REF-AI

Exploring the potential of generative AI for REF2029

Professor Richard Watermeyer, Professor Lawrie Phipps,
Rodolfo Benites & Professor Tom Crick

Funded by:

University of BRISTOL

CHET
CENTRE FOR HIGHER EDUCATION TRANSFORMATIONS

Jisc

UKRI Research England

Contents

Foreword

Generative artificial intelligence (GenAI) is transforming how research is developed, delivered, and assessed. Its capacity to analyse data, summarise complex ideas, and accelerate writing brings opportunities and profound disruption. As this report shows, universities are already experimenting with these tools, both with confidence and often with caution, whilst at the same time grappling with questions of integrity, authorship, ethics, and trust. This fast-changing landscape has implications for the whole of the UK research systems, including the Research Excellence Framework (REF).

The evidence gathered here gives a picture of optimism tempered with realism. Across disciplines, academics and professional staff recognise GenAI's potential to enhance efficiency, reduce administrative burden, and support fairer, more consistent REF preparation. Yet they also voice deep concern about data security, plagiarism, bias, and the risk of diminishing human creativity. This report does not call for rejection or uncritical adoption but for governed innovation, a deliberate, transparent, and equitable approach that preserves the integrity of UK research. The recommendations in the report set out practical steps for consideration by all actors in the system. They urge the development of formal governance frameworks, sector-wide standards for transparency and disclosure, and equitable access to safe, secure AI tools. They recognise differences in institutional capacity and call for shared solutions, such as national access to AI platforms, consistent training, and cross-sector oversight to prevent inequality and ensure accountability.

The challenges are significant, the technology evolves faster than regulation, behaviours are shifting rapidly, and the geopolitical context of AI development raises unresolved questions about sovereignty and data hosting. Yet if approached with integrity and foresight, GenAI could strengthen, rather than weaken the REF, modernising assessment without compromising human judgement.

This independent report offers both a caution and a call to action. It warns against haste and complacency alike, while inviting the sector to lead with principle, collaboration, and critical literacy. With the right safeguards, the integration of GenAI can help us uphold excellence, fairness, and trust in the assessment of UK research.

Dr Steven Hill
Director of Research, Research England

Acknowledgements

The research team would like to extend its thanks to all those who have been instrumental to the success of the REF:AI project.

Thanks to all of our research participants who were so generous with their time and accommodating of our own restricted timeframe: the 200+ focus group participants across our 16 sampled institutions in England, Wales, Scotland and Northern Ireland; our 16 institutional REF-Leads and 16 Pro Vice-Chancellors (or equivalents) who engaged so thoughtfully and deeply with our interview questions; and the 400+ respondents to our national survey. We hope to have given fair representations to all your views.

Additional thanks to our 16 institutional REF-Leads who did such a fantastic job in organising our campus visits.

Thanks to Research England for responding positively to our initial research proposal and funding this independent study, and Myles Furr, of the same, for being a constant source of advice and support.

To Medr and the Scottish Funding Council for supporting the study in Wales and Scotland.

To our colleagues at the University of Bristol and at Jisc for their various contributions in supporting this study. Special thanks to Verena Weigert and Dr Donna Lanclos.

To our colleagues at ARMA for all their help in the distribution of our survey.

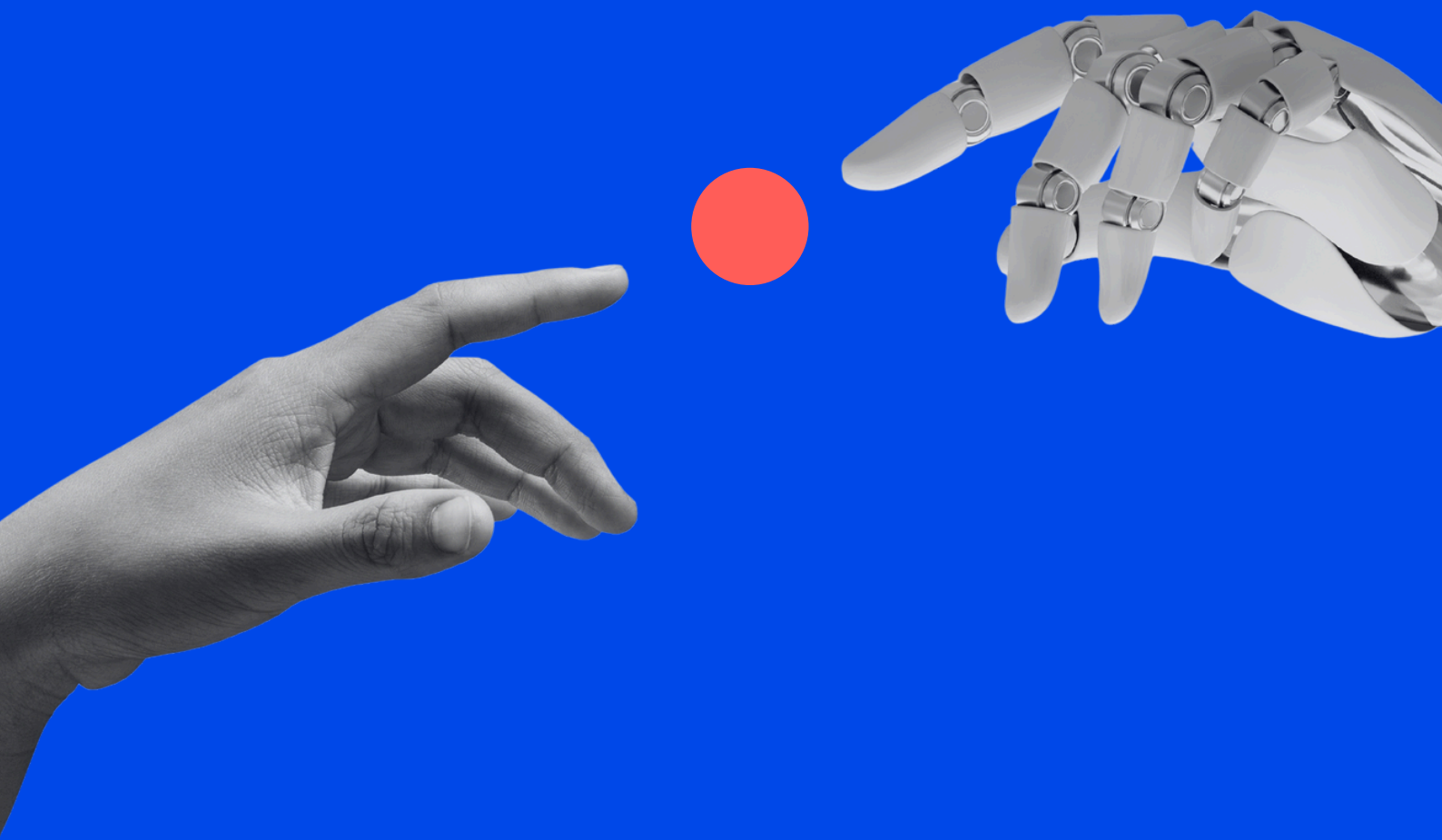
And finally, prospectively, to all those who will engage with the findings of our study.

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that influenced or could have appeared to influence the research herein reported

Richard Watermeyer, Lawrie Phipps, Rodolfo Benites & Tom Crick
December, 2025

Introduction



The REF:AI project is an independent research study funded by Research England. It has been undertaken as a rapid qualitative investigation of existing, emerging and prospective applications of and attitudes towards the use of generative artificial intelligence (GenAI) tools for purposes of institutional Research Excellence Framework (REF) preparations and assessment.

It has sought to respond to a significant knowledge gap pertaining to how current (and rapidly evolving) automative and agentic technology is/can/might be used in supporting higher education institutions (HEIs) and REF panels better manage the time and resource-demanding work of REF submissions and assessments, respectively.

Much is made of the resource demands of the REF and the significant financial outlay committed by HEIs. While REF2021 is calculated to have accrued a total cost of £471m, at an average cost of £3m for each submitting HEI, some have forecast that the various adjustments being proposed to the 2029 instalment might culminate in expenditure of £1bn. Though arguments are made that the REF offers good value for money, the well-reported financial frailty of the UK's higher education sector, and by extension, concerns of substantial job losses and prospect of nationwide strikes, appear to undermine the palatability and plausibility of such claims. Defending the REF's 'value for money' thesis has been further confounded by the discontinuation of other comparator performance-based research funding models like Excellence Research Australia and the Performance Based Research Fund in New Zealand and calls in the former's case for its update with 'a modern, data-driven approach, informed by peer review'. In such context, the REF:AI project was conceived to understand how technology might facilitate a similar ambition of a less-burdensome, more agile and responsive REF. Building on previous recent research examining the impact of GenAI tools on academic work (Watermeyer, 2024a, 2024b), the research team set out to understand the extent to which GenAI is being used or might be used in the context of the REF as a data-led process.

Our initial provocation that GenAI might be used for the REF drew something of a backlash with opinion that integration of AI into the REF is a non-starter and would be 'fraught with danger'. Other opinions however emerged that were more varied in considering whether GenAI might 'improve the REF'. More recently voices across the sector, including the President of the Royal Society, have questioned whether the REF, in its current form, offers the 'best use of brainpower' and advocate for the introduction of AI tools into its re-architecture.

Through a consultation of 16 HEIs across all four corners of the UK (and a national attitudinal survey) REF:AI has identified current and potential future use of GenAI tools for the REF; perceived contributions of GenAI to REF processes; conditions for the appropriate use of GenAI tools in REF; concerns and challenges for GenAI integration in REF processes; and prognoses for future GenAI use in the REF. In light of the recent hiatus to REF planning, our hope is that this report – as a representation of multiple voices from across the sector – directly feeds into fuller critical consideration of the contribution of AI technology to 'simplify, streamline and reduce burden on the sector' and 'ensure alignment with government priorities and vision for higher education.

This report provides a comprehensive critical review of our findings. It establishes sector wide attitudinal trends in relation to GenAI use; triangulated with the perspectives and experiences of those at the frontline of REF preparations within universities. It is organised into an executive summary and offering of key recommendations, which are followed by extensive presentation and analysis of survey, interview and focus group data. It is also supported by a critical review of an extant scholarly literature which considers the efficacy of GenAI as a disruptive technology to research assessment and governance.

Executive Summary

1. GenAI adoption for REF purposes is shown to be relatively shallow and/or otherwise patchy across HEIs. There is, however, variation in the extent to which institutions are already experimenting or considering experimenting with GenAI tools for the purposes of their REF submissions.

2. Variation in experimentation with GenAI tools is mainly influenced and explained by institutional resource capacity. HEIs with a less developed or more modest infrastructure for supporting REF, are less likely to have experimented with GenAI tools for REF purposes and are more agnostic to their value.

3. There is evidence that some institutions have/are in the process of creating in-house GenAI tools for REF purposes, and in some instances have already attained proof-of-concept as relates to credible use. Some institutions also report use of algorithmic tools in the context of REF2021 output selection.

4. Proof-of-concept in the use of GenAI tools for REF purposes is restricted to only a handful of institutional examples and by extension therefore, proof-of-concept as achieved only at the level of institutional data.

5. The credibility and sector-wide integration of GenAI tools for REF purposes is argued to be contingent upon achieving proof-of-concept via the national REF dataset. GenAI tools might be piloted using existing REF2021 data and further refined through exposure to REF2029 data.

6. GenAI tools are viewed for their contribution in facilitating institutions' management of the REF's considerable data demands and as offering an opportunity for unburdening from the REF's substantial regulatory burden.

7. GenAI tools should not, however, be confused as a silver bullet for the REF's inefficiencies and shortcomings. GenAI is instead an unavoidable – perhaps indivisible – dimension of the REF's data(fied) architecture.

8. GenAI tools are predominantly viewed, in the current context, as providing a potentially valuable addition to the REF's human-based activities and for purposes of inter alia output selection; output assessment; evidence reconnaissance, content and narrative generation for impact and environment (people, culture) dimensions of the REF; calibration and validation.

9. In the present context, GenAI tools are viewed as a supplement to REF processes and not (currently) as a replacement for human input and oversight.

10. A blended approach is recommended for the integration of GenAI tools into the various aspects of the REF, and thus REF personnel (within institutions and populating panels) working with and alongside GenAI.

11. The integration of GenAI into the REF will have a substantial impact on REF related labour at institutional and sector levels and will instigate a re-rationalisation of REF related job functions affecting HEI's professional services staff and adjunct labour (e.g. impact consultants and evaluation specialists) servicing HEIs' REF needs.

12. There is broad consensus that the REF will inevitably become increasingly AI infused and ultimately, fully automated.

13. There is wide belief that GenAI tools will be utilised by REF2029 panellists. Such use is rationalised because of the (uneven) assessment burden placed on panellists and in terms of a perceived value of GenAI tools in validating panel judgements[1]. Assertions that panellists will not use GenAI tools (particularly given the potential improved capability of such tools by the beginning of the assessment period in 2029) will be treated with scepticism and suspicion – potentially impairing claims to the authenticity of the REF as informed exclusively by human judgement.

14. Use of GenAI tools by REF panellists and/or the prospect of a partially or fully automated REF is also viewed positively on the basis of allowing those populating panels – many leading UK researchers – to commit to their research and thus the UK's 'science' basis without the disruption caused by the REF's obligation (and as understood both in the terms of their host institution's REF preparations and REF panel assessments).

15. There is a strong belief that embargoing the use of GenAI tools for REF purposes will culminate in tacit, unregulated and inappropriate use.

16. There is broad consensus that GenAI tools will not only help to alleviate what is reported to be REF's unsustainable financial burden on submitting HEIs and the broader UK HE sector but will provide for much more efficient and 'in-time' reporting of research excellence, enabling a more accurate and credible distribution of QR funding.

---

[1] Relatedly it is worth noting that 11 sub panels made use of citation data to inform assessments of research quality in REF2021: https://2021.ref.ac.uk/guidance/citation-and-contextual-data-guidance/index.html

17. HEIs that are leaning towards the use of GenAI for REF purposes are doing so not only on the basis of alleviating academic staff from REF-related workload but on the basis that the integration of GenAI tools into institutional REF preparations will lead to more accurate and therefore more competitive submissions. GenAI use in REF preparations is consequently rationalised as providing institutions with a competitive edge.

18. In corollary, uneven investment in GenAI tools by HEIs is seen as unbalancing the capacity of HEIs, especially smaller and less well REF resourced institutions to return competitive REF submissions and fairly participate within the REF (where consistently understood primarily as a competition for Quality Related (QR) funding and positional status in REF performance league tables).

19. A standardised REF tool (underpinned by standardised practice) is advocated as a necessary condition to what many view as the REF's necessary integration of GenAI tools and for achieving (i) more equitable participation among HEIs of varying research power and resource capability (ii) a solution to data privacy and security concerns. However, there is also fear that a subscription model of usage, would disadvantage less well-resourced HEIs and might compound a digital divide within the REF. Moreover, there are concerns that a standardised GenAI REF tool might not meet the expectations/requirements of some (more AI mature) HEIs and face resistance.

20. Some institutions and disciplines will be better prepared to pivot towards GenAI integration in the REF.

21. AI literacy in the context of REF (and wider research) applications is low and appears to correspond to low trust and therefore adoption and/or avoidance of GenAI tools for REF purposes.

22. Usage and related acceptance/disavowal of GenAI for REF purposes is also attributed to variations of 'generational attitudes' towards technology adoption.

23. There is greater acceptance of the value of GenAI tools for REF processes by professional services staff than academics – even though the former's role is ostensibly most risked by the technology, or susceptible to significant change.

24. Organisational/cultural maturity in relation to GenAI tools is depicted as low in HEIs and in need of urgent redress.

25. Despite what appears as concentrated opposition to the integration of GenAI tools in the REF, we find reported in many larger HEIs (and those typically making larger institutional REF submissions) a burgeoning appetite for the deployment of GenAI tools for research assessment and wider research practices.

26. Resistance to the integration of GenAI tools into REF processes (institutional preparations and panel assessment) is primarily attributed to concerns that they are corruptive to peer-review and challenge the primacy of human expertise in the formation of scholarly judgements. Where GenAI is observed as a threat to the REF as a process of peer-review it is also seen to diminish and marginalise the contribution of academic researchers, and more explicitly, their ability to curate, control and/or have some form of influence over the UK's system of research governance.

27. There is (arguably relatedly) a strong feeling that human-based peer-review must remain central to the REF's evolution.

28. Concurrently, the REF is frequently characterised as poorly representative of a gold-standard of peer-review and that instead as a 'comfort blanket', no longer provides a justifiable reason for preventing the REF's innovation by GenAI.

29. Instead, GenAI tools are viewed as facilitative to the preparation of REF submissions, where the depth and accuracy of internal 'peer-review' is questioned, and where the focus of institutional assessments are geared towards predicting how panels will score.

30. There is an existing dearth if not general absence of formal guidance (and/or policy) for the 'appropriate' application of GenAI tools for REF purposes in HEIs (and at sector level), and practice involving GenAI tools is running significantly ahead of policy formation. Institutional policy for GenAI use in the REF is found to be generally overlooked (and viewed as inappropriate/unsustainable given the speed of technology innovation and corresponding policy redundancy). Discussions pertaining to the use of GenAI tools for institutional submissions are also found to be in their infancy. The REF:AI study as a national consultation is reported to have kick-started and accelerated internal discussions as to the appropriate use of GenAI tools in institutional REF preparations.

31. Within the course of our consultation seldom mention was made of responsible research assessment initiatives such as DORA[1] and CoARA[2]. There is no clear view that the deployment of GenAI tools in the REF is antithetical to the ambitions of such initiatives.

32. Concerns of bias and inaccuracy related to GenAI tools are widely acknowledged yet so too is the speed by which AI technology is improving, leading to forecasts that the presence of human (and machine) bias, error and omissions in the REF should decrease.

33. The deployment of GenAI tools in institutional REF preparations are also considered for enabling HEIs to better know themselves as datafied organisations and as a useful prop for institutional memory.

34. Opinion exists that the advancements of GenAI (and other digital technologies) are making the REF (in its current form) appear outdated and unfit for purpose in a milieu of data-led research governance.

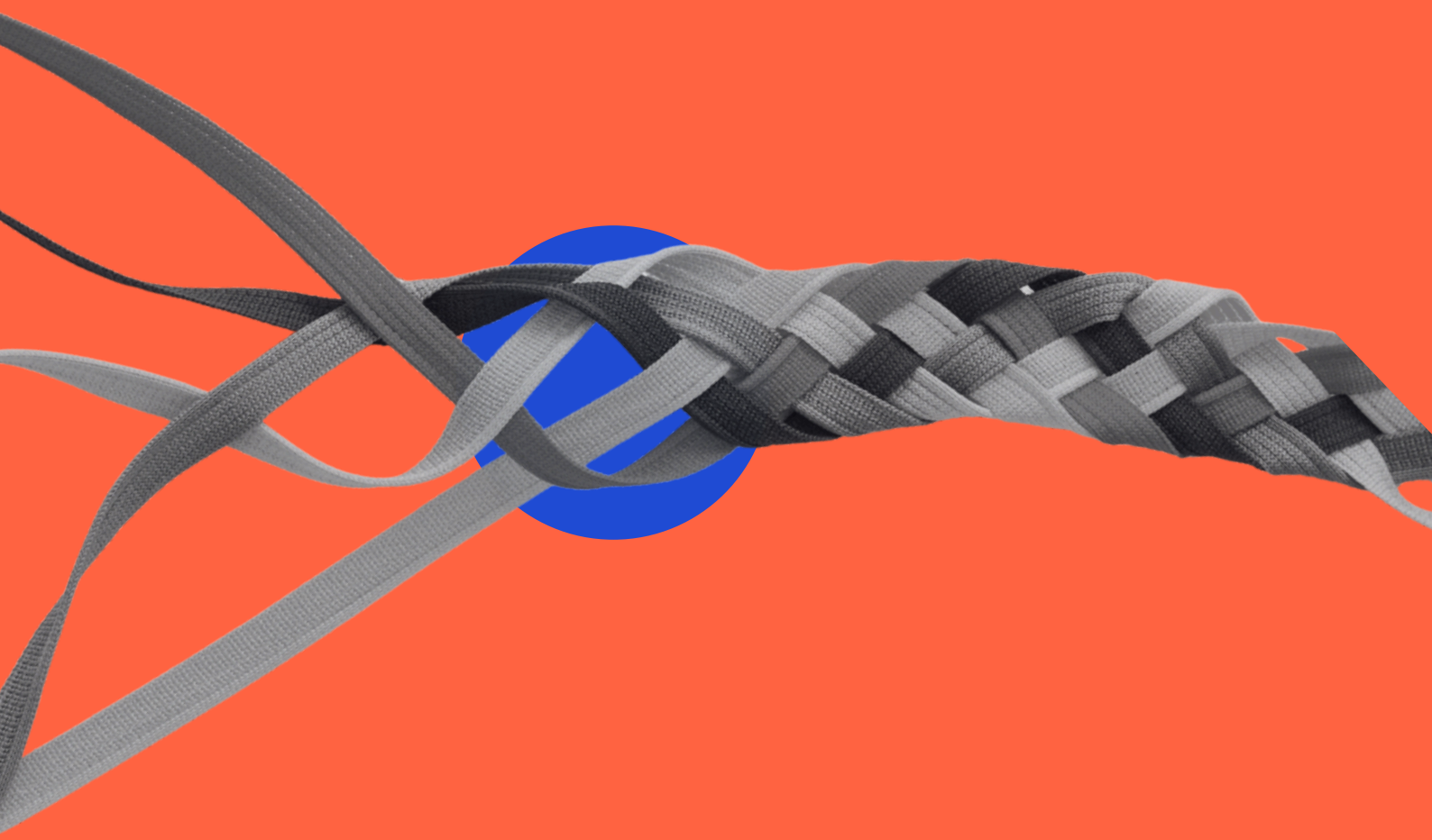
35. There is argument that the REF is compelled to lean into GenAI tools to maintain its international credibility and the reputation of UK academia as a science leader.

36. A concern also emerges that GenAI integration will depreciate the contribution of the REF to the development of UK researchers (through their diminished involvement in output assessment, development of impact case studies etc.), though we also find scepticism as to the extent to which such developmental opportunities exist within submitting HEIs.

37. HEIs' adoption of GenAI tools for the REF is in likely conflict with the UK HE sector's commitment to environmental sustainability, where GenAI tools (are presently and foreseeably) characterised by their high natural resource needs.

[2] DORA is the Declaration on Research Assessment: https://sfdora.org
[3] CoARA is the Coalition for Advancing Research Assessment: https://www.coara.org

Recommendations



AI Policy and Governance in Institutions

1. Every university should establish and publish a formal policy on the use of GenAI for research purposes and as specific to REF, defining permissible use and requiring transparent disclosure of any AI involvement. Policies should be embedded within existing research integrity and ethics frameworks to ensure consistent, accountable governance.

2. Senior Management Teams should designate a senior committee or responsible officer to oversee AI use in research and as specific to REF, ensuring compliance, regular review, and alignment with evolving legal, ethical, and technological standards.

3. AI literacy is essential. All staff involved in REF2029, both academic and professional, should receive structured training on the responsible and effective use of GenAI tools for research and assessment purposes.

4. Approved and secure AI environments should be mandated for REF-related work. Confidential or unpublished research content should not be entered into public or unvetted AI systems.

5. Institutions should maintain an internal AI Use Log documenting all GenAI involvement in REF preparation, including prompts, models used, purpose, and human verification, retained for audit and accountability.

6. Each institution should apply a Quality Assurance Checklist for AI-assisted outputs, verifying factual accuracy, references, and bias, with final sign-off by an academic lead.

7. Clear rules for IP ownership, attribution, and acknowledgement must be set, ensuring that any AI-generated text or analysis is cited and attributed according to research integrity standards.

8. Universities should conduct risk and ethics reviews before adopting AI workflows that handle sensitive, personal, or unpublished research data.

9. Institutional policies should ensure consistent AI governance across all Units of Assessment (UoAs), prohibiting local deviations unless justified and documented.

REF Guidance and Auditing Mechanisms

10. The REF Steering Group should work with the main and sub panels to provide formal, sector-wide guidance on GenAI use in REF2029, mandating transparency and defining acceptable forms of AI assistance (for example, editorial or summarising support, not unacknowledged content generation).

11. A comprehensive REF AI Governance Framework should be established and published, covering both institutional preparation and panel assessment.

12. REF submissions should include a standardised AI Disclosure Statement, using a sector-wide template to ensure consistency and comparability.

13. REF main and sub-panels when publishing their panel criteria and working methods, should include how GenAI tools will be deployed, reviewed, and moderated in assessment.

14. Completion of AI literacy training should be a precondition for all REF2029 panellists, ensuring informed and ethical use of AI in assessment and for underpinning judgements about the (non)selection of GenAI tools by panels.

15. REF assessments should include a human verification step for any AI-assisted analyses or summaries, confirming that final judgements rest on human academic expertise.

16. Panels and institutions should disclose and document all AI use during preparation and assessment, supported by an auditable trail and accountability mechanisms.

17. REF guidance should include discipline-specific guardrails, clarifying how AI may be applied in varied research contexts.

Equity, Capability, and Sustainability

18. Equitable access to GenAI tools should be provided for all HEIs and REF panels. Sector leaders could collaborate to develop or procure a shared, high-quality AI platform accessible to all institutions.

19. Institutions and disciplines less able to adopt AI at scale must be supported through shared infrastructure, training, and policy flexibility to preserve parity in assessment.

20. Sector leaders must introduce periodic bias and EDI monitoring of AI-assisted processes, publish findings and implement mitigations where inequities are detected.

21. A capability baseline exercise should be introduced, requiring HEIs to self-declare AI readiness (policy coverage, training, secure tooling) to promote transparency and parity.

22. The environmental impact of AI integration must be modelled, monitored, and mitigated through responsible technology choices and transparent sustainability reporting.

23. Ongoing research is needed to establish safe, secure, sector-level access to GenAI tools that keeps pace with the rapid evolution of AI technologies, changing user behaviours, and the geopolitical uncertainty surrounding data hosting and model ownership. Future analysis of GenAI use is also warranted at two critical junctures: (i) further to institutional REF submissions in late 2028, and (ii) the culmination of REF2029 formal assessment in late 2029.

Statement of methods

The REF-AI study consisted of: (i) focus groups of up to 2 hour 30 duration, across 16 discrete institutional settings, with personnel (academics and professional services staff) with formal responsibility for the development of institutional REF submissions; (ii) semi-structured interviews with 16 institutional REF-leads and 16 pro vice-chancellors for research (PVCRs); (iii) an online survey of n=386 respondents (academics and professional services staff).

The interviews and focus groups were undertaken with staff from different types of HEIs spanning the entirety of the UK and including (n=8) large Russell Group (RG) universities; (n=1) a research-intensive non-Russell Group university; (n=6) post-92 universities, and (n=1) specialist institution. Interview and focus group data was intended to baseline existing/emerging activity and attitudes involving the use of GenAI tools for REF preparations across HEIs with both broad and more narrow representation across Units of Assessment (UoA), while also scenario-building for appropriate institutional application. Crucially, the interviews/focus groups were envisaged to identify not only inter-institutional but intra-institutional commonality and variation and thus how GenAI tools are/might be variously used – and used optimally – across both different institutional and disciplinary contexts.

The interviews/focus groups were designed to identify the extent of use or experimentation with GenAI tools –identifying what kinds of tools are being used and/or considered (by whom) and (prospectively) benefitted from – in tandem with analysis of institutional policy (however nascent and emergent) legislating the use of such tools for the purpose of internal institutional evaluation and generation of particulars related to REF2029's three core (current) dimensions of assessment: Knowledge and Understanding; Engagement and Impact; and People, Culture and Environment.

The research was focused on eliciting how HEIs interpret the value proposition of GenAI tools to their REF preparations and how they define, embed and promote (or possibly neglect) the parameters and conditions of responsible use – including ethical appraisal and safeguarding and linked forms of AI literacy.

The research sought to assess the potential of such tools in making REF preparations more efficient, cost effective and less resource burdensome according to varying applications by different user groups. Moreover, the research was intended to analyse the extent to which GenAI tools may not only support but enhance institutional capacity in making REF submissions.

Interview and focus group data was also intended to inform how GenAI tools might support and enhance the work of REF panellists and feed into recommendations for panellists' appropriate use.

Prior to the onset of research, a formal submission was made for ethical approval, subsequently granted by the University of Bristol's School of Education Research Ethics Committee. The REF:AI project's ethics ID is 25175.

The research began with an inception meeting with Research England, where the research methodology was further considered and confirmed and milestones agreed. The latter changed in so much as our reporting was been brought forward to offer greatest use to stakeholders in light of a pause on the development of REF2029 guidelines.

Fieldwork commenced in June (2025). The vast majority of data was collected in the June-July period. Three further institutional visits occurred in September after the summer break. The bulk of data analysis occurred from mid-September through to the end of October (2025).

Interviews and focus groups were established in conjunction with an institutional gatekeeper who was drawn upon for the identification and recruitment of key university personnel. A roughly even split of academics and professional service staff, all with REF related roles, were selected for the focus groups. Academic participants represented a broad spectrum of disciplinary domains and were typically senior scholars with previous track record of involvement in REF institutional preparations. Some also had prior experience of serving on REF panels. Some are due to serve on REF2029 panels.

Access to institutions was facilitated by Research England, Medr and the Scottish Funding Council who also advised on the selection of institutions. An initial sample of 10 HEIs was selected though this ultimately increased to 16.

Letters of invitation were distributed to the PVCRS of selected institutions. Our invitations to participate within the study were broadly well received and accepted. Only in the case of one institution did we receive no response. At one other institution, multiple efforts to arrange a focus group and interview were ultimately abandoned due to a lack of progress. In every instance we arranged an initial briefing call with the participating institution to further explain the purpose of the study and our research requirements.

All institutional visits were made in person. They typically involved three members of the research team, though on occasion, given the concentrated timetabling of fieldwork and geographical range of visits, only two researchers were able to attend.

Where possible interviews with PVCRs and institutional REF-Leads occurred on the same day. Due to the busyness of such persons and restrictions of diary availability, some of the interviews were undertaken online and via video conference (these represented roughly a third of our interviews). We note no discernible difference in the quality of these online discussions compared to those conducted in person.

All the interviews and focus groups were recorded with the permission of participants and auto transcribed. Transcriptions were checked by members of the research team.

A process of reflexive thematic analysis was then undertaken in making sense of a considerable data set of approximately 40 hours of recorded focus group data and 30 hours of interview data. For the purposes of this report, survey data is treated only as descriptive statistics which identify basic attitudinal trends. The survey itself was arranged as a series of Likert scale agreement questions and also included two open-text questions. This report only considers Likert scale data.

The identity of our participants and the institutions they work for has been strictly anonymised.

Reviewing the literature

In this part of the report, we explore the applications of GenAI tools in academic research assessment processes, specifically, the literature related to Large Language Models (LLMs); a subset of generative artificial intelligence. The research literature is while burgeoning still juvenescent.

It spans academic literature (journal articles, preprints, and conference publications), grey literature (scientific magazine articles, editorials, and institutional reports from international organisations), and pre-published material, all written in the English language.

Explorations of the potential use of LLMs

Not until the release of open-access LLMs, such as GPT-4 Turbo, Gemini 1, and Claude 3, have LLMs featured as tools for academic research (Liang et al., 2025; Liang, Zhang, Wu, et al., 2024). Currently, their uses vary and extend inter alia to academic writing, automated language, plagiarism detection, format compliance, and preliminary evaluations of research quality and significance (Carobene et al., 2024; Doskaliuk et al., 2025; Joachim et al., 2025; Liang et al., 2025; Liang, Izzo, et al., 2024). Extended use has prompted evaluation of their potential and limitations. Today, the literature examining these issues is inconclusive, with only a modest number of empirical studies having been conducted and published. Notwithstanding the paucity of an extant literature, what empirical evidence there is of LLMs is valuable to ascertaining their value proposition for the REF.

In respect of the REF submission processes, LLMs are advocated in supporting/managing tasks that demand the **processing of large amounts of research data**. In the first case, based on the foundations of Natural Language Processing (NLP) and Machine Learning (ML), LLMs are assisting with scoping and systematic reviews of the literature. In the case of systematic review assisted by GPT-4, one study found positive results during the data extraction process, which represents “arguably the most complex and time-consuming stage of any review” (Scherbakov et al., 2025: 1079). The authors concluded that “(...) in this review, LLMs achieved remarkable results in accuracy, making it possible to delegate time-consuming phases of review to LLMs (...) and human effort should be redirected to supervision of the review process.” (Scherbakov et al., 2025: 1980).

Similarly, positive results were found in a study comparing literature reviews from GPT-4 and humans. In this study, the GPT-4 model “generated results much faster, provided an impressive breadth of content, and was reasonably accurate” (Mostafapour et al., 2024: 7).

The same was true in a study that explored the performance of nine open-access LLMs in the task of abstract screening in systematic review and meta-analysis studies, reporting that GPT-4 showed "stellar performance, achieving an accuracy of at least 85% (...) [attaining] sensitivity and specificity rates ranging from 80% to an impressive 95%" (Li et al., 2024: 15). Despite such positive outcomes, these studies also report concerning limitations of the tools such as for instance inaccurate responses and fake references (hallucinations) (Mostafapour et al., 2024; Scherbakov et al., 2025).

More positive results of LLM use are reported in a recent study using a purpose-built medical-specialised LLM called LEADSInstruct (Wang et al., 2025). This model was trained with 633,759 curated samples, including systematic reviews, clinical trial publications, and clinical trial registries, to carry out tasks such as study search, screening and data extraction (Wang et al., 2025). Compared to generic, open-source and specialised medical LLMs, LEADS consistently outperformed them across the six validations undertaken in the study. LEADS shows that "when trained on high-quality, curated data with a tailored training process, smaller models can surpass much larger generic models in domain-specific tasks" (Wang et al., 2025: 9). What is more, the study showed that the collaboration human-AI tool for literature mining tasks resulted in time savings, recall improvement, and recall increase for more challenging review topics, data extraction efficiency and accuracy, compared to tasks developed by only humans (Wang et al., 2025: 10). The study emphasises the importance of the quality of the training data and rigorous expert human oversight.

LLMs have also been used for **academic writing**. An analysis of LLM-modified content in academic writing across 1,121,912 preprints and published papers from arXiv, bioRxiv and the Nature portfolio reveal an increase in LLM use after the release of ChatGPT. According to the authors, certain groups of researchers are more likely to engage with ChatGPT for assistance with academic writing. For instance, a larger and faster growth of their use was detected in computer science papers; in papers whose first authors share preprints more often, aligned to more competitive or crowded research fields; and papers of shorter lengths (Liang et al., 2025). Also, Liang et al. (2025) have observed higher estimated usage rates "in bioRxiv papers from regions with lower populations of English-language speakers, including China and Continental Europe, compared with those from North America and the UK. This difference may be attributed to authors using ChatGPT for "English-language assistance" (Liang et al., 2025, 6-7), particularly for 'polishing' writing by multilingual scientists. This is consistent with research suggesting that artificial intelligence can support non-native English-speaking researchers in the clarity, style, and coherence of their scientific writing (Giglio & Costa, 2023).

In addition, the authors found more use of LLMs in "abstracts, introductions, related works and conclusions compared with the experiment and method sections [...], [suggesting] that researchers may be more comfortable using LLM for summarisation tasks, such as writing abstracts, which traditionally provide a concise overview of the entire paper" (Liang et al., 2025: 7).

Explorations of the potential use of LLMs in research assessment processes

Regarding tasks related to research assessment processes, current research explores the use of LLMs for **reviewer selection**. A mixed-methods study evaluating the efficacy of artificial intelligence-assisted reviewer selection in academic publishing by comparing GPT-4-generated recommendations with traditional selection methods undertaken by experienced journal editors from ten academic disciplines found three advantages: "AI significantly enhances efficiency, resulting in substantial time savings during the reviewer selection process; (...) it expands the reviewer pool by identifying relevant experts beyond editors' immediate networks; (and it) ensures consistency by applying uniform criteria across all manuscripts" (Farber, 2024: 7). The study also warned about the potential limitations of GPT-4 when conducting the reviewer selection process, which includes AI's understanding of interdisciplinary work, methodological alignment and theoretical perspective, hallucinations in the form of fictional reviewer suggestions, and bias in the form of over-suggestion of senior researchers, researchers from well-known institutions in North America and Europe (Farber, 2024: 6).

Other studies have explored the use of LLMs for **initial screening and consistency checks of research to be evaluated**. A literature review examining perceptions within academia about the uses, benefits, limitations and ethical implications of integrating AI peer review and academic publishing identified potential for streamlining repetitive tasks. These tasks include, for instance, correcting language issues (flagging grammatical errors, spelling mistakes and awkward phrasing); detecting inconsistencies in terminology, references and data reporting; and identifying potential ethical issues (plagiarism or data manipulation) (Doskaliuk et al., 2025: 4-7). A similar suggestion was made by another study that compared manuscript review performances of four LLMs and human reviewers. The study found that AI models differ systematically from human reviewers in assessing research novelty and methodological rigour, and in rejection rates, highlighting "that these systems currently lack crucial aspects of expert judgment." (Joachim et al., 2025: 1049). For this reason, the researchers suggest that AI may be ideal for "identifying content suitable for specialised journals, supporting their potential role in initial manuscripts screening (...), (and) technical

checks, while reserving critical novelty and methodology assessments for human reviewers." (Joachim et al., 2025: 1048).

Regarding the use of LLMs for conducting academic review processes, there are a few empirical studies. A study conducted to evaluate ChatGPT's ability to **develop editorial decisions and produce peer reviews** for eleven surgery-related manuscripts found that "ChatGPT can accelerate the peer review process by conducting an initial article analysis and providing the human reviewer with suggestions for improvement or rejection" (Marrella et al., 2025: 3). The study also identified several concerns regarding the lack of specificity, the generation of hallucinations, and the reproduction of human bias, suggesting the importance of human oversight, training data, and prompt engineering skills (Marrella et al., 2025: 4).

A large-scale study comparing GPT-4-generated and human-generated peer review feedback of scientific and conference papers found that LLMs can produce valuable and timely expert feedback for authors seeking constructive feedback and recommendations to improve their manuscripts (Liang, Zhang, Cao, et al., 2024). The study showed that GPT-4 and human reviewers' overlap rate is comparable with the overlap between two human reviewers and revealed that the main limitation is the AI's lack of "ability to generate specific and actionable feedback" (Liang, Zhang, Cao, et al., 2024: 10).

Another study comparing ChatGPT-generated and human-generated peer reviews found that "ChatGPT consistently exhibited low levels of agreement with human reviewers", "(it) cannot reflect the selectivity of a journal or grasp the intricacies of the editorial vision for the journal", "(and that) despite its promptness in generating reviews for submitted articles andits potential as a tool to enhance the quality of reviewer feedback, it falls short of being suitable for the review process as a whole" (Saad et al., 2024: 3). The authors conclude that in its current form, ChatGPT cannot be used to substitute the peer-review process, but "integrating a ChatGPT-like tool as an adjunct in the review process does present some potentially significant opportunities to streamline and enhance manuscript evaluation" (Saad et al., 2024: 4). According to the study, ChatGPT can offer feedback to improve "manuscript clarity, organisation, writing style, (and can contribute) to plagiarism detection, ensuring research originality and integrity" (Saad et al., 2024: 4).

Regarding the possibility of ChatGPT **predicting peer review decisions**, a study explored original submitted manuscripts alongside the peer review outcomes available in F1000 Research, the International Conference on Learning Representations (ICLR), and SciPost Physics (Thelwall & Yaghi, 2025).

The study found that “averaging multiple ChatGPT predictions is more effective than relying on individual predictions when assessing quality-related aspects of academic publications” (Thelwall & Yaghi, 2025: 14). The authors also report that the latter was only true for two (ICLR and SciPost Physics) of the three platforms evaluated, and that the best predictions were obtained when the full text of papers was processed. The authors suggest that ChatGPT should not be used for actual peer review tasks. Instead, ChatGPT predictions might be useful in cases where, for instance, “human experts disagree and a decision must be made, such as for the final few papers that should be accepted for a conference, or where two editors agree that a journal submission is basically sound but disagree on whether is meets the threshold for acceptance” (Thelwall & Yaghi, 2025: 16).

Some studies have explored the **potential use of LLMs for the REF**. Using a sample of 200 articles from 34 Units of Assessment (UoAs) in the REF2021, a study used ChatGPT-4o-mini to estimate the quality of journal articles based on their titles and abstracts. The study compares ChatGPT and departmental average scores and found that an almost universally positive correlation between scores, with relevant variations between disciplines: “the correlations in the current article are higher for 28 out of 34 UoAs and substantially higher in many cases” (Thelwall & Yagui, 2025: 17). The authors suggest that, if proceeding with caution, ChatGPT can be considered “in contexts where citation data is currently used to support post-publication peer or expert review, it would be plausible to supplement or replace it with ChatGPT estimates, including for fields where citation data is close to useless and for articles that are too new for citation analysis” (Thelwall & Yagui, 2025: 19).

Finally, another study exploring the potential use of ChatGPT in the evaluation of more than six thousand **REF 2021’s impact case studies (ICS)** found that, even when departmental REF average and GPT average score are highly correlated, the “scores are not high enough to consider replacing expert evaluations of ICS with AI evaluations” (Kousha & Thelwall, 2024b: 15). Thus, the authors recommend that ICS scores can perform a supporting role, by “cross-checking expert scores or as a second opinion or (together with feedback) to support internal university reviews of potential ICS submissions” (Kousha & Thelwall, 2024b: 15). If this idea is pursued, the authors recommend not to use GPT score in mixed academic discipline areas. In both studies, because of the approach taken (e.g. the use of titles and abstracts rather the full text for analysis), it should be clear that ChatGPT is providing an “intelligent guess” rather than fully assessing the research quality of the documents analysed.

Identified patterns in the literature reviewed

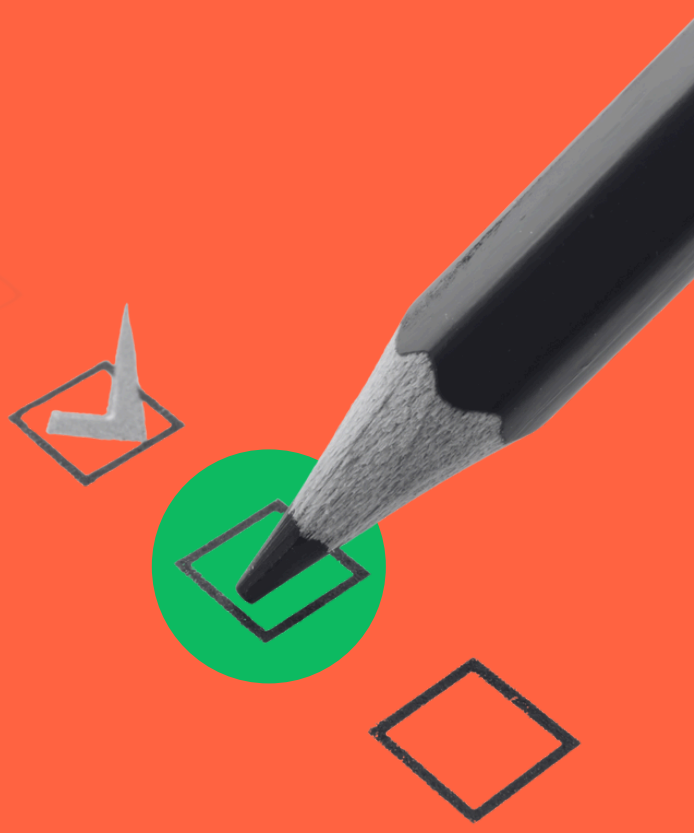
Scholars investigating the potentials and limitations of the use of LLMs for academic research evaluations, agree on the following points.

First, it seems that there is an explicit universal agreement over **the role LLMs** might play in research assessment processes, either at the submission and evaluation stages: LLMs can be a useful supporting tool for researchers, editors and reviewers. There is no evidence in the literature reviewed that supports the complete replacement of humans throughout research assessment processes.

Second, while this review highlights the existing and potential uses of LLMs in many stages and tasks involved in the research assessment process, **LLMs carry several limitations**. The most important ones seems to be their lack of capacity to deal with and produce academic expert analysis (Bhattacharya, 2024; Farber, 2024; Joachim et al., 2025); a tendency to hallucinate (Mostafapour et al., 2024; Scherbakov et al., 2025) and reproduce human bias (Bentley et al., 2025; Thelwall & Kurt, 2025; Vincent-Lamarre & Larivière, 2021; Ye et al., 2024); a lack of transparency (at the backend of the models) (Mostafapour et al., 2024; Thelwall, 2024); and pronounced ethical and legal concerns associated with their application (intellectual property, originality, manipulation, data licensing, among others (Carobene et al., 2024; Thelwall & Kurt, 2025; Ye et al., 2024).

While some of these limitations have been discussed in the literature, suggesting the need for bespoke closed-LLMs, specific training data, prompt engineering training for researchers, editors and reviewers, and clear guidelines (Latona et al., 2024; Liang, Zhang, Cao, et al., 2024; Wang et al., 2025), there is no evidence that these or other solutions might fully reduce these issues. For this reason, **human oversight** with specialised training has always been recommended (Doskaliuk et al., 2025; Farber, 2024; Kousha & Thelwall, 2024a; Mann et al., 2025). Finally, what is clear in the literature, is the continuous interest of the research community to explore the potential of LLMs to deal with the serious existing issues and challenges faced by peer-review research assessment processes.

Survey findings



The survey (sample size 386) was carried out between April and June 2025. Considering the valid values, the sample is organised as follows. It comprises **58.5% of academics and 34.5% of professional services staff**. 40% of the total sample hold the title Professor position, and 37% hold the title Associate Professor one. **53.1% of the total sample reported having a formal role in preparing for the REF2029.** Regarding the disciplinary areas, **the majority of respondents were from Arts & Humanities** (36.5%), followed by STEMM (19.3%) and Social Sciences (16.0%)[4]. **The majority of respondents belong to a Russell Group University (39.7%)**, and only 23.8% reported affiliation to a Post-1992 University.

The substantive survey questions asked about the perceptions of the use of GenAI tools for the REF2029. A summary of these responses is shown in Figure 1. While **the majority of respondents disagree with the use of Generative AI tools for the REF2029**, differences were found between statements. **The majority of positive responses were found in the calibration and validation of assessment outcomes by REF2029 panels** (15%), in supporting **the development of REF2029 People, Culture and Environment (PCE) statements by universities** (22%), and in supporting **the development of REF2029 impact case studies** (23%).

Significant statistical differences were found in three categories: job family, type of university, subject discipline and whether the respondent had previously used Generative AI tools for work purposes. Table 1. shows the total percentage of respondents who disagree with each of the proposed statements. Based on the survey sample, respondents in an academic role from Russell Group Universities, from Arts and Humanities domains, and who have never used Generative AI tools for work purposes showed more disagreement towards the use of Generative AI tools for REF2029 than other respondents. Respondents in professional service roles and from STEMM disciplines were less likely to disagree with their use than their peers.

[4] 'Other' disciplinary affiliations constituted 28.3% of the total sample.

Figure 1 Survey. Results (in % response) for question: “Generative AI tools should be used...”

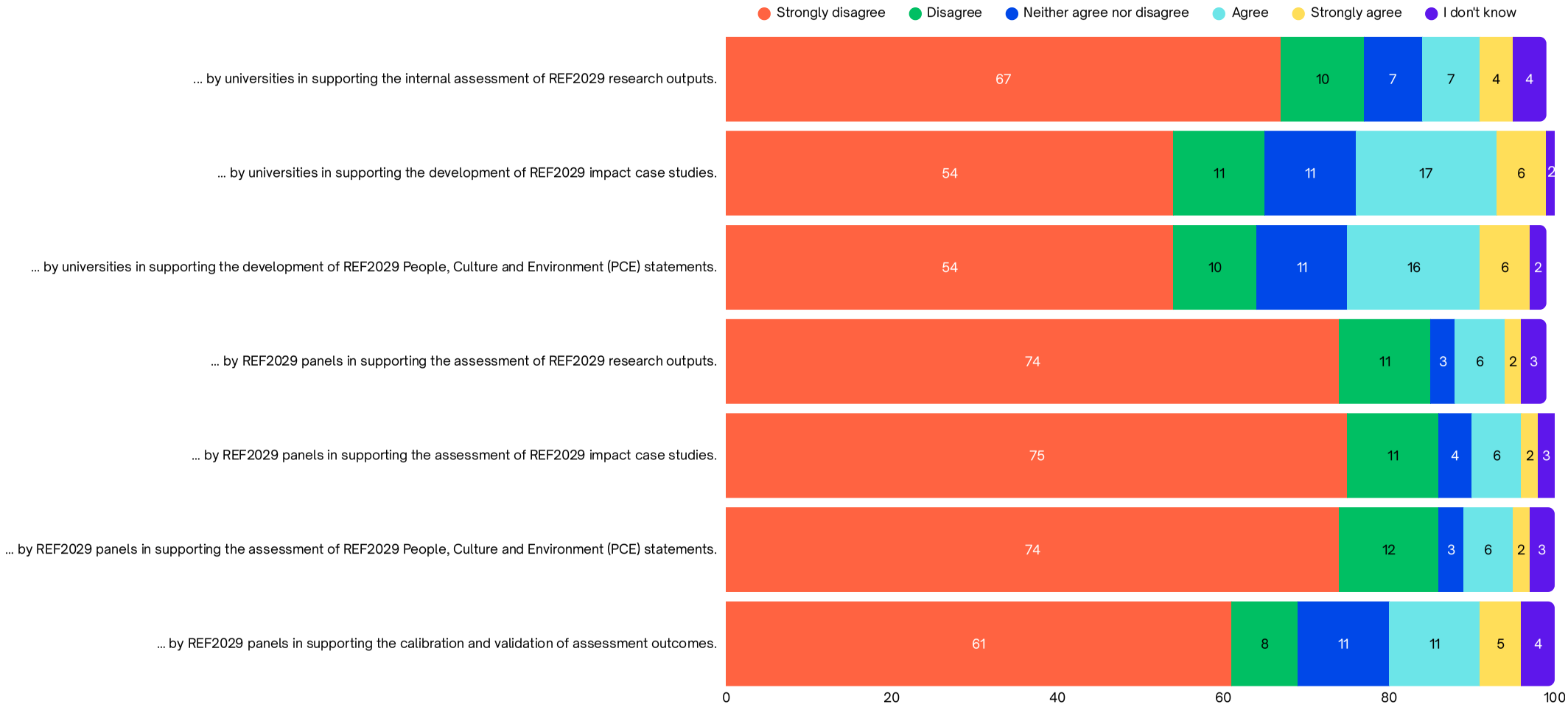


Table 1 Disagreement with the use of Generative AI tools for REF2029. Responses to the statements starting “Generative AI tools should be used…”. Total percentages shown (%)

Statement	Job Family		Type of University		Academic Discipline			Use of Generative AI tools	
Generative AI tool should be used by . . .	*Academic*	*Professional Services*	*Russell Group HEI*	*Non-Russell Group HEI*	*STEMM (1)*	*Social Sciences (2)*	*Arts & Humanities (3)*	*Have never used GenAI tools for work*	*Have used GenAI for work*
… universities in supporting the internal assessment of REF2029 research outputs.	85.8**	63.9**	82.2**	72.3**	63.8**	79.5**	95.5**	95.9**	66.4**
.. universities in supporting the development of REF2029 impact case studies.	77.0**	42.1**	70.5*	60.4*	56.5**	68.4**	88.8**	94.5**	45.9**
… universities in supporting the development of REF2029 People, Culture and Environment (PCE) statements.	77.6**	41.4**	70.7*	60.9*	58.7**	65.8**	91.0**	92.4**	47.9**
.. REF2029 panels in supporting the assessment of REF2029 research outputs.	91.9**	73.7**	90.7**	81.2**	76.1**	84.6**	98.9**	97.2**	77.4**
.. REF2029 panels in supporting the assessment of REF2029 impact case studies.	92.3**	73.7**	88.6	83.3	78.3**	84.2**	98.9**	97.9**	77.7**
.. REF2029 panels in supporting the assessment of REF2029 People, Culture and Environment (PCE) statements.	93.2**	75.2**	89.3	83.8	80.4**	84.2**	98.9**	98.6**	78.1**
.. REF2029 panels in supporting the calibration and validation of assessment outcomes.	80.9**	51.1**	75.0*	65.4*	60.9**	63.2**	93.1**	91.7**	55.2**

Note: "Total disagreement" to each statement is calculated by considering the sum of "Strongly disagree" and "Disagree" responses. Significant differences are reported as follows: * p<0.05, **p<0.01. Academic disciplines were clustered as follows: (1) STEMM = Agriculture & Related Subjects, Architecture, Biological Sciences, Computer Sciences, Engineering & Technology, Mathematical Sciences, Subjects Allied to Medicine. (2) Social Sciences = Business and Administrative Studies, Law, Social Studies. (3) Arts & Humanities = Creative Arts & Design, Education, Historical & Philosophical Studies, Languages, Mass Communications & Documentation.

Interview findings

Normalising GenAI for the REF

1. Our interviewees tended to exhibit **poor institutional knowledge** (at strategic and executive levels) of the extent of use of GenAI tools, primarily among academics for purposes of REF output review (but suspected that use was widespread):

> *We don't have sight in terms of what's going on. So, for instance, if we're doing output review, we don't know how individuals who have been asked to peer review are using them . . .*
>
> *(REF-Lead RG Institution)*

2. In some institutions, where GenAI use was discernible to PVCRs and REF-Leads, interviewees sensed that GenAI tools were mainly being **explored at the 'margins'** and were **far from being mainstreamed** into research (and indeed) REF practices:

> *We've seen some mobilization across this university, both in researching AI or thinking about researching AI and then engaging with kind of ethical practices in AI, which has floated into the research space. But what I don't think we've seen yet is that kind of transformative potential of where AI might be being built into research bids, research projects as part of kind of future practice . . . I think we're seeing it at the margins, almost kind of a very pragmatic way of just kind of playing around with. I think we're not seeing it as part of mainstream research practice.*
>
> *(PVCR Post-92 Institution)*

3. Despite a lack of clarity on the pervasiveness of GenAI tools for research and REF purposes our interviewees acknowledged that the application of GenAI was being explored (in some institutions more tentatively than others) in respect of REF-related tasks, including **output selection, narrative development, and administrative tasks** like taxonomy classification.

4. While many of the institutions we visited were characterised by **low levels of organisational maturity in the application of GenAI tools for REF purposes**, a few bucked the trend, though these were institutions identifiable for being larger research intensive and ostensibly better REF resourced institutions. **Positivity towards GenAI tools among HEI executive leadership** also emerges from the interviews as a significant factor influencing the extent of experimentality with and resourcing for GenAI tools within institutions.

We found for instance one research intensive institution which had shifted from a cautious to much more experimental approach in the application of GenAI tools in the last 6-12 months, as driven by new leadership and strategic vision. Clear institutional sanctioning and support structures have been established, including Copilot licenses, pathfinder projects, and AI apprenticeship training programs for staff.

5. In some institutions with large REF submissions, we found **previous use of algorithms** (in REF2021) having been deployed for purposes of output selection. We also found from the interviews, evidence of the development of in-house GenAI tools for REF purposes. In one institution, an LLM had been developed which had largely achieved proof-of-concept in undertaking the various parts of REF assessment (at the institutional preparation stage).

6. A general sense from interviews was that **experimenting** with GenAI technology is key to HEIs retaining their **institutional competitiveness**:

> *This is the future. This is going to change our operations. We need to lean into it . . . We need to keep our institutional competitiveness. We've got to understand it and experiment a bit.*
>
> *(REF-Lead RG Institution)*

7. In a few institutions we found interviewees reflect a **growing appetite** for the deployment of GenAI tools for research and REF purposes:

> *We have a huge appetite in some areas of research who are already doing it, want to do more, want the university.*
>
> *(PVCR RG Institution)*

8. We also encountered broad consensus from our interviewees that HEIs would be **mistaken if they were to 'shy away'** from GenAI use for REF and wider research purposes:

> *I think we have to lean into them, not shy away from them. I think both in terms of generative and agentic AI. I think we as a sector need to lean in and embrace them and of course be critical. Not critical as in negative, but I think to be critical consumers.*

[Continued on the following page]

> *And I think it will change and transform the ways in which we do our work. And I do think that to just put our heads in the sand and say it's not going to happen or not on our watch I think is very limiting of what the future might look like . . . I think there's a lot of moral panic about and I think we're not going to stop it happening. So, it's how do we work in the academic space? How do we work with and influence in order to both improve our practice and be efficient and effective with our academic results?*
>
> *(PVCR Post-92 Institution)*

9. Application of GenAI tools was advocated by institutional REF-Leads for handling the **drudge dimensions** of REF preparations like for instance, 'output taxonomy data' (REF-Lead RG Institution).

10. Interviewees reported significant **variation in respect of GenAI acceptance and adoption across academic disciplines**, with science-based subjects reported to be more receptive than humanities and practice-based subjects.

11. Disciplinary variation was also reported in the context of **experimentality with GenAI tools.** An interviewee in one RG institution reported that there is significant appetite for greater application of GenAI for REF purposes in disciplines which fall under the REF's Main Panels A and B, with more modest appetite in Panel C disciplines. Panel D disciplines were seen to be 'behind the curve'.

12. Variation of technology acceptance at the UoA level was further expounded upon by one REF-Lead in respect of some panels' avoidance of supplemental tools to peer-review:

> *I can't see it [GenAI] supplanting peer review. Even augmenting peer review might be difficult in some panels. For example, some panels don't even take use of citations at the moment. So, for those panels, that's going to be a massive leap when they don't even consider what is, you know, standardized metrics and they only use peer review.*
>
> *(REF-Lead RG Institution)*

13. The **black box nature** of GenAI tools was propelling some institutions to develop their own tools, through which decision-making processes are **made transparent.** At time of writing we are also aware of the development of 'ethical' AI agents which offer transparency to an LLM's 'decision-making' and consequently elucidate and by extension potentially mitigate concerns of bias that are unavailable in human forms of (black boxed) peer-review.

14. There was a sense emergent from the interviews that **professional service staff** (ironically despite fears of their AI replacement) are **more accepting** of the introduction of GenAI tools in REF processes than academics, and that (elevating) adoption of GenAI tools might also be explained as a **generational trend**:

> *I think probably the PS staff are more on board with change but it's just, it's such a fundamental part of kind of being an academic and kind of peer review. But it would be naive to think that it wouldn't or shouldn't change it . . . Maybe it's a generational thing. We've got new, generations that are more open to this and maybe that will accelerate it.*
>
> *(REF-Lead RG Institution)*

> *I think professional services staff are more likely to see what it [GenAI] could do. An academic wants you to sit down and read their output from cover to cover. You don't need to do that to understand where an output might sit in terms of a REF assessment. You can often pick it up very quickly from an abstract. But I know from the panels we have here to assess internally that academics want you to really engage with it. They want to know that you've understood it before coming to an outcome. So the idea that a tool could come to the same outcome as someone reading it in great detail is something they will really struggle with, particularly for things like long form outputs, practice based research, where there isn't that same association with, you know, impact factors and other kind of numerical measures that people would associate with quality. I think we're more pragmatic, we're more looking at the operation and it's not our material, it's not our work. So that personal side of it isn't there. So that's why I think we would engage with it.*
>
> *(REF-Lead Specialist Institution)*

15. The inevitability of an AI infused REF is associated with the REF having been constantly tweaked over successive exercises. The evolution of the REF in tandem with the proliferation of GenAI technologies across UK universities is seen as a given:

> *I think one of the key arguments that we have had over the last two exercises has been, "Please don't change it, please don't change it, because it allows us to actually have some consistency to understand how we're doing"... Let's just say it's something we've been expecting because that's the way that we expect REF to develop.*
>
> *(REF-Lead RG Institution)*

16. Notwithstanding it was stated by interviewees that a wholesale delegation to GenAI would cause **significant sector disquiet**, with **distrust** for GenAI tools compared to distrust in metrics as a barometer of research excellence. GenAI tools for REF purposes might only thus be accepted where configured as a **guide and as supplemental** to the REF as a process of assessment **underpinned by human expertise and judgement**:

> *I think we're still battling with metrics generally and the fact that individuals really don't actually trust citations, for example, or even category normalized citation counts. To then say we've used AI to pick our outputs for ref, I think would cause an absolute uproar. So as a guide, Potentially, yes, but not necessarily as something which we're going to completely cut out the middle.*
>
> *(REF-lead RG Institution)*

Inhibitors to GenAI adoption in the REF

17. Some of our interviewees questioned whether the present milieu is an **AI bubble**? Some considered whether rather than increasing AI adoption, the sector may become more cautious about AI use in research assessment as **unintended consequences** become apparent, similar to previous rollback on citation metrics.

> *I do just wonder whether we're just in a bit of an AI bubble at the moment, where we think everything's going to be driven by AI and suddenly over the next six, seven years, actually we're going to much have a much greater clarity of what the limitations of AI are. And therefore, rather than it going to be more driven by AI, actually we would be going, oh, there's all these unintended consequences of AI, and therefore we row back further from the use of AI in REF. And I think it's really difficult to see it could go one of two ways. I suspect we will be more suspicious of AI going forwards when people see how it shapes the agenda, shapes the argument, than more embracing of it going into Ref 20:36. I might be wrong, but I suspect it will be a little bit about, like, citations and impact factors where we see the unintended consequences.*
>
> *(PVCR Non-RG Research Intensive Institution)*

This view however may be seen to bypass the significance of metrics use in value determinations of research quality (whether licensed/disclosed or not) and the application diversity of GenAI as a multi-purpose technology that greatly exceeds the function of a digital abacus.

18. While interviewees communicated their openness to the use of AI tools for administrative tasks like matching outputs to reviewers, they were largely unanimous in articulating a **strong commitment to maintaining human peer review** as integral part of robust system of research assessment.

19. Peer-review was seen to provide the REF with its academic license to operate – and as also providing a modicum of control over the research assessment process, which AI might otherwise weaken:

> *The principles of the REF exercise were that academics designed it and it's. Okay. Not owned by it, but it's controlled by academics. If you take that away from them, I suspect it would probably be the end of REF and it would just end up being a scaling question about how much funding you get.*
>
> *(REF-Lead RG Institution)*

20. There is recognition that resistance to the adoption of GenAI tools for the REF stems more from **comfort with traditional methods** than valid concerns about effectiveness. An insistence on peer-review as gold standard in the REF was compared to a 'comfort blanket'.

21. Scepticism exists about centralized sector-wide AI tools, with concerns about **timeliness and adaptability** across institutions:

> *I think we would have scepticism that there would be a central tool that would work across the complete breadth of the sector. I think also, you know, previous ref exercises, symplectic type Things would also give us a certain level of cynicism that relying on a tool that somebody else is managing to update it in a timely manner that they say they would probably be something that we would be concerned about.*
>
> *(REF-Lead RG institution)*

22. **Trust in GenAI tools** remains a key issue inhibiting wider or fuller implementation of the tools in REF preparation processes, with a **lack of trust** linked to **low usage or experience** of using AI tools and as maybe in corollary, **limited AI literacy** (there is currently poor evidence to demonstrate a correlation between high AI literacy and low AI use):

> *We don't trust them enough in order to really start backing away from conventional tools and supplementing them with AI tools. I can see that happening but we're not at that point yet. Although I've talked about how much is going on in the university, there are still very large numbers of administrators and researchers who have had no real experience of AI yet.*
>
> *(PVCR RG Institution)*

However, there is a clear sense from participants – that corresponds to previous findings from Watermeyer et al. 2024, 2025) – that GenAI tools will **unequivocally transform research practices**, especially in institutions taking a more **proactive and experimental** approach, and that the integration of AI tools into REF processes is **guaranteed** (if likely to be incrementally phased).

23. Some of our interviewees saw GenAI as being in conflict with the REF as a **socialised process** or more specifically as a 'carefully socialised learned understanding of the process' (PVCR RG Institution). However, it is worth noting that these individuals were also individuals talking from the perspective of large and well-resourced institutions, where access to former panel members and REF intelligence is (more) readily available.

24. Many among our interviewees advocating the benefits of an existing human model of institutional REF preparations, did so while also claiming that they were also likely 'outliers', 'old fashioned in their views', or 'not fully versant in the capabilities of the technology'.

25. Some argued that while there remain many unknowns concerning the credibility of GenAI for REF purposes, the existing REF systems is at least a largely **trusted system**.

26. A **lack of precedence** in the use of GenAI by panels, was also identified as a major obstacle to confident application of GenAI tools within institutional preparations, where most institutions were said to be guided by panel behaviour from REF previous to their current preparations:

> *In terms of being able to embrace this, I think it's going to be a very mixed bag this time, particularly until institutions have some assurance as to how it might be received by the panels or what the panels are going to do with it. So, I think there's always a lag between what the feedback that comes back from panels and trickles back into the academic consciousness as to what is actually done with the information that goes into the reference.*
>
> *(REF-Lead Post-92 Institution)*

27. In some institutions, a proposal of GenAI benefitting the competitiveness of submissions is **unconsidered and untested**, though is often hypothesised in the context of impact case study development and in helping to identify evidence of impact that might otherwise be missed:

> *We've got experts who work really closely with our researchers to try and find the best impact case studies. That's a fallible system because it relies on my team talking to researchers and people are fallible. If there was a way of identifying that data through generative AI going out and coming back and suggesting an impact case study for us, would that give us a competitive advantage? It might do if it means that we find out about different case studies that we weren't aware of.*
>
> *(REF-Lead Specialist Institution)*

28. Through the interviews we established that most institutions (especially less research intensive) were at an 'early stage' of AI adoption, with initial exploration of **LLMs to supplement workflows** and **varying levels of AI literacy** across departments. AI literacy in the terms of GenAI for REF appears limited to pockets of expertise (and interest) across our sampled institutions.

29. While there is **growing appetite for AI technology in REF processes**, the exact implementation remains uncertain and will likely evolve with technological capabilities.

30. In some institutions there is **limited formal adoption of generative AI tools for REF purposes**, though there is recognition of potential future benefits, particularly around research administration and data consolidation.

31. **Practice involving GenAI tools is generally running ahead of policy or more general guidelines** for appropriate use for research/REF activity. HEIs are typically operating on a catch-up basis with their staff.

32. In all but one institution (a large RG institution) we found **no current institutional policies on generative AI use for REF** though interviewees suggested that some individual departments within their institutions were active in developing their own guidelines and approaches.

33. Institutions are focusing on producing **guidelines over policy** given the speed of technological innovation causing policy to become equally as quickly outdated:

It's intentionally guidance rather than policy. Because it needs to adjust, it needs to constantly evolve. A policy feels very, you know, kind of preserved.

(REF-Lead RG Institution)

34. In some very large research HEIs there was reported an almost total absence of centralised policy, with the suggestion that it would be **very difficult to standardize AI practice at a whole institutional level**:

We have almost no institutional policy. Unless it's absolutely necessary for legal reasons or just, you know, because it's operationally essential. We have very little sort of institution wide positions on stuff which actually makes it quite hard for people like me who run big departments to do that because I can't rely on central policies for stuff.

(REF-Lead RG Institution).

The contributions of GenAI tools to the REF

In this section we consider the various justifications put forward by our interviewees for the use of GenAI tools in both (i) institutional REF preparations, and (ii) REF panel assessments.

35. In overview, GenAI tools were acknowledged by interviewees for their potential **in reducing REF preparation burden through evidence checking, quality validation, cross-referral identification, standardizing assessment processes and providing more objective validation, particularly for non-traditional outputs; through dialogical reasoning; auditing, mapping, sense-checking and validating associated data.**

36. It was suggested that a system level GenAI tool would significantly help address what is reported to be a huge burden placed on academics in the process of reviewing outputs for REF institutional selections. However, it was also felt that on the current REF schedule there is insufficient time for piloting a tool, gaining sector-level confidence in a tool and for its systemic embedding. There is some suggestion that an existing schedule for the REF might be reconsidered to allow for AI integration:

It all comes down to the quality of the tool and the biggest risk for REF is going to be the tool being developed midway through and there won't be time to pilot it if people don't have confidence in it. We'll just be running double systems and increasing the workload. If we had a tool in a timelier way that had been developed earlier in the REF cycle, like so many bits of the REF, and had built community confidence, then it would have saved the big-time consuming part.

(PVCR RG Institution)

37. Interviewees advocated AI tools as means of reducing the labour intensity of REF assessment, for handling "heavy lifting" tasks and in undertaking initial assessment triage:

Anything which could enable you to either get your life back and or perhaps think that you're not in a situation that you know the last 40 papers that you've done in the last two hours, you know, have they all morphed into one in terms of your scoring and stuff would be helpful.

(PVCR RG Institution)

38. However, it was also considered (though among representatives of larger and better resourced institutions) that the REF should **stick to its current timeline** and that the **experimental work of integrating GenAI should stay with the panels**, whose experience would provide a vital source of learning for whatever REF (or research assessment system) follows 2029:

There isn't going to be really time for the sector in its preparation unless a standard tool can be designed very quickly and rolled out, and then the universities have got to be absorbed and embedded, but they can't take much more change at the moment, so I would focus on it for the panels and make it clear and transparent about the principles the panels are using, and let us know how they're doing it and that will tell us more about the panels than we've ever known before, frankly. And then look to the next REF.

(PVCR RG Institution)

39. GenAI tools were viewed for their potential in **levelling a playing field** for the REF and in so much as they might scaffold typically 'newer' and less well-resourced institutions (with smaller REF teams) in making submissions that might be as equally competitive as institutions with a greater wealth of finance directed at supporting REF submissions:

For research intensives, it's a whole different game. The turnovers are different, but the quality of the research isn't different and the impact that that research can have is not different. And the kind of experiences we can give researchers is different but is no lesser. So yeah, in terms of that levelling, you know, you know, if I had 12 people on to support 'Hugo', we might be thinking differently. So, we're having to think about how we best use our resource. So yeah, it would level potentially.

(PVCR, Post-92 Institution)

40. Our interviews elucidated the impact of **resource constraints** for some HEIs in respect of their REF preparations and how a lack of internal capacity (exacerbated by a sector wide trend of work intensification) meant that their **REF preparations became unmanageable** – in corollary, intimating the potential contribution of GenAI in resolving the constraints of manual tasks:

I did a REF outputs review working with the REF planning group and our unit of assessment leads and we wanted all our research active staff to submit up to 10 of their best publications that they thought would be going to, you know, meet the REF eligibility criteria. We had to extend the deadline on that three times because people just weren't able to meet it.

(REF-Lead, Post-92 Institution)

41. Interviewees routinely described AI tools as **additive rather than substitutive**. Many emphasised the continued **importance of human judgment and discrimination** in the preparation of institutional submissions and assessments of REF panels:

What we wouldn't want to do is trust entirely that that would be our guide as to what is high quality and what isn't high quality because clearly we could get it very badly wrong. And some of the stories that we are hearing about GenAI, particularly around where it kind of goes into a bit of an 'Alice in Wonderland' world where it's all kind of bit fantasy would concern me. Which is why I think we take the view that it can be a tool to help, but that we retain the cynicism to make sure that we are then not just blindly following a particular set of guidance from a generated outcome, but that we are at least putting the human view on it. I think that will be our overall approach.

(REF-Lead RG institution)

I think there is a general belief that it [GenAI] is an additional and helpful tool rather than a replacement tool.

(REF-Lead Non-RG Research Intensive Institution)

42. Some of our interviewees mooted the possibilitythat **GenAI tools might be applied selectively** and only in relation to specific aspects of the REF, for instance the proposed people culture and environment statements or impact case studies as dimensions of the REF that are ostensibly not quite so closely prescribed by assertions of peer-review emphasised in the assessment of research outputs.

43. Interviewees discussed the primary potential application of GenAI tools in reference to output ranking (at the level of institution's making their submission choices). They also considered the value of GenAI tools in ranking outputs whose value as set against the REF's assessment criteria is more ambivalent:

Each of the UOAs are going to have to make a decision on what's their top 150 and what's in scope. I think the top 50 and the bottom 50 are easy. The ones in the middle are really, really challenging. And I think there will be a natural drift for people to try and use something like AI to help them make those decisions.

(PVCR Post-92 Institution)

44. Despite concerns of the idiosyncrasy of panel behaviour and the challenges of second guessing how submissions would be scored, there was recognition among our interviewees that GenAI **might attend to issues of imbalance (dangers of partial/partisan judgements) in internal scoring and inflated/deflated internal assessments** of REF items.

45. Internal validation of scores depends on the quality and size of academic resource within an institution. Where there is **less internal academic resource, institutions are forced to go outside of their institutions for checking**, which may have a significant financial cost, that it was felt might be **compensated for by GenAI**.

46. Some of our institutional REF-Leads with prior experience of supporting REF panels spoke of how GenAI tools might be **exploited by the REF secretariat.**

47. Concurrently, while GenAI tools were recognised for their potential in assisting with administrative tasks and output assessment, some of our interviewees (predominantly REF-Leads) expressed **reservation about their use for narrative creation**, preferring that the tools might instead be oriented towards technical checks and data cleaning. Interviewees also, however, acknowledged that GenAI tools hold significant value as **early initiators of narrative development**.

48. We found within our interviewees' accounts evidence of GenAI tools being used for **checking references, sentence structure, and readability**, as well as analysing impact narratives for **evidence and narrative structure**.

> *We are going to use, and we are using some AI tools to interrogate all of our impact stories to go, you know, is there clear research evidence there, you know, clear strategy to communication and impact?*
>
> *(REF-Lead RG Institution)*

49. In reference to impact narratives, it was also suggested by interviewees that GenAI tools might be **trained on previous high scoring impact case studies**:

> *The narrative bit is interesting though, isn't it? Because if you basically trained the AI tool on all the last impact case studies and whatnot, and how they were essentially scored well, you've got your perfect narrative, and you could do it according to subject discipline.*
>
> *(REF-Lead RG Institution)*

50. A key potential benefit identified by interviewees is using AI tools to generate **coherent narratives from disparate data sources**. Our interviewees discussed how HEIs face challenges with data being dispersed across systems and consequently having to rely heavily on institutional memory:

> *I think that the benefit [of GenAI] would be in helping us to put a coherent narrative together and not miss anything. The difficulty we have is hundreds of different points of data or information coming from all different sources that some of which are verified and some of which aren't, and they're hugely reliant on institutional memory . . . I think certainly there's a plethora of potential benefits and using GenAI in development of impact case studies and because again, they rely on memory or interpretation or value judgements about what's worth pursuing and what's not, what's worth mentioning and what's not.*
>
> **[Continued on the following page]**

So, I think having a safety net on all of the evidence and activity related to a particular theme of impact would be really beneficial because it does seem very risky and prone to error to be relying on people, single people, single points of failure.

(REF-Lead Post-92 Institution)

51. Where concerns surfaced in the interviews (as per the focus groups) that GenAI tools harnessed in the generation of narrative content, would produce **bland and formulaic, and even 'unscientific' prose**, such considerations neglect advancements in **AI agents which are evolving to replicate human and scientific 'voice'** and which might also be trained on the narrative construction of exemplar impact and environment narratives from previous REFs.

52. One interviewee also argued that one of the major advantages of GenAI tools in the context of narrative generation was in **harmonising multiple voices** in the development of UoA and wider institutional accounts.

53. Another key role for GenAI tools was presented in the terms of **output allocation**:

The thing that jumped out that we are hoping AI can help with is allocating outputs to reviewers. So not obviously, not getting AI to review them, but saying this is a close match with this academic expertise. And you can see how that would be helpful to REF panels as well.

(REF-Lead RG Institution)

54. Others among our sample considered that with ongoing technological developments, the capacity of GenAI tools to provide **objective review** (deprogrammed of bias, hallucinations etc) **would only improve and thus so too the wider objectivity** of the assessment process:

I think if you can have an unbiased platform or software that could actually robustly evaluate the quality of an output and lead to no kind of, of questions on, well, how the decision would work, that'd be great. And just maybe to come up with that more objective process.

(REF-Lead Post-92)

55. One of our interviewees anticipated the (future) contribution of GenAI tools both in addressing what was often reflected by our research participants as **the profligacy of the REF process** with an assessment process, potentially **free of bias**:

> *You think of the panellists and all of the time and the money, and all the other costs that are actually within that process. If it was a robust algorithm that could unbiasedly evaluate every output, that would be great. That might be a dream scenario.*
>
> *(REF-Lead Post-92 Institution)*

56. Relatedly, the use of GenAI was advocated by interviewees as an **additional layer of calibration** in the process of institutions trying to validate their internal scoring of research outputs, impact narratives, etc.:

> *We try to watch out for biases or problems by having what we call external calibration, and I think everybody around the country does that as their as their way of checking. We could look at a further layer of quality checking using GenAI tools.*
>
> *(REF-Lead Non-RG Research Intensive Institution)*

57. Discussion with our interviewees delved into technological solutionism and concerns that replacing GenAI with human expertise in the process of peer-review would set a dangerous precedence in terms of exaggerating human fallibility. However, there was also a strong sentiment evidenced that GenAI might **not replace human peer-review but extend it**, and that GenAI has an obvious role in **sense-checking and validating human judgement**:

> *It would give a different flavour . . . I think it would maybe go towards more of an objective analysis rather than a very highly subjective analysis, which is what we currently have both at the internal and the external view. You've got to remember that the panel, the panel's assessment of this is, subjective. So, therefore, it's what two, possibly three individual panel members think of a particular paper and if they all agree then happy days. But if they have somewhat of a disparate take on it, then who knows what the outcome would be. But we've never seen a parity really between the internal and external . . . So, it depends on what the internal assessors are doing. They could be, you know, having a glass half empty day. It goes in anyway and then ends up being a four-star paper depending on which way the winds blow. But yes, I would say it would give another view.*
>
> *(REF-lead RG institution)*

58. In the terms of output selection by institutions, interviewees also identified the potential contribution of GenAI in the **identification and avoidance of unclassified outputs**:

> *One of the things that we're trying to avoid is unclassified outputs. I could see using AI and trying to remove mistakes that would lead to unclassified outputs. They actually become really quite expensive in terms of reputational ranking and in terms of GPA league tables, and they're a missed opportunity because you could have put something else in that was potentially really good. And being able to check for those, it's really burdensome.*
>
> *(REF-Lead RG Institution)*

59. GenAI tools were considered by some as potentially providing an additional layer of 'reassurance' to panel decisions:

> *We've been very reluctant to not let go of peer review in many of the processes that we adopt, whether it's REF or competitive grant funding, or many other things. But there's always this kind of nagging concern about bias. And so, a check, I think it [GenAI] could be helpful and perhaps provide some reassurance.*
>
> *(REF-Lead Post-92 Institution)*

60. A concern of not using GenAI by panels as tools of sense-checking, validating scores is also predicated on scepticism that **REF peer review is not as is typically portrayed, gold standard**, and that the argument of such for excluding GenAI tools from REF processes is thus flawed:

> *Is REF peer review really this gold standard? If you've got 20 people from a subject area, do you have enough expertise to completely assess that area? I've heard people saying, "Well actually compared to something that I would peer review for a journal, I read a much wider range of things". And on a REF panel people talk about, you know having ten minutes to assess an output and having the stopwatch on. So, I don't think there's absolute confidence in the REF assessment process for outputs, as exists at the moment. So, we could be shifting into a different domain.*
>
> *(PVCR RG Institution)*

61. Some of our interviewees even described REF peer-review as an outdated and profligate process that ought to be replaced by a combination of bibliometric and GenAI tools:

> *I think peer review is on the way out . . . Do we really need to read all these outputs? The correlation, you know, we could easily find a bibliometric way of doing it, that for some subjects would gain the confidence of the sector. I think you put AI on top of that. You know we are very close I think for the sciences on output assessment to just have that done automatically. And I think the evidence or being worried about that in terms of peer review is not really there. I mean, I'm sure there'll be wrinkles but maybe you'd handle it by doing some dip test calibration so you wouldn't read everything, but some stuff would get read centrally just to check that the results are consistent.*
>
> *(PVCR RG Institution)*

62. GenAI tools were recognised as providing **finer-grained analysis of output quality at the panel level**, enabling peer-reviewers whose time constraints and volume of outputs to consider were felt to limit their ability to offer a **deep review** of the submitted work:

> *We're looking for tools that help us minimise the human side of peer reviewing now. At this stage we've deliberately done it through a community engagement model of tell us your best and we'll review. Next round, just because of volume, because we tracked the volume of prep for the last REF, we will be guiding peer reviewers to select some of that based on citations. But we're starting to explore how we might use some AI tools to help with that.*
>
> *(PVCR RG Institution)*

63. Interviewees referenced that **not all REF panellists are given time off by their institutions, providing significant unevenness of burden and individual resourcing**. GenAI tools were thus considered for stabilising the unevenness of individual capacity among REF panel members, potentially **scaffolding those without sufficient release from their home institutions** in concentrating their focus on the assessment process.

64. Our interviewees also signposted **significant variation of REF expertise across institutions.** Where in RG institutions, we found that many core staff had gone through multiple REFs (and even RAEs), in many of our post-92 institutions we consulted, there was far less experience evidenced of having supported previous REF submissions. In these institutions, **GenAI tools were recommended for compensating for gaps in REF expertise.**

65. Some of our interviewees stated using GenAI tools for the purpose of **interpreting REF guidance** – an application likely to be more prevalent in institutions with lower levels of staff experience of previous REFs.

66. Analogously, some of our interviewees identified opportunities for GenAI in **scaffolding submitting institutions' adaptation** to the various changes to REF2021 being proposed for REF2029:

> *I think it's hard to disentangle the benefits that AI might bring from the changes in the assessment exercise. So, a lot of the areas where we think it might be really useful is actually addressing the new challenges in REF 2029. So, things like the output decoupling and having to submit outputs into the correct UoA rather than it being associated with the staff member who's in that UoA.*
>
> *(REF-lead RG institution)*

67. Rather than viewing GenAI tools as catalysts for individual institutional competitive advantage, interviewees placed an emphasis on **sector-wide collaboration and shared development of GenAI resources**. Such an approach was recommended on the basis of benefitting the wider UK research ecosystem:

> *We're in a massive learning phase at the moment and we'll come out at the end of it through sharing as a sector. That's not going to be overnight, and all these institutions are developing their own things. I think we're at a point where are we putting lots of effort into individually developing stuff that could be developed as a sector to help and support us all and that we evolve together . . . [there] is so much wasted effort at the moment because we're all doing things differently and thinking we need to invest for competitive advantage.*
>
> *(REF-Lead RG Institution)*

68. However, it was considered by some of our interviewees that a system level tool would **only be necessary for output review and selection** and that other aspects of REF preparations might be farmed out to existing GenAI tools.

69. Interviewees stressed the importance that **to embargo use of GenAI for the REF would be to send it underground** and to provoke unregulated and potentially inappropriate use:

If you push it underground, it will have negative consequences and people won't engage with it fully . . . I think trying to engage and think through how we can do this together; addressing any concerns mitigating any risks we see with it and supporting a kind of responsible process and facilitating good use, I think is definitely the way to go.

(PVCR RG Institution)

70. Another interviewee who identified the inevitability of GenAI as a ubiquitous technology argued strongly that the **REF cannot be silent on AI** use and ways would have to be identified for its appropriate integration:

I think AI is here to stay and so we need to work out how we adopt it and in a way that we're comfortable with. I think being upfront about that is helpful in the way that our institution is doing and many others are. So, you can't be silent on the topic of AI in any process anymore.

(REF-Lead Post-92 Institution)

71. Interviewees highlighted how REF submissions are **data driven and therefore demanding of data analytics** in evaluating, sense-checking and cross-checking:

We really want to utilise data as much as we possibly can in planning our submission, evaluating our submission as we put the submission together . . . We're at the early stages of developing REF dashboards. So, we are going to be setting up a series of REF checkpoints from the autumn, looking at the data you'd expect us to be looking at for REF. So, income, numbers of Ph.Ds and output evaluation. We're pulling together a series of REF dashboards for that information. We also want to pull together some REF dashboards for our people, culture and environment data as well.

(REF-Lead Post-92 Institution)

72. Relatedly, GenAI was recognised by interviewees for its potential **in mitigating data errors associated with the limitations of human recall and value judgements.**

73. Interviewees identified significant potential for AI tools in helping to **consolidate disparate research data and automate manual processes**, particularly around capturing research outputs and impact evidence that currently **exist across multiple data systems**:

We have a multiplicity of data pots of information that will support the impact narrative that we want to present to the world in relation to the impact of our research on the world. That is not in one particular system. It is all over the place. I see this being a tool which could potentially help us in it, just to pull that all together into one place so we can actually visualize it in one particular area.

(REF-Lead RG Institution)

74. In recognising REF as a data-led process, some of our interviewees acknowledged that **by not adopting GenAI tools for REF purposes they risked a suboptimal REF submission**; particularly in the event that they might fail to optimise their material selection.

75. One of our interviewees described the potential contribution of GenAI tools in the terms of **institutional strategy building**:

I think AI tools could work in help helping us to understand how we're supporting a particular strategic imperative. So, for example, how we're promoting equality, diversity and inclusion? Because we, again, we do have the data . . . but what we're not very good at doing is evaluating broadly our approach.

(REF-Lead Post-92 Institution)

76. REF preparations were viewed by some of our interviewees not as a process of internal peer review but of **predictive analytics**, a function some argued would be more efficiently/accurately performed by GenAI tools (in managing and making sense of multiple data points).

77. The integration of GenAI tools into REF processes was viewed not so much as has been recently reported for doing away with REF related personnel but **improving both the efficiency and quality of their contributions by generating more comprehensive data sets.** Thus, the potential of GenAI at an institutional level is also argued for helping HEIs **getting to know themselves better through more wide-ranging data sets that are not limited by the vagaries of institutional memory**:

The person that would have been doing that task [data management] will be more effective, more useful, more purposeful, and therefore more valuable to not just the organization, but the whole exercise. So, it's not just useful for the exercise [REF] but actually institutionally useful because we will now have a better sense of how all these things are working. So, I wouldn't say it will be like, hooray, we can save some money . . . I think we will actually probably still have a person, but their work will be more meaningful.

(PVCR RG Institution)

78. GenAI was viewed by interviewees as providing **a necessary scaffold to institutional memory** and in respect of staff turnover affecting recall of what's happened within an institution over a prolonged period of time:

I think it can be a definite aid in terms of institutional memory. That's one of the big kind of costs of REF, particularly around outputs, but also when you're trying to describe an environment across a 7-year period and how it has evolved. And I'm thinking here of institutions where they will have had significant turnover of staff and significant loss of staff over the period. Then being able to essentially, ask AI to ask the internet what had happened at their institution is actually a cost saving compared to you trying to find someone who remembers.

(REF-Lead RG Institution)

79. Interviewees argued that the use of GenAI tools for selecting and assessing outputs was seen as **a necessary step forward in institutions making large submissions**.

80. There was a major concern that by not leaning into GenAI, **REF would appear on an international stage to be an outdated system of research assessment**:

If we don't use AI, we're going to look like a very old-fashioned assessment exercise, which doesn't really demonstrate the dynamism of a community that is producing cutting edge research but won't use one of the tools that's coming out of its cutting-edge research.

(PVCR RG Institution)

81. There was strong opinion that the REF had entered a period where it **has to be less human in its implementation, more agile and efficient** in terms of how it reports research excellence and provides a much **more accurate and up-to-date account of research** across the UK:

> *I do think we are at a moment where it's got to be less human, more automated. We did quite a bit of this discussion with FRAP[5] and fed all this in. There was anticipation then about how much can we automate if we took outputs out and did that as a rolling cycle. You could then focus on the impact, the environment, the other bits in a lighter touch, and perhaps even more frequent exercise if the output stuff was just running along in the background. And if we can get a machine to do that for us, that would be really great. I think that really would open up REF.*
>
> *(PVCR RG Institution)*

82. An argument was also made for GenAI **enabling a faster feedback loop so that QR was invested into research areas that were demonstrating efficacy/success** (and as part of institutional review processes) not on historical but current data, and therefore **REF as a streamlined annual or biennial audit**:

> *I think the effort and hoo ha and focus that we put on the REF will gradually become diluted if it's an annual process or by once every two years, and it becomes more automated. It will become a bit like maybe the reports that we get on our access agenda for undergraduate access or other sorts of business that we do with the Office for Students. I think also in reality, QR has become a reducing fraction of university budgets, and if it became more algorithmical more frequent, I think that again would support a gradual sort of less of a big deal around the REF process. What would it mean for university strategy? You might be able to have a slightly stronger feedback loop on the actions that you take and whether they're making any difference. You'd get a faster feedback on whether that investment is paying off in terms of your quality and your QR, rather than having to wait seven or eight years to see whether it's working.*
>
> *(PVCR RG Institution)*

Mitigating side-effects of an annual review in terms of QR distribution was discussed on the basis of a financial settlement organised as a rolling average.

[5] The FRAP of Future Research Assessment Programme was undertaken to 'to explore possible approaches to the assessment of UK higher education research performance': Other' disciplinary affiliations constituted 28.3% of the total sample.

83. During a period of particular financial stress for the UK's HE sector it was stated by interviewees that GenAI would provide **maximum value extraction from limited resource in the form of REF panels**:

> *It's all about using the resource effectively. So, you know, if you've got a reduced academic resource across the sector, you point that in the direction where those academics have the most value. The same thing holds for REF panels, surely.*
>
> *(PVCR Post-92 Institution)*

84. Some of our interviewees described the contribution of GenAI to the REF as **a necessary disruptor**, that might provoke disquiet from certain parts of the sector yet provide a necessary jolt in reforming the REF:

> *There will be lots in the sector who won't be happy, but that's because they dragged their feet on everything frankly, which is partly why we haven't agreed half of the stuff we should have done by now. So, I think go for it. I think we need to disrupt a little bit how we do REF. If we disrupt transparently and we give everyone the same tool, then we're satisfying principles of open research practises. We apply those principles to how we do REF. I think that gives us a legitimate basis. But there will still be people who squeal. I'm tired of squealing. Let's just get on and do it.*
>
> *(PVCR RG Institution)*

85. A need for reform was also registered by our interviewees in acknowledgement that despite changes in the overall composition of the REF in terms of what it assesses, the **exercise is for the most part largely unchanged**:

> *"30 or 40 years ago, people had to do it in a manual, very straight forward way and basically, that's what we still do."*
>
> *(PVCR RG Institution)*

86. Despite many of the espoused benefits of the REF to institutional research culture and behaviour, interviewees expressed their concern that in its current (analogue) form, the **REF exerted a handbrake on the UK's cutting edge 'science'** – and in corollary influence as a research power – by co-opting leading researchers into the hugely burdensome work of panels – an obligation they might be freed from or otherwise at least alleviated through the greater infusion of GenAI tools:

I think a whole bunch of countries think we're mad. "You take your leading researchers, and you take them out of research for a year in order for them to evaluate a whole bunch of things that people have already evaluated? Are you mad?" So, you know, I think that while I agree that lots of places think that this is a great way to do it, there are a lot of other places that would not want to do it themselves. And I think the technology might help with that, because if you said in actual fact this is going to somehow reduce the amount of time a panel has to take to do this, so that they actually can keep doing their research and we haven't taken all these excellent researchers away from doing their research. I think that might actually come across as being a more sensible thing to do.

(PVCR RG Institution)

Conditions for GenAI adoption in the REF

While our interviewees identified a myriad of ways in which GenAI tools could contribute to the REF, they also applied a series of conditions which extend to a series of guardrails for the appropriate use of GenAI tools.

87. If there was to be a motto common to institutions in response to GenAI use for the REF (and wider research processes) it is 'cautious, judicious, embracing'.

88. A perspective of AI as neither 'evil incarnate' **nor a complete solution** emerges strongly from our interviews. Interviewees instead spoke of appropriate and responsible integration of GenAI/LLMs into REF processes; urging the establishment of explicit guidelines for GenAI use in the REF, while also stressing the need for **human responsibility and oversight** in the process of AI adoption and integration.

89. It was felt that effective use of GenAI by panels hinged on the degree to which they could **trust the tool** and that such trust would be u**nderpinned by training**:

It's going to be building the trust of a panel. And that will come down to individuals how much they're trained.

(PVCR RG Institution)

90. Interviewees also considered that future AI implementation in the REF would require **clear governance frameworks and transparency in its usage** – incorporated into institutional codes of practice – that would also be responsive to different institutional contexts.

91. There was an appeal among interviewees for a **REF standard tool** and one that would avoid some institutions finding an advantage in making their submissions by being better REF-resourced and/or benefitting from in-house expertise:

> *We need a REF standard [GenAI] tool. We don't need everybody customising and producing their own.*
>
> *(PVCR RG Institution)*

92. In fact, the very efficacy of GenAI for REF was seen to rest with the development and distribution of **a standardised sector tool enabling standardised practice** in the application of GenAI tools for institutional REF preparations and panel assessments:

> *The more we have a transparency and a shared tool it will help the sector. We're a sector of different sizes. You know, my internal reassurance to my teams is kind of whatever we get to ask to do, we'll be able to do because we're big and we've got lots of bright people here and we've got teams we can pivot. But the conservatoire next to me can't. So yeah, tools which help people and give reassurance across the field and proscribe people from using other tools. You've got to use the standard.*
>
> *(PVCR RG Institution)*

93. Interviewees argued for **complete transparency** in how panels might potentially use GenAI tools on the basis of a strong expectation that panellists will. In this respect, many of our interviewees, yet not, it should be said, all, considered that REF panellists had insufficient time or lacked capacity to read everything asked of them, intimating therefore the value of GenAI in providing necessary short-cuts within the assessment process:

> *Nobody believes the panel reads everything. They don't have the capacity, if you take the time involved in the volume. So, let's not pretend everything is read thoroughly. Let's get an AI tool in to help us do the first batches. And we'll dipstick, sample and tweak the results accordingly.*
>
> *(PVCR RG institution)*

94. **Transparency** about how AI tools are being used and **clear accountability mechanisms** were relatedly advocated as being essential for the preservation of integrity in the assessment process:

> *It's about transparency. It's how are using this? How are checking that it is being used in the right way? And are we all comfortable with that.*
>
> *(PVCR RG Institution)*

95. There was another appeal that the sector needs to bypass those 'dragging their heels' on GenAI and that a more progressive stance was necessary to securing an open and transparent research (assessment) processes.

96. However, at panel levels, interviewees considered that variation of practice (and personalities in the leadership of panels) would mean that **standardisation of GenAI across the REF would prove challenging**:

> *I supported two panels last time. I would say that one might have embraced it, whilst the second would have had more hesitation . . . And I guess some of this is also going to depend upon the personality of the chair and deputy chair of the panel and where they get to informing the panel's own sub panel's criteria. And panels have their own nuances and their own criteria that they set over and above the generic criteria and that will be quite a hard one for them to shift. So, I'm guessing standardising it across all the panels is going to be very tricky.*
>
> *(REF-Lead Post-92 Institution)*

97. 4.1 Interviewees identified that the value of GenAI as a tool for evidence reconnaissance and collection is dependent on the **quality of institutional data systems** and is simultaneously compromised by academics being characteristically poor at record management.

98. Special mention was also given to the need for rules for panel membership as relates to **institutional buy-out**. This was argued on the basis that not all panel members receive relief from their institutional duties while serving on REF panels. While the integrity of panellists was routinely defended by our interviewees, it cannot be ignored that GenAI tools may provide a necessary coping mechanism for those struggling to concurrently manage a surplus of REF and institutional demands.

99. **Human oversight** and/or integration into all AI infused processes was seen to be at this stage (of AI capability/proof of concept) **a vital dimension** and one which would differentiate the REF from a purely data driven process. Human responsibility and accountability in use of GenAI was also touted as key condition of use:

> *I am totally comfortable with people using it as a tool, assuming that there is a way that they ultimately take responsibility if they use that tool wrongly.*
>
> *(PVCR RG Institution)*

100. While interviewees viewed AI integration across research funding processes as inevitable, they asserted the **integralism of human judgment in peer review and cautioned against a reduction of research assessment to purely algorithmic decisions**:

> *I think there definitely will be ways that I think the process could be easier and the burden lifted which would be in, in, in keeping with what they want to do. But I think as you say, that fundamental of peer review on the assessment side, I can't really see that changing unless the whole thing gets scrapped and turned to an algorithm to just distribute funding and we don't submit anything.*
>
> *(REF-Lead RG Institution)*

101. Where there is a common view that GenAI tools will be used by panels in the assessment phase of REF2029, there was strong feeling among interviewees that this should have good **symmetry with GenAI use in the preparations phase.** However, there was a sense that this may not happen and consequently cause **some UoAs to enjoy unfair advantage over others**:

> *If the panels are going to be using AI to support their decision making, we should be using AI to support our decision making, so we're judging it on the same basis. So, I think it will be diffuse in this way, and I think some subjects, some UoAs will be ahead. So, no doubt the computer scientists will be ahead on this, and you know, modern languages will be in a very different place. Arguably the selection process should be operating to the same protocols that the assessment process is going to otherwise you know, you're not picking the right stuff.*
>
> *(PVCR RG Institution)*

102. Correspondingly, interviewees identified urgent need for universities to receive and benefit from panel criteria and working methods in relation to GenAI use to cater for greater complementarity between submission and assessment phases:

> *We don't have any panel criteria and working methods yet, so from my perspective it would be really good if the working methods were specific enough to talk about how they expected the panel members to use AI.*
>
> *(REF-Lead RG Institution)*

103. GenAI adaption was related to resource capability and **larger institutions being better able to pivot to the affordances of GenAI to REF preparations**.

104. It was acknowledged by our interviewees that university professional services staff – and particularly those whose job function is in supporting institutional REF processes need to **quickly skill in AI tools** for their own job retention:

> *I think, certainly from a professional services point of view, the best way to keep your job and not get replaced by AI is to be the person that can manage the AI. And so, there's a decent amount of not necessarily reskilling, but just skilling involved at this point. Know what it is that's going to be asked of you and know when you can use not necessarily shortcuts, but tools that make your life easier and then you can do other things and won't get stuck in the weeds of administrative tasks.*
>
> *(REF-Lead RG Institution)*

105. How to make good prompts was seen as a key condition of good practice and would become a major part of support for REF in submitting institutions:

> *How to create prompts would be an enormous advantage and could become an industry; it could become a professional services cottage industry by itself.*
>
> *(REF-Lead Non-RG Research Intensive Institution)*

106. One of our participants (the PVCR of a RG institution) also spoke of the need for "a massive cultural shift not just a technology shift . . . a culture shift in general societal acceptance of AI in order to justify that level of funding allocation. And I can't see that happening within the next decade". A **cultural shift** has already occurred in very many sectors outside of higher education, which are shown to display high levels of GenAI investment and adoption[6]. This leaves us asking whether the perceived arduousness for cultural change in respect of GenAI and the REF reflects a general risk aversion of the UK HE sector (or excessive accommodation of tech-related sensitivities among its community) in response to GenAI or the necessary judiciousness of a sector carefully thinking through the implications of AI infusion (though insighted of the extent to which infusion is already occurring).

Concerns in the adoption of GenAI for REF

While our interviewees discussed their views on the contributions of GenAI (and conditions for its appropriate use for the REF), they also articulated a variety of concerns, here discussed.

107. It was argued by interviewees that the manual review process **provides valuable professional development benefits that could be lost with AI automation** and undergirds many assertions of the intrinsic value of the REF:

> *There is this sort of learning process and self-improvement and reflection that goes on as part of preparing those submissions . . . I'm not quite sure you would learn in the same way and, call me old fashioned, but going through those steps of writing, reading the peer review, the drafting, just writing the thing down, I think is quite significant.*
>
> *(PVCR RG Institution)*

> *I remember the old days when you submitted a load of work and 18 months later a pro vice-chancellor would send you a letter saying you're not in the REF. You know, that was the old days and we don't do that now. It's a much more engaged thing. We see it as being part of a professional academic development. I think if we move to, for want of a better term, a sort of more automated kind of approach, I'm not sure we'd ever get the developmental agenda. I'm not sure how we'd say to anybody, REF helps you.*
>
> *(PVCR Non-RG Research Intensive Institution)*

[6] https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work

However, the extent to which the REF provides developmental opportunity was also acknowledged to be uneven and related to variation of institutional resourcing – not least the availability of staff with related subject expertise and direct experience of REF processes, or more specifically learned experience of the behaviour of REF panels in the assessment of research outputs, impact and environment statements.

108. Interviewees raised significant concerns about **data security** and institutional adoption of AI tools, with **a preference for perceived closed systems** like Microsoft Copilot over open platforms like ChatGPT.

109. Without exception, interviewees discussed **the ethics of GenAI use**, particularly in respect of corporate control, privacy, and lack of regulation. Our conversations reflected deep reservations about trusting major technology companies with research data and AI systems.

> *I worry about, you know, theft effectively by corporations. I worry about privacy. It's corporations all involved in a kind of arms race to be at the forefront of AI. And you know, do I trust any of these corporations? Do we really trust Google? Do we really trust, you know, any of them? Apple, Amazon? We all act like we trust them.*
>
> *(REF-lead RG institution)*

110. The majority of our interviewees stated that without a **standardised tool across the sector**, the inevitable use of GenAI in REF preparations will bake in structural inequalities for poorer resourced institutions, less able to pivot to REF rules and requirements:

> *I think any kind of changes to the system, of any system always, benefit the most privileged first because they can put the resources in to adapt to it. And I think that's true of AI as well. Again, it's going to be your AI savvy, well-resourced universities that prioritize it that will adapt first . . . We all find a way to prioritize embedding AI in our submission. But the really small universities are never going to be able to have it. They don't have an equivalent of (in house AI tool) or anything like that.*
>
> *(REF-Lead RG Institution)*

111. An alternative view was that the associated costs of subscribing to such a sector-level tool would prohibit less affluent institutions from benefitting:

I would expect a company will become the preferred partner with enough time in advance that if people wanted to pay for it [a standardised GenAI tool] for a year, they could get that kind of tool usage in advance of the submission. So again, that comes down to budgets and financial capability of the institution. So even if there was the offering of a tool, it wouldn't necessarily provide a level playing field. Someone has to pay for the tool. It's unlikely that Research England are going to provide it pro bono to everybody because they've never done it before.

(REF-Lead RG Institution)

112. Interviewees also identified the **expense of AI literacy development and those using GenAI tools requiring training to ensure appropriate application**; costs that HEIs of different sizes and resource would not be able to equally meet. Relatedly, interviewees stated their concern of **implementing GenAI tools in a piecemeal fashion without proper coordination and guidance.** They stressed an urgent need for clear, sector-wide processes with appropriate support structures to ensure effective and responsible use.

113. Other equity concerns were raised by our interviewees in apropos of how any reduction in administrative burden through GenAI deployment **would result in new human resource needs** – that not all HEIs would be able to accommodate – in the form for instance of human oversight and managing of GenAI systems.

114. Sharing an LLM was also considered by some of our interviewees to potentially **risk their institutional competitiveness** (though a centralised system with institutional specific and gatekeeped entry points might provide an option), while a **shared data system for narrative generation was felt would be a more acceptable option**:

If you're putting things into a large language model, that's publicly available, how do we know that we're not, you know, giving people an advantage. It's so competitive, isn't it? The REF is a competitive exercise and so sharing a large, large language model, I think I'd feel that was a risk. What I think would be helpful though is sharing a system that would allow you to take multiple sources of data or information and convert it into a meaningful narrative.

(REF-Lead Post-92 institution)

115. Interviewees also raised concerns about who would 'own' and administer a centralised tool and the implications of its non-use where it failed to meet any given institution's AI standards.

116. Interviewees voiced concern that where use of GenAI tools by panellists was taken to be au fait accompli, **inconsistent use would be problematic**:

> *Some panellists are going to use AI anyway. So, I think REF needs to get its position in order and be really super clear one way or the other. And if their view is you don't use it, then that just needs to be really clear and told to every panel member from day one, because otherwise there's going to be a mixed economy and obviously certain outputs you can't review using AI. So, you'll be in the situation that some have been reviewed using AI and some haven't, which I think is potentially quite problematic.*
>
> *(REF-Lead Russell Group Institution)*

117. **Environmental sustainability** emerged as a major concern related to the escalated use of GenAI tools for REF purposes, highlighting the conflict between universities' sustainability goals and AI usage. One interviewee noted that HEIs lack policies around large language models despite their obvious environmental impact as a major oversight (and in turn intimated the organisational immaturity in thinking through GenAI integration):

> *Some of us are kind of trying to be in part of the sustainability rankings and all that sort of stuff, but yet none of us, as far as I know, have any policies around the use of large language models or any kind of AI. So, I think it's a real deep conflict of interest and I don't think it's been given any thought whatsoever.*
>
> *(REF-lead RG Institution)*

118. Interviewees also raised concerns about:

- Maintaining transparency and preventing political influence in AI-driven assessment systems.
- Maintaining academic integrity while leveraging technology.
- Maintaining research quality assessment rigor and avoiding over-reliance on AI tools (especially where over-use might compromise the credibility of institutional ownership of REF submissions).
- The assessment of non-traditional outputs using GenAI and the urgent need for trialling in such respect.
- How a turn towards a fully automated REF would remove academics' control over the REF process, and reduce the role of the REF to QR distribution only (though most of our interviewees recognised this as already being the main function of the REF).

- A lack of disclosure concerning the use of GenAI tools in REF preparations, where it was not a reporting requirement in the REF's Code of Practice.
- Whether the adoption of GenAI in the REF would overlook/miss aspects of research excellence.

Prognoses

119. Some of our interviewees foresee the future of the REF, beyond the currently slated 2029 iteration, being **fully automated**:

> *I certainly think for a successor to REF 2029, I would envisage a very high degree of AI used in the assessment which could very significantly relieve burden on universities. I mean, I can imagine a REF which is run essentially by Research England using an AI to gather information. And I think the sciences would probably have to be a pilot, but I think it could be done in a way that would gain the trust of the university sector and that is my anticipation of what post 2029 will be . . . We just have to understand that that's what's going to happen. It's probably a bit like the use of citation data where in previous REFs it was not trusted and it was looked down upon or even discouraged, actively discouraged. Now you know citation data is provided to the panels and they're told to use it responsibly. I suspect we'll end up with something along those lines for panel instructions on AI.*
>
> *(PVCR RG Institution)*

120. It was also suggested by our interviewees that an AI infused or automated REF would hold the potential of **providing 'current' insights into research quality and future research directions**:

> *I would expect the entire nature of the exercise to be completely changed if it exists in its current format in future. Because it could well be that if we go to this kind of model then the whole idea of the 7, 8-year cycle would probably disappear. We'd be on probably a continual annual roll. I would think that has been mooted in the past that we replace the entire exercise with something where there's just a light touch, annual review. It depends what the main focus of REF is, and I think this is probably the nub of the issue is what is REF actually for? Is it to judge, is it the litmus paper on the quality of UK research? Is it just to divvy up QR? Is it to get a view on the extent to which UK money is being spent, what money is being spent by the institutions? What is the main focus for this? And I think that will be the key driver because at this point in time it's just seen as a burdensome exercise which costs a huge amount of money on a seven-year cycle against which then a vast amount of UK funding is spent. Is that the only reason we're doing it or do we really want to know how good UK research is?*
>
> **[Continued on the following page]**

In which case then I think possibly that might flavour how you might approach this.

(REF-Lead RG Institution)

REF is a long, drawn out process, isn't it? It doesn't really necessarily tell you where we are now and what it doesn't ever do, although we have the opportunity to do it in the narrative, but only a little bit is, it doesn't tell you what's next. It's all retrospective. So, there's almost something, isn't there, about working out what there is and then also working out what's coming. That's where you could imagine AI in the future getting even more sophisticated about being able to plot the future research landscape based on what's there and what's up and coming.

(PVCR Post-92 Institution)

121. For some of our PVCRs, the use of GenAI by REF2029 panellists was viewed as a **near certainty** and was rationalised in relation to how further advanced they anticipated AI capability will be by 2029.

122. Interviewees also considered that for UoAs receiving **an abundance of books as research outputs, the temptation to use GenAI will be too great to discount,** given the time intensity associated with their assessment.

123. The potential of GenAI **reviewing all of an institution's outputs**, and not just those selected and submitted for the REF was also discussed by interviewees but concomitantly critiqued for the way in which a system of 'total assessment' would have a huge (managerial) effect on research practices.

124. A link was also made to the greater significance of **QS rankings** in terms of institutional funding and an automated **REF paving the way for a greater focusing on managing research citations**:

If they ditched the REF, my job as head of research performance would probably suddenly be to care about research citations in a way that I've never cared about before . . . refocusing on a game that's already been played, but perhaps in a much more generous way.

(REF-Lead RG Institution)

125. Where so much value is attributed to how the current system of REF (positively) affects institutional behaviour, one of our participants at a large RG institution, queried whether a proven REF-AI would 'influence or differently influence the behaviours of HEIs'?

Focus group findings

Discipline and demographic differences

1. Some participants observed that attitudes and usage patterns vary by academic discipline and demographic. It was suggested that tech-oriented fields (computer science, engineering, some sciences) are more inclined to embrace GenAI tools, often with greater understanding of how they work, whereas others (for example, traditionally less tech-focused fields like law or certain humanities) might lag or even resist. As one person half-joked, "***scientists are likely to embrace it more.. lawyers tend to be technophobes and old fuddy-duddies who don't want to engage***" (this was followed by a tongue-in-cheek comment that indeed the legal profession's regulator had just cautioned lawyers against using these tools, even as elsewhere a law firm was proudly touting an GenAI-driven service). Beyond discipline, digital literacy appears to cut across age and seniority in unpredictable ways: "***I'm shocked at how much some people know.. and equally shocked by people who say, I don't even know what that is***". One participant remarked, noting that familiarity with GenAI doesn't map neatly to one's field or career stage. This observation underscores the importance of not stereotyping, and that targeted training may be needed in pockets where awareness is low. Generally, attitudes from the focus groups toward GenAI in research are heterogeneous and in flux. There is a cautious optimism that these tools can play a useful, perhaps transformative, role, but tempered by real concerns about maintaining integrity, quality, and control.

Attitudes Toward Generative AI in Research

Mixed Reactions: Curiosity, Optimism and Caution

2. Across all institutions interviewed, participants reported a wide spectrum of attitudes toward using GenAI in research. Many described their communities as divided between early adopters excited by the potential and sceptics worried about its risks. "***Like most of the academic community, people are either for or against it – people are either 'it's a tool.. to speed up summaries.. to do analysis faster than us,' or 'no, we don't want to use it.' It's early days for us.***".

3. In most focus groups, there was a general level of curiosity of what could be done with the tools, coupled with uncertainty. One participant noted that when the first generation of ChatGPT appeared it felt like a novelty or "***parlour trick***," but newer iterations have advanced dramatically in knowledge and nuance, arriving "***very, very quickly and we're not entirely sure what to do with it, how to get the best out of it***". Many admitted to only limited direct experience with the tools. For example, one academic participant confessed, "***I know very, very little.. I've used ChatGPT about 10 times***".

They expressed finding it "***amazing for drafting emails***" but not yet integral to their workflow.

Overall, no institution reported a singular stance; attitudes were varied, and most often dependent on individuals' disciplines, technical literacy, or personal philosophy.

Early Adopters and Everyday Uses

4. Although there were various levels of resistance to the tools, numerous participants shared how they or colleagues have begun experimenting with AI to streamline research-related tasks. Several early adopters use generative AI as a writing assistant for mundane or laborious chores. At several universities we find staff members who have fed completed research papers into ChatGPT to "***generate an abstract***", finding that "f***or a first stab at an abstract, it's pretty good***". Others have used GenAI to draft lay summaries for grant applications or to extract key objectives for proposal forms. A few academics praised the tools for polishing language or editing: "***I've used it a couple of times when I've written something that's just unreadable, to get it to improve the English.. I've actually been quite impressed***". One researcher described using ChatGPT to précis a 100-page document into a concise summary, which "***speeded up my life***". These examples illustrate a pragmatic strand of optimism, the sense that the tools can (or might at some point) "***help speed up some groundwork***" and alleviate routine burdens. Participants noted that these efficiencies, in theory, free up time for more "***substantive intellectual work***".

Scepticism, "Beigeification" and Critical Caution

5. As well as enthusiasm for the tools' convenience, there were strong notes of caution. A recurring concern from academics is that heavy reliance on the tools for writing could dull the originality and quality of scholarly work. One arts academic, for instance, warned that using AI to craft narratives can result in "***what I call beigeification – you end up with [writing where].. you spot a lot of mistakes***", essentially bland and homogenised prose that lacks the critical edge or creativity of human-authored text. Others agreed that GenAI tends to produce generic outputs which, if used uncritically, might lower the standard of academic writing.

6. This ties into a broader concern about critical engagement: librarians in one focus group likened today's GenAI to Wikipedia in its early days, a useful resource if approached with scepticism. "***It's a critical literacy tool.. We can't stop it.. so yeah, use it, but this is how you engage with it - critically***".

Several institutions reported efforts to build such critical capacity (especially among students), emphasising that users must "***still bring that academic perspective.. still [be] critical about what they see and what they get out of it***". While the key message is that GenAI is here to stay and should be neither ignored nor blindly trusted, academics are grappling with how to maintain scholarly standards. As one participant commented, "***we cannot stop the flood.. [but] engage with it critically***".

Ethics and Security

7. Common across the focus groups was anxiety over the ethical pitfalls of GenAI in research. One major issue is data confidentiality and intellectual property (IP). Researchers worry that inputting unpublished research or sensitive data into an external tool (such as ChatGPT) could amount to inadvertently sharing it with the world. At one research intensive university, it was observed that some academics had naively pasted confidential information into ChatGPT to get a "***nice report***", not realising they might be exposing data externally. Comments from one participant shared worries about how the prompts they use are also "***feeding the beast***" questioning "***if we use it before our work is made public, anyone can pretty much know what we're working on, that's the problem***", the participant concluded. Many agreed this is a "***grey area***" that institutions have yet to clarify.

8. IP ownership is likewise murky when GenAI is involved: who is the author of AI-generated text or images, and can such outputs be plagiarised? An arts academic noted the irony that "***if you're using it correctly, you're not sharing intellectual property***", yet acknowledged "***there's still a lot of information and intellectual property.. going into something we don't own***" when we use commercial AI platforms. This external control of tools "***none of these systems are owned by us***" was flagged as a strategic concern for universities. Another ethical dimension cited was bias and accuracy, where participants recognise that GenAI can hallucinate false content (e.g. nonsensical or fake references in academic style) and often reflects the biases of its training data. One focus group mentioned a study finding that GenAI tools summarising research literature were "***generally not credible.. and even when they are, they do so inconsistently.. a serious threat to credibility***". This reinforces a widespread sentiment that any use of AI must be accompanied by human verification and sound judgment.

Attitudes Toward Generative AI in Research

AI in REF Preparation

9. A central topic to our study was whether and how generative AI might assist universities in preparing for the next REF. Participants noted a range of potential

uses: drafting standardized text (for example, the brief 100-word research output summaries required in REF submissions), proofreading for style consistency, generating first drafts of narrative sections (like impact case studies or unit "environment" statements), and analysing large volumes of outputs or data for REF strategy. Indeed, some academics who had struggled with writing concise summaries in REF2021 welcomed the idea of an algorithmic helper, "***we'd really like to be able to use it for that.. take academic writing and turn it into something homogeneous that would work***", one said, adding "***I think [AI] would be perfect***" for compressing and simplifying technical descriptions. Similarly, focus group members at one arts university suggested that a custom-trained AI could aid in describing artistic research outputs or proofreading portfolios to meet REF formatting guidelines, tasks that currently consume significant staff time. These examples indicate that efficiency gains in the REF process are widely anticipated. If an AI tool can produce a decent first draft or checklist, research offices can then refine the text rather than start from scratch. One impact officer noted that at a granular level, using an AI early in the drafting process might "***get [researchers] over that first hurdle***" of articulating their work in REF-friendly language, thus speeding up an often "***painful***" iterative writing process.

Quality and Oversight

10. Participants were quick to balance hopes of efficiency with caution about quality and oversight. There was consensus that AI-generated REF material would require careful human curation. As one person put it, these tools "***don't take away what you've got***" (i.e. scholarly substance), they might accelerate preparation, but the intellectual input and polishing still rest with researchers. A recurring term was "***AI is an assistive technology, not a substitute***". One participant stressed that even if AI helps produce a draft, "***you've still got to be you.. the human element, the judgment, the empathy, still has to lead***" in writing and reviewing REF narratives. The general approach floated was to use AI for preliminary work (summaries, structure, basic language) and then rely on human expertise to ensure accuracy, coherence, and nuance in the final submission.

Competitive Edge

11. A lively debate emerged on whether using AI confers any competitive advantage in REF outcomes. Some participants were sceptical that it would meaningfully boost a university's REF performance. "***I don't think our REF return is going to stand or fall on how much we engage with AI over the next few years***", argued one research lead. For REF2029 many felt it was "***too late in the process***" for AI to make more than marginal differences. Several reasons underpinned this view. Firstly, leveraging AI might require widespread staff training and culture change that cannot be accomplished in a short time frame. One REF

manager estimated that training all their Unit of Assessment leads to use AI effectively now "***would offset any gains in efficiency.. a bad use of time***" given how busy those people already are. Secondly, an arms-race dynamic was noted: once AI tools are widely available, any edge gained is likely to be short-lived as everyone will use similar methods. "***Could it not be argued that by speeding up the work we give better quality.. [but] everyone's going to do that.. don't put more hay on the wagon***", argued one participant, implying that across the sector AI's role may just raise the baseline rather than distinguish one institution. Others agreed: by the next cycle, these tools (or their successors) might be as common as word processors, making them an expected part of the process rather than a secret weapon.

12. A few focus group participants, however, suggested that not embracing AI could itself become a disadvantage. When at one university we as asked if there were risks to not using the tools, a response was: "***if we can do more work with less, then that's good***", i.e. any institution that refuses to improve efficiency will lag behind those that do. This fear of missing out was articulated at a more teaching focused university, where a senior manager described "***a real kind of big FOMO jetpack on a lot of institutions.. in the rush and the fear of missing out, you're more likely to [make] errors of judgment.. so how do you pace it?***". The balance, in their view, was to explore AI's advantages without succumbing to irrational hype or cutting corners. In practical terms, many saw AI as providing "***marginal gains***" in REF preparations; helpful efficiencies in drafting and analysis; but not a substitute for research quality or a revolutionary shortcut to higher scores. As one REF director put it, "***I think that really oversells what [the] potential is.. there might be some things, but.. I'm with [others], it's the other way***", cautioning not to overestimate AI's impact on the outcome.

Transparency, Credibility and REF Policy Guidance

13. A prominent theme to emerge from the focus groups was how Research England and the REF panels might respond to generative AI. Would they expect disclosure if AI was used in preparing submissions or even in generating research content? Many participants advocated for transparency and "***disclosure***" as key principles of good practice. The notion is that if a narrative, impact case, or even research output was produced with AI assistance, this fact should be made clear to maintain integrity. "***Transparency: have you used it? how did you use it?***" was the kind of guideline several groups felt should be in any sector-wide policy. Some noted that journals have already begun to establish policies (for example, requiring authors to acknowledge AI assistance in writing, and forbidding listing an AI as co-author). In the REF context, however, standards are yet undefined. A Russell Group participant involved in drafting their university's policy explained that it was "***commissioned [partly] to ensure that.. we capture the opportunities***

[AI] provides", but equally to "***minimise the risks***", for instance, by guiding staff on responsible use so that REF outputs remain credible. This reflects a delicate balance that institutions are struggling with; encouraging researchers to embrace AI where it helps, but safely. The institutional message, they said, would be "***please do engage and use it safely, and don't avoid using it for fear.. [instead] increase the literacy, grab the opportunities***". This institutional policy was in final approval stages at the time of the focus group, "***approved with minor amendments.. not quite published yet***", but aiming to tell researchers how to use AI in compliance with both ethics and likely REF expectations.

14. Participants clearly stated a desire for guidance from Research England and REF panels, with some posing fundamental questions: "***If AI co-produces research text or analysis, what exactly is REF assessing, human scholarly ability, or the quality of the research output irrespective of authorship?***" One arts participant suggested that by 2036 we might even see "***two REF[s]: one for human research and one for AI-supported research***", asking whether REF is meant to measure "***human capacity to produce research, or.. the knowledge base that the planet has, to which humans and AI both contribute***". While most found the idea of separate AI vs non-AI tracks fanciful (and adding even more burden to an "***already onerous process***"), the underlying point resonated, REF criteria may need evolution to address scenarios where AI is deeply embedded in research production. In the near term, simpler steps like requiring a statement of AI use in submissions could be considered; participants repeatedly came back to credibility. "***Credibility***," one said when asked what they would want Research England to ensure. There is a fear that undisclosed AI-generated content could undermine confidence in REF outcomes. The Dean of a Russell Group institution pointed out that unless AI's current limitations are overcome, any heavy reliance without disclosure risks flooding REF with low-credibility material: "***We find [current AI tools] generally are not [credible]. And even when they are, they do so inconsistently.. a serious threat to credibility***". The recommendation emerging from the focus groups is that sector-wide REF guidelines must be developed before the next cycle, promoting transparency, setting boundaries (e.g. AI may assist with style or summaries but not generate novel research content without acknowledgement), and training REF assessors to recognise appropriately disclosed AI use.

Fairness and the Future – Levelling the Playing Field?

15. A final consideration was whether GenAI will level inequalities between institutions or exacerbate them. On one hand, some argued that widely available AI tools could democratise certain aspects of REF prep. For smaller or teaching-focused universities that lack large teams of REF support staff, a good AI could act as a "***force-multiplier: a standardized tool that everyone signed up to and***

used" might reduce the gap in polishing capacity. We directly asked: "***will increased use of AI lead to more of a level playing field, or.. further competitive inequality?***". The responses were split. Some felt it could level the field if, for example, UK universities collectively had access to a high-quality large language model trained on academic content, a resource any institution could use to enhance writing and analysis. In theory, this might reduce reliance on having the most experienced human staff for narrative crafting, thus benefiting those with fewer resources. However, others were more pessimistic. They noted that developing and deploying AI itself requires investment, and richer institutions are already pulling ahead. It was revealed that some well-resourced universities are "***already in the process of developing their own in-house LLMs exactly for this purpose***". This suggests a new kind of stratification: those who can build or heavily customise AI tools versus those who must rely on generic public ones. "***Is it going to level the playing field or make it even more unequal?***" a participant from a smaller university cautioned that "***we've seen hype cycles before***", referencing how early adopters of past tech (like virtual worlds in the Second Life era) spent money but ultimately "***got their fingers burned***". The danger is assuming the massive current investment in AI will inevitably pay off for academia. If the technology's value plateaus or its proprietors (big tech companies) charge high fees, those betting on AI might not reap proportional rewards.

16. In summary, whilst GenAI is widely seen as a tool that will be used in REF preparations, its precise role remains undefined and somewhat contentious. The focus groups urge clarity on ethical use and call for a measured approach: use AI to assist and augment human effort, but do not compromise the credibility of REF by abandoning human judgment. Most do not see AI radically changing REF outcomes in the immediate cycle, but they are keenly aware that by the 2030s the landscape could evolve drastically, necessitating ongoing dialogue between HEIs, the UK funding bodies and REF Steering Group.

Institutional Readiness and Responses to AI Adoption

Policy Development

17. One striking finding is that, as of the focus group conversations, few universities had formal policies in place for GenAI in research, though many were in the process of developing them. In several institutions, participants noted that initial official guidance focused on student use of AI (for example, rules around AI and plagiarism in coursework), while policies for staff research use lagged behind. "***At the institutional level, a lot of the guidance has been focused on students.. how they can or can't use it in assessments***" said one academic, "***There is a policy being developed on generative AI in research. We don't have it yet***".

This was echoed across multiple sites. One research-intensive university focus group described their institutional stance as essentially "***blind***" and that leadership had "***no idea what staff were doing with AI, and no clear policies or principles had been issued***". Another institutional focus group agreed "***There's no university policy***" on AI in research at present, characterising their approach as "***emergent and decentralised***": "***administratively, we are a lot more comfortable being fast followers than leaders.. someone puts their neck out and then you follow.. when it works***". In other words, they were letting usage grow organically and intended to codify best practices once they become clearer.

18. By contrast, a couple of research-intensive universities were further along. One stood out for having a nearly finished policy at the time of discussion, a focus group participant reported that a formal "***policy for the use of GenAI in research***" had been "***commissioned***" by the institution's Research and Innovation Committee. This policy had passed committee approval with minor amendments and was "***ready to be published***" aiming for rollout and publicity at the start of the new academic year (September). Oxford's forthcoming policy intends to encourage researchers to "embrace [AI] safely" and will include guidance to "***ensure.. on one hand we're minimising the risks.. but also.. capture the opportunities***". It will likely cover issues like data security, ethical use, and disclosure, thus providing a framework for staff. An arts university indicated it is not creating a standalone new policy from scratch but instead augmenting existing ethics guidelines: "***We're not going to make a new policy, but work within our current guidelines***" said one senior manager, referring to an internal "AI and ethics" working group shaping recommendations. This suggests a trend of embedding AI considerations into pre-existing research ethics or integrity frameworks.

19. Institutional policy readiness varies widely: a few universities are on the cusp of issuing comprehensive policies for AI use in research, several have ad-hoc or in-progress efforts, and some largely rely on external policies (journal guidelines, funder rules) or local common sense in absence of top-down direction. The participants generally expressed a desire for clearer institutional guidance, not to police researchers punitively but to give them confidence about what is allowed and advisable. As one person noted, "***without guidance, researchers may either refrain from beneficial uses out of fear or charge ahead in risky ways out of ignorance – neither scenario is ideal***". A well-communicated policy can "***ensure that.. people.. don't avoid using [AI] for fear of doing things you shouldn't and [also] use it safely***".

Training and Capacity-Building

20. Alongside formal policy, training and support emerged as critical components of institutional readiness. Many participants observed that researchers need upskilling to use GenAI effectively and ethically. At one research intensive, for example, the academic development team has "***started putting in training sessions around the use of generative AI in research***", primarily to highlight innovative uses by those academics already experimenting with the technology. However, this was described as early and voluntary: "***showcasing [keen] academics.. but not [yet] anything institutionally.. that we're promoting***". In other words, they were beginning to cultivate communities of practice, though without mandating broad training.

21. Elsewhere, participants complained of gaps in awareness that training could fill. One participant in a teaching focused institution noted "***large chunks of the institution need training, or if we do [have training resources], nobody knows it's there***". They felt that achieving any competitive edge or even full uptake of GenAI's benefits would require much more investment in staff development. In their view, at the current stage "***we're too far behind***" to leverage GenAI widely for REF without first educating many colleagues. This aligns with earlier observations that digital literacy with GenAI varies greatly; some academics don't even know what tools exist, while others might be overconfident and unaware of pitfalls like data leakage. Structured workshops, guidance documents, and perhaps integration of AI literacy into existing professional development were all implied needs.

22. Some HEIs are leveraging their library and IT staff to lead the way. A research-intensive institution's librarians have treated AI as an extension of information literacy training: "***It's an excellent tool for teaching critical literacy.. there are pockets of training where librarians are.. showing students how to engage with the tools in a critical way***". Though they didn't have a central directive, these grassroots training efforts acknowledge that "***they're going to use it – we cannot stop the flood.. what we can do is [teach], this is how you engage with it critically***". There was also mention of practical guidance such as testing assignments against AI (e.g. running an essay question through ChatGPT to see if the answer would pass, then adjusting the question) as a new standard practice in some departments. This shows staff proactively adapting teaching and assessment in response to AI – a form of training by doing. For research specifically, training might include how to prompt AI effectively, how to fact-check AI outputs, understanding AI's limitations, and data security protocols (e.g. using approved tools or anonymising inputs). Notably, at one university, a participant called for training that addresses "***what IP is in the first place.. what you can't disclose [to AI].. there's a really big gap around that***" which currently either

"***puts people off using [AI] or they're using it naively***". Filling such knowledge gaps is clearly on the institutional to-do list.

Technical Infrastructure

23. While most institutions are currently relying on public tools (e.g. OpenAI's ChatGPT, Microsoft's Bing/CoPilot, Anthropic's Claude), there was reference to some universities building in-house large language models (LLMs) for REF and research support. Although details were scant, this suggests that richer institutions (perhaps those with strong computer science departments or resources) are investing in AI systems trained on their own data or tuned to academic use-cases. The potential advantage is control: an in-house AI could be used on confidential research material without it leaving the university, addressing the data privacy issue. It could also be customised to UK research context, REF terminology, etc. The downside is the cost and expertise required, which only some universities can marshal. If this trend continues, it might widen the gap between AI-haves and have-nots in the sector. The levelling solution proposed by some was a sector-wide approach where a sector body could potentially host or negotiate a common AI service for universities.

Current Stopgaps and Local Practices

24. In absence of comprehensive institutional frameworks, many academics have defaulted to using their own judgment or following external guidelines. For instance, some participants mentioned relying on publisher or professional association guidance. An example given was the British Sociological Association's ethical guidelines, one participant noted they look to such bodies, but admitted there is "***no wider policy.. within our subject groups***" yet on AI use. In terms of local practice, we heard instances of departments devising their own interim rules. One academic described how different programmes within the same school took varied approaches to student use of AI, resulting in inconsistency in handling AI-related plagiarism or misconduct. This ad-hoc approach is clearly not ideal and participants called for more coordination: "***there's an area there where we really need to.. work harder***" to harmonise policies, they said.

25. Some universities have issued high-level statements or principles as a stopgap. One mentioned that one of their Pro-Vice-Chancellors had outlined "***six ways to adopt GenAI to support the activities of [the university]***" as a strategic direction. While details of those six points were not given in the focus group, the mention indicates that at least at the leadership vision level, institutions are acknowledging AI. The actual implementation of such vision, however, was seen as slow: "***the university spent a long-time dancing around this without producing much.. things supposedly in the works, and yet we haven't really had***

anything come to us. We've had to develop [practices] by ourselves". This quote likely resonates across many campuses, the sense of waiting for central guidance that is late to arrive.

26. Institutional responses are in an early, uneven stage. A few universities are relatively advanced in crafting policy and guidance, many are still formulating their approach or extending student-focused rules to research contexts, and in the meantime, academics and professional staff are improvising and sharing knowledge on the ground. There is a clear call for more communication from leadership: as one manager reassured colleagues, if they hadn't yet seen the new policy, it was only because it was brand new "***if you've never heard of this policy.. it's [because] it isn't quite published yet.. It's in a very interesting and exciting state***". Going forward, universities will need to ensure that once policies or guidelines are approved, they are disseminated and translated into practice through training sessions, resource toolkits, and integration into research management processes (e.g. REF planning meetings explicitly covering AI dos and don'ts).

Shared Concerns and Divergent Outlooks Across Institutions

27. Across all the focus groups, despite differences in emphasis depending on the type of institution, several shared concerns consistently emerged. Academic staff everywhere worry about maintaining research integrity in the face of AI-generated content. Plagiarism and attribution are universal issues: whether students handing in AI-written essays with fake references (where one participant exclaimed that a student's bibliography "***contained a bunch of references that simply don't exist***", presumed to be AI-invented), or researchers potentially failing to credit AI contributions in research articles. The consensus is that transparency and clear rules are needed. Data privacy and security is another common thread: every institution expressed concern about sensitive research data being exposed via AI tools. This is driving calls for either safer tools (onshore servers, in-house models) or user education to avoid inputting confidential information. Reliability and accuracy of AI outputs troubled academics across the board, no one is ready to trust GenAI unchecked. Many gave examples of AI confidently outputting erroneous information (the "hallucination" problem), reinforcing the notion that human expertise must validate all results. Bias and ethics also featured: participants recognise that AI can reproduce societal biases present in training data, and that widespread use of AI might have ramifications for diversity and originality in research.

28. Finally, all institutions share the recognition that GenAI is an external disruptor to which they must respond, but none claim to have fully figured it out yet. Humility and uncertainty pervade the discussions. As one participant aptly put it,

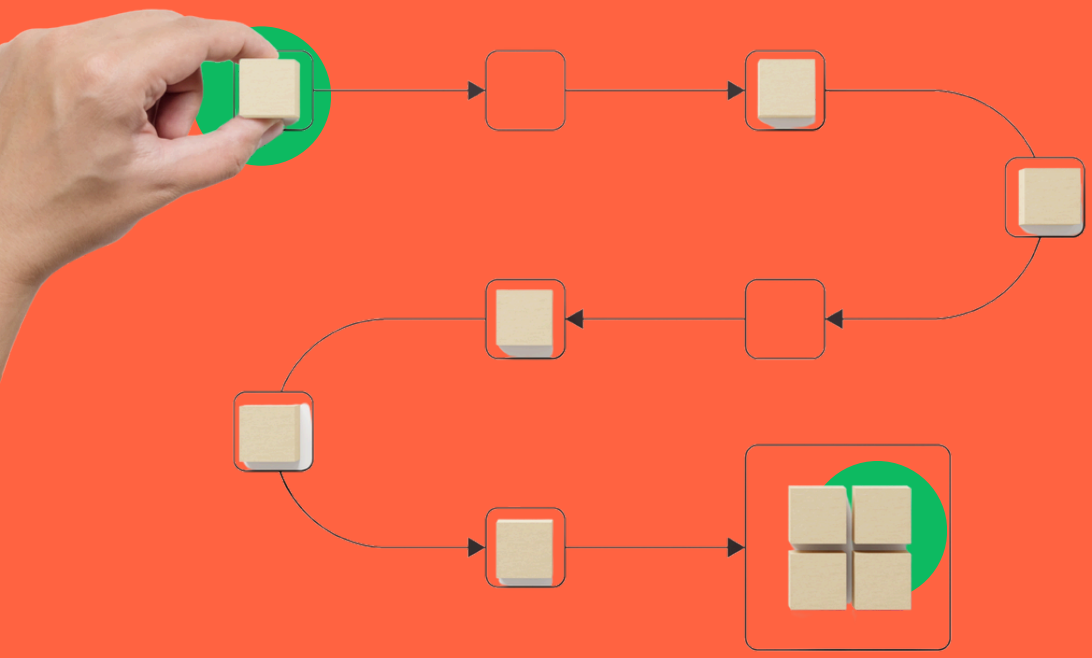
"***we just don't know yet***", much of this area is "***a grey area***", and experimentation, evidence gathering, and ongoing dialogue will be needed. The focus groups themselves are part of that dialogue, and participants frequently expressed appreciation that this conversation was happening at a sector level, allowing them to compare notes.

Summary observations

29. The focus groups show that the UK higher education sector is in a moment of negation with the realities of GenAI, trying to confront the disruption and realise the promise. Participants' responses are on a wide spectrum in favour and against and recognise that many GenAI tools and practices are already woven into systems and behaviours of both staff and students, and increasingly into research practice. However, their use remains uneven and often improvised. Early adopters have realised some benefits in respect of efficiency, such as drafting summaries, refining text, and supporting REF preparation. Concurrently participants vocalise legitimate concerns about intellectual erosion, data security, and loss of scholarly distinctiveness. There is a tension between innovation and integrity that permeated discussions. Staff are aware that ignoring AI is not feasible but adopting it without critical oversight poses real risks to research credibility. The dominant tone throughout the focus groups was one of cautious pragmatism: seeking opportunities for AI to enhance productivity and ease administrative burden, on the proviso that institutions and individuals retain responsibility for judgement, accuracy, and ethical conduct.

30. The focus group findings also highlight an urgent need for strong governance and shared standards. It is clear without institutional policies and national guidance, universities risk fragmented approaches, uneven capacity, and growing inequity between those able to develop bespoke systems and those reliant on public tools. Participants consistently called for transparent REF guidelines to define what constitutes responsible AI use and how such use should be disclosed. Equally, there is consensus that human oversight must remain central to all evaluative processes, ensuring that any automation supports rather than supplants academic reasoning. The sector's readiness for GenAI appears from their vantage to depend less substantively on the sophistication of the tools themselves and more on the culture, training, and ethical frameworks built around them. If GenAI has a place in the Academy, then it must be adopted critically, collaboratively, and ethically, shaping its integration in ways that uphold scholarly rigour, trust, and fairness across the research community.

Conclusion



1. The REF:AI project has been a significant undertaking given the compressed timeframe for and scale of data collection. In the space of only a few months the research team has managed to consult with many leading voices across UK HEIs and gained access to a variety of institutional, disciplinary and role-based perspectives and experiences that have significantly enriched what was previously an impoverished understanding of how GenAI tools are being and could be used for the REF.

2. We have ended up with strong symmetry across our three datasets, though we acknowledge that attitudes in opposition to GenAI use for the REF are most prevalent in our survey data. On this we reflect upon REF:AI not only as a learning journey for the research team but all those who participated in its conversations. Many of our focus group participants, for instance, began at the **outset of discussions with a fairly rigid outlook that the use of GenAI tools for the REF is not a particularly good idea.** These, however, were positions that in many cases **mellowed and became much more accommodating** as ideas and experiences were shared and discussion developed. In nearly every case, we found that focus group discussion culminated with participants **being no less critical yet far less closed off to the potential use of GenAI for the REF**. In many instances we also enjoyed the thanks of participants who congratulated the research team for providing (in many cases the first) opportunity for substantive discussion of the use of GenAI in their institutional REF preparations. We have also come to know, from those who have remained in touch with us since the focus groups, that REF:AI has been instrumental in either kick-starting or providing necessary momentum to institutional conversations and strategy building concerning GenAI tools.

3. **This momentum cannot be lost**. GenAI and its role in the REF **must not be kicked down the road or dismissed**. GenAI's influence over the REF **is no future imaginary but present reality**. Further work and, dare we say, **time** must now be afforded to allow for REF to appropriately respond to the recommendations we have provided. The credibility and continuance of it depends on substituting hysteria towards GenAI use with critical maturity, policy resourcing and shared infrastructure that allows for those who participate within it to maximise equally from the affordances of GenAI while being protected of its potential misapplications.

References

Bentley, S. V., Evans, D., & Naughtin, C. K. (2025). What social stratifications in bias blind spot can tell us about implicit social bias in both LLMs and humans. Scientific Reports, 15(1), 30429. https://doi.org/10.1038/s41598-025-14875-3

Bhattacharya, A. (2024, September 23). AI in Peer Review: The Positive, The Negative and Insights from the Research Integrity Team. Sage Blogs. https://www.sagepub.com/explore-our-content/blogs/posts/sage-perspectives/2024/09/23/ai-in-peer-review-the-positive-the-negative-and-insights-from-the-research-integrity-team

Carobene, A., Padoan, A., Cabitza, F., Banfi, G., & Plebani, M. (2024). Rising adoption of artificial intelligence in scientific publishing: evaluating the role, risks, and ethical implications in paper drafting and review process. Clinical Chemistry and Laboratory Medicine (CCLM), 62(5), 835–843. https://doi.org/10.1515/cclm-2023-1136

Doskaliuk, B., Zimba, O., Yessirkepov, M., Klishch, I., & Yatsyshyn, R. (2025). Artificial Intelligence in Peer Review: Enhancing Efficiency While Preserving Integrity. Journal of Korean Medical Science, 40(7). https://doi.org/10.3346/jkms.2025.40.e92

Farber, S. (2024). Enhancing peer review efficiency: A mixed-methods analysis of artificial intelligence-assisted reviewer selection across academic disciplines. Learned Publishing, 37(4). https://doi.org/10.1002/leap.1638

Giglio, A. Del, & Costa, M. U. P. da. (2023). The use of artificial intelligence to improve the scientific writing of non-native English speakers. Revista Da Associação Médica Brasileira, 69(9). https://doi.org/10.1590/1806-9282.20230560

Joachim, M. V., Dodson, T. B., & Laviv, A. (2025). How artificial intelligence differs from humans in peer review. Journal of Oral and Maxillofacial Surgery, 83(8), 1040–1050. https://doi.org/10.1016/j.joms.2025.03.015

Kousha, K., & Thelwall, M. (2024a). Artificial intelligence to support publishing and peer review: A summary and review. Learned Publishing, 37(1), 4–12. https://doi.org/10.1002/leap.1570

Kousha, K., & Thelwall, M. (2024b). Assessing the societal influence of academic research with ChatGPT: Impact case study evaluations. https://arxiv.org/abs/2410.19948

Latona, G. R., Ribeiro, M. H., Davidson, T. R., Veselovsky, V., & West, R. (2024). The AI review lottery: Widespread AI-Assisted peer reviews boost paper scores and acceptance rates. https://arxiv.org/abs/2405.02150

Li, M., Sun, J., & Tan, X. (2024). Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Systematic Reviews, 13(1), 219. https://doi.org/10.1186/s13643-024-02609-x

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. https://arxiv.org/abs/2403.07183

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., Yin, Y., McFarland, D. A., & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI, 1(8). https://doi.org/10.1056/Aloa2400196

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou, J. (2025). Quantifying large language model usage in scientific papers. Nature Human Behaviour. https://doi.org/10.1038/s41562-025-02273-8

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou, J. Y. (2024). Mapping the Increasing Use of LLMs in Scientific Papers. https://arxiv.org/abs/2404.01268

Mann, S. P., Aboy, M., Seah, J. J., Lin, Z., Luo, X., Rodger, D., Zohny, H., Minssen, T., Savulescu, J., & Earp, B. D. (2025). AI and the future of academic peer review. 1–34. https://arxiv.org/abs/2509.14189

Marrella, D., Jiang, S., Ipaktchi, K., & Liverneaux, P. (2025). Comparing AI-generated and human peer reviews: A study on 11 articles. Hand Surgery and Rehabilitation, 44(4), 102225. https://doi.org/10.1016/j.hansur.2025.102225

Mostafapour, M., Fortier, J. H., Pacheco, K., Murray, H., & Garber, G. (2024). Evaluating Literature Reviews Conducted by Humans Versus ChatGPT: Comparative Study. JMIR AI, 3, e56537. https://doi.org/10.2196/56537

Saad, A., Jenko, N., Ariyaratne, S., Birch, N., Iyengar, K. P., Davies, A. M., Vaishya, R., & Botchu, R. (2024). Exploring the potential of ChatGPT in the peer review process: An observational study. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 18(2), 102946. https://doi.org/10.1016/j.dsx.2024.102946

Scherbakov, D., Hubig, N., Jansari, V., Bakumenko, A., & Lenert, L. A. (2025). The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. Journal of the American Medical Informatics Association, 32(6), 1071–1086. https://doi.org/10.1093/jamia/ocaf063

Thelwall, M. (2024, October 30). Can ChatGPT run the REF? Not yet—but it might help. Research Professional News.https://www.researchprofessionalnews.com/rr-news-uk-views-of-the-uk-2024-october-can-chatgpt-run-the-ref-not-yet-but-it-might-help/

Thelwall, M., & Kurt, Z. (2025). Research evaluation with ChatGPT: is it age, country, length, or field biased? Scientometrics, 1–21. https://doi.org/10.1007/s11192-025-05393-0
Thelwall, M., & Yaghi, A. (2025). Evaluating the predictive capacity of ChatGPT for academic peer review outcomes across multiple platforms. Scientometrics. https://doi.org/10.1007/s11192-025-05287-1

Thelwall, M., & Yagui, A. (2025). In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results. Trends in Information Management, 13(1), 1–29.
Vincent-Lamarre, P., & Larivière, V. (2021). Textual analysis of artificial intelligence manuscripts reveals features associated with peer review outcome. Quantitative Science Studies, 2(2), 662–677. https://doi.org/10.1162/qss_a_00125

Wang, Z., Cao, L., Jin, Q., Chan, J., Wan, N., Afzali, B., Cho, H.-J., Choi, C.-I., Emamverdi, M., Gill, M. K., Kim, S.-H., Li, Y., Liu, Y., Luo, Y., Ong, H., Rousseau, J. F., Sheikh, I., Wei, J. J., Xu, Z., … Sun, J. (2025). A foundation model for human-AI collaboration in medical literature mining. Nature Communications, 16(1), 8361. https://doi.org/10.1038/s41467-025-62058-5

Watermeyer, R., Phipps, L., Lanclos, D. and Knight, C. (2024a) Generative AI and the automating of academia. Postdigital Science and Education. 6, 446–466. https://doi.org/10.1007/s42438-023-00440-6

Watermeyer, R., Lanclos, D., Phipps, L., Shapiro, H., Guizzo, D., and Knight, C. (2024b). Academics' Weak(ening) Resistance to generative AI: The cause and cost of prestige? Postdigital Science and Education. https://doi.org/10.1007/s42438-024-00524-x

Ye, R., Pang, X., Chai, J., Chen, J., Yin, Z., Xiang, Z., Dong, X., Shao, J., & Chen, S. (2024). Are We There Yet? Revealing the Risks of Utilizing Large Language Models in Scholarly Peer Review. https://arxiv.org/abs/2412.01708

Authors

Richard Watermeyer

is Professor of Higher Education and Co-Director of the Centre for Higher Education Transformations at the University of Bristol

Lawrie Phipps

is Senior Research Lead at Jisc and Visiting Professor of Digital Leadership at the University of Chester

Rodolfo Benites

is the REF-AI Research Fellow and Doctoral Researcher within the Centre for Higher Education Transformations at the University of Bristol

Tom Crick

is Professor of Digital Policy at Swansea University

Citation

Watermeyer, R., Phipps, L., Benites, R., and Crick, T. (2025). REF:AI: Exploring the potential of generative AI for REF2029. Centre for Higher Education Transformations and Jisc, Bristol.

About REF-AI

If you want to know more about the REF-AI project, scan the QR code below or click here.

CHET
CENTRE FOR HIGHER EDUCATION TRANSFORMATIONS

Helen Wodehouse Building
35 Berkeley Square
Bristol
BS8 1JA
chet-centre@bristol.ac.uk
https://chet.bristol.ac.uk
Centre for Higher Education Transformations - CHET

Jisc

4 Portwall Lane
Bristol
BS1 6NB
0300 300 2212
help@jisc.ac.uk
https://jisc.ac.uk
Jisc