-01108-5
14. Slade R, Pathak M, Connors S, Tignor M, Okem AE, Leprince-Ringuet N. Back to basics for the IPCC: applying lessons from AR6 to the Seventh Assessment Cycle. npj Clim Action. 2024;3(1). https://doi.org/10.1038/s44168-024-00130-4
15. IPCC. Appendix A to the Principles Governing IPCC Work: Procedures for the Preparation, Review, Acceptance, Adoption, Approval and Publication of IPCC Reports. [cited 2024 February] https://www.ipcc.ch/site/assets/uploads/2018/09/ipcc-principles-appendix-a-final.pdf
16. Ford JD, Biesbroek R, Ford LB, Creutzig F, Haddaway N, Harper S, et al. Recommendations for producing knowledge syntheses to inform climate change assessments. Nat Clim Chang. 2025. https://doi.org/10.1038/s41558-025-02354-6
17. Montfort S, Callaghan M, Creutzig F, Lamb WF, Lu C, Repke T, et al. Systematic global stocktake of over 50,000 urban climate change studies. Nat Cities. 2025;2(7):613–25. https://doi.org/10.1038/s44284-025-00260-8
18. De-Gol AJ, Le Quéré C, Smith AJP, Aubin Le Quéré M. Broadening scientific engagement and inclusivity in IPCC reports through collaborative technology platforms. npj Clim Action. 2023;2(1). https://doi.org/10.1038/s44168-023-00072-3
19. Callaghan M, Schleussner C-F, Nath S, Lejeune Q, Knutson TR, Reichstein M, et al. Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. Nat Clim Chang. 2021;11(11):966–72. https://doi.org/10.1038/s41558-021-01168-6
20. Minx JC, Callaghan M, Lamb WF, Garard J, Edenhofer O. Learning about climate change solutions in the IPCC and beyond. Environmental Science & Policy. 2017;77:252–9. https://doi.org/10.1016/j.envsci.2017.05.014
21. Pollitt H, Mercure J-F, Barker T, Salas P, Scrieciu S. The role of the IPCC in assessing actionable evidence for climate policymaking. npj Clim Action. 2024;3(1). https://doi.org/10.1038/s44168-023-00094-x
22. Carmona R, Reed G, Thorsell S, Dorough DS, MacDonald JP, Rai TB, et al. Analysing engagement with Indigenous Peoples in the Intergovernmental Panel on Climate Change's Sixth Assessment Report. npj Clim Action. 2023;2(1). https://doi.org/10.1038/s44168-023-00048-3
23. Skea J, Shukla P, Al Khourdajie A, McCollum D. Intergovernmental Panel on Climate Change: Transparency and integrated assessment modeling. WIREs Climate Change. 2021;12(5). https://doi.org/10.1002/wcc.727
24. Cointe B. The AR6 Scenario Explorer and the history of IPCC Scenarios Databases: evolutions and challenges for transparency, pluralism and policy-relevance. npj Clim Action. 2024;3(1). https://doi.org/10.1038/s44168-023-00075-0
25. Hulme M, Zorita E, Stocker TF, Price J, Christy JR. IPCC: cherish it, tweak it or scrap it?. Nature. 2010;463(7282):730–2. https://doi.org/10.1038/463730a PMID: 20148014
26. IPCC. Decisions adopted by the Panel. Sixty-Second Session of the IPCC, Hangzhou, China, 24 – 28 February 2025. IPCC-LXII/Doc.1, Rev. 1 and IPCC-LXII/Doc.1, Rev.1, Add.1. 2025. https://www.ipcc.ch/site/assets/uploads/2025/03/IPCC-62-Decisions.pdf

27. IPCC. Options for expert meetings and workshops for the seventh assessment cycle. Sixty-First Session of the IPCC, Sofia, Bulgaria, 27 July – 2 August 2024. 2024b. [cited February 2025], Available from: https://apps.ipcc.ch/eventmanager/documents/87/050720240428-Doc.%207%20-%20Options%20for%20Expert%20Meetings.pdf

28. Muccione V, Vaghefi SA, Bingler J, Allen SK, Kraus M, Gostlow G, et al. Integrating artificial intelligence with expert knowledge in global environmental assessments: opportunities, challenges and the way ahead. Reg Environ Change. 2024;24(3). https://doi.org/10.1007/s10113-024-02283-8

29. Van Noorden R, Perkel JM. AI and science: what 1,600 researchers think. Nature. 2023;621(7980):672–5. https://doi.org/10.1038/d41586-023-02980-0 PMID: 37758894

30. Pawlicka A, Pawlicki M, Kozik R, Choraś M. The rise of AI-powered writing: How ChatGPT is revolutionizing scientific communication for better or for worse. Communications in Computer and Information Science. 2024;2014:317–27.

31. IEA. IEA launches new GPT tool to explore flagship energy data and analysis using artificial intelligence. International Energy Agency. 2024. [cited 2025 April 23]. https://www.iea.org/news/iea-launches-new-gpt-tool-to-explore-flagship-energy-data-and-analysis-using-artificial-intelligence

32. Vaghefi SA, Stammbach D, Muccione V, Bingler J, Ni J, Kraus M, et al. ChatClimate: Grounding conversational AI in climate science. Commun Earth Environ. 2023;4(1). https://doi.org/10.1038/s43247-023-01084-x

33. Lelong J, Achache N, Olympie G, Chesneau N, De la Calzada N. ClimateQ&A - a hugging face space by ekimetrics. 2023. Available from: https://huggingface.co/spaces/Ekimetrics/climate-question-answering

34. Berrang-Ford L, Siders AR, Lesnikowski A, Fischer AP, Callaghan MW, Haddaway NR, et al. A systematic global stocktake of evidence on human adaptation to climate change. Nat Clim Chang. 2021;11(11):989–1000. https://doi.org/10.1038/s41558-021-01170-y

35. Lamb WF, Creutzig F, Callaghan MW, Minx JC. Learning about urban climate solutions from case studies. Nat Clim Chang. 2019;9(4):279–87. https://doi.org/10.1038/s41558-019-0440-x

36. Wang B, Pan S-Y, Ke R-Y, Wang K, Wei Y-M. An overview of climate change vulnerability: a bibliometric analysis based on Web of Science database. Nat Hazards. 2014;74(3):1649–66. https://doi.org/10.1007/s11069-014-1260-y

37. Creutzig F, Callaghan M, Ramakrishnan A, Javaid A, Niamir L, Minx J, et al. Reviewing the scope and thematic focus of 100 000 publications on energy consumption, services and social aspects of climate change: a big data approach to demand-side mitigation *. Environ Res Lett. 2021;16(3):033001. https://doi.org/10.1088/1748-9326/abd78b

38. Callaghan MW, Minx JC, Forster PM. A topography of climate change research. Nat Clim Chang. 2020;10(2):118–23. https://doi.org/10.1038/s41558-019-0684-5

39. Đukić M, Škembarević M, Jejić O, Luković I. Towards the utilization of AI-powered assistance for systematic literature review. In: New Trends in Database and Information Systems (ADBIS 2024), 2024. 195–205.

40. Berrang-Ford L, Sietsma AJ, Callaghan M, Minx JC, Scheelbeek PFD, Haddaway NR, et al. Systematic mapping of global research on climate and health: a machine learning review. Lancet Planet Health. 2021;5(8):e514–25. https://doi.org/10.1016/S2542-5196(21)00179-0 PMID: 34270917

41. Sietsma AJ, Theokritoff E, Biesbroek R, Canosa IV, Thomas A, Callaghan M, et al. Machine learning evidence map reveals global differences in adaptation action. One Earth. 2024;7(2):280–92. https://doi.org/10.1016/j.oneear.2023.12.011

42. Callaghan M, Banisch L, Doebbeling-Hildebrandt N, Edmondson D, Flachsland C, Lamb WF, et al. Machine learning map of climate policy literature reveals disparities between scientific attention, policy density, and emissions. npj Clim Action. 2025;4(1). https://doi.org/10.1038/s44168-024-00196-0

43. Haman M, Školník M. Using ChatGPT to conduct a literature review. Account Res. 2024;31(8):1244–6. https://doi.org/10.1080/08989621.2023.2185514 PMID: 36879536

44. Jones N. OpenAI's "deep research" tool: is it useful for scientists?. Nature News. 2025. https://doi.org/10.1038/d41586-025-00377-9 PMID: 39915598

45. Haddaway NR, Callaghan MW, Collins AM, Lamb WF, Minx JC, Thomas J, et al. On the use of computer-assistance to facilitate systematic mapping. Campbell Syst Rev. 2020;16(4):e1129. https://doi.org/10.1002/cl2.1129 PMID: 37016615

46. Liverman D, von Hedemann N, Nying'uro P, Rummukainen M, Stendahl K, Gay-Antaki M, et al. Survey of gender bias in the IPCC. Nature Comment, 01 February. 2022. https://www.nature.com/articles/d41586-022-00208-1

47. Tyler C, Akerlof KL, Allegra A, Arnold Z, Canino H, Doornenbal MA, et al. AI tools as science policy advisers? The potential and the pitfalls. Nature Comment. 2023.

48. Canali S, Barone-Adesi F. Can AI deliver advice that is judgement-free for science policy?. Nature Correspondence. 2023. https://www.nature.com/articles/d41586-023-03949-9

49. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. arXiv. 2024. https://arxiv.org/abs/2202.03629

50. Zhao L, Nguyen K, Daumé H. Hallucination detection for grounded instruction generation. In Bouamor H, Pino J, Bali K. Editors, Findings of the Association for Computational Linguistics: EMNLP 2023. 2023; (pp. 4044–4053). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.266

51. Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit: leveraging artificial intelligence to combat climate change-opportunities, challenges, and recommendations. AI Soc. 2023;38(1):283–307. https://doi.org/10.1007/s00146-021-01294-x PMID: 34690449

52. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv. 2021. https://arxiv.org/abs/2005.11401

53. Chen Y, Benton J, Radhakrishnan A, Uesato J, Denison C, Schulman J, et al. Reasoning models don't always say what they think. Anthropic. 2025. https://assets.anthropic.com/m/71876fabef0f0ed4/original/reasoning_models_paper.pdf

54. Perez E, Huang S, Song F, Cai T, Ring R, Aslanides J, et al. Red teaming language models with language models. arXiv preprint. 2022. https://doi.org/10.48550/arXiv.2202.03286

55. Yang Y, Li Z, Dong Q, Xia H, Sui Z. Can large multimodal models uncover deep semantics behind images?. arXiv preprint. 2024. https://doi.org/10.48550/arXiv.2402.11281

56. Edge D, Trinh H, Cheng N, Bradley J, Chao A, Mody A, et al. From local to global: A graph RAG approach to query-focused summarization. arXiv preprint. 2025. https://doi.org/10.48550/arXiv.2404.16130

57. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. arXiv. 2017. https://arxiv.org/abs/1706.03762

58. Holtzman A, Buys J, Du L, Forbes M, Choi Y. The curious case of neural text degeneration. In: International Conference on Learning Representations (ICLR), 2020. https://arxiv.org/abs/1904.09751

59. Kaufmann T, Weng P, Bengs V, Hüllermeier E. A Survey of Reinforcement Learning from Human Feedback. arXiv. 2023. https://doi.org/10.48550/arXiv.2312.14925

60. Stanford CRFM. HELM: A Holistic Framework for Evaluating Foundation Models. Stanford University, Center for Research on Foundation Models. 2024. https://crfm.stanford.edu/helm/latest/

61. White C, Dooley S, Roberts M. LiveBench: A challenging, contamination-free LLM benchmark. arXiv preprint. 2024. https://doi.org/10.48550/arXiv.2406.19314

62. Chiang WL, Zheng L, Sheng Y. Chatbot Arena: An open platform for evaluating LLMs by human preference. 2024. https://doi.org/10.48550/arXiv.2403.04132

63. Gu J, Jiang X, Shi Z. A survey on LLM-as-a-Judge: Advancements, applications, and challenges. arXiv preprint arXiv 2025. https://doi.org/10.48550/arXiv.2411.15594

64. Zheng L, Chiang WL, Sheng Y. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv 2023. https://doi.org/arXiv:2306.05685