
Article:
Kousha, K. and Thelwall, M. orcid.org/0000-0001-6065-205X (2025) Assessing the societal influence of academic research with ChatGPT: Impact case study evaluations. Journal of the Association for Information Science and Technology (JASIST). ISSN 2330-1635

https://doi.org/10.1002/asi.25021

Received: 25 October 2024 | Revised: 30 April 2025 | Accepted: 5 May 2025

DOI: 10.1002/asi.25021

RESEARCH ARTICLE

JASIST WILEY

Assessing the societal influence of academic research with ChatGPT: Impact case study evaluations

Kayvan Kousha[1] | **Mike Thelwall**[2]

[1]Wolverhampton Business School, University of Wolverhampton, Wolverhampton, UK

[2]Information School, University of Sheffield, Sheffield, UK

Correspondence
Mike Thelwall, Information School, University of Sheffield, Sheffield S10 2TN, UK.
Email: m.a.thelwall@sheffield.ac.uk

Funding information
Economic and Social Research Council, Grant/Award Number: APP43146

Abstract

Academics and departments are sometimes judged by how their research has benefited society. For example, the UK's Research Excellence Framework (REF) assesses Impact Case Studies (ICSs), which are five-page evidence-based claims of societal impacts. This article investigates whether ChatGPT can evaluate societal impact claims and therefore potentially support expert human assessors. For this, various parts of 6220 public ICSs from REF2021 were fed to ChatGPT 4o-mini along with the REF2021 evaluation guidelines, comparing ChatGPT's predictions with published departmental average ICS scores. The results suggest that the optimal strategy for high correlations with expert scores is to input the title and summary of an ICS but not the remaining text and to modify the original REF guidelines to encourage a stricter evaluation. The scores generated by this approach correlated positively with departmental average scores in all 34 Units of Assessment (UoAs), with values between 0.18 (Economics and Econometrics) and 0.56 (Psychology, Psychiatry and Neuroscience). At the departmental level, the corresponding correlations were higher, reaching 0.71 for Sport and Exercise Sciences, Leisure and Tourism. Thus, ChatGPT-based ICS evaluations are simple and viable to support or cross-check expert judgments, although their value varies substantially between fields.

1 | INTRODUCTION

Governments and research funders sometimes ask institutions to explain how their research benefits society. This can take many forms, from informal discussions between civil servants and academic leaders to structured periodic requests for descriptions of the societal benefits generated. This is part of the managerial turn of academia (Raaper & Olssen, 2015), with increased accountability for public spending. The process is perhaps most structured and systematic in the UK, where the 2021 Research Excellence Framework (REF) national assessment included 6781 Impact Case Studies (ICSs), which are five-page evidence-based claims of the societal impact achieved by submitting units (approximately departments). Each ICS is unique in terms of the nature of the impact claimed, the underpinning research, and the impact evidence presented. Nevertheless, nearly two-thirds (68%) of respondents in a REF2014 survey reported difficulties in providing impact evidence (Morgan Jones et al., 2017) as highlighted in other qualitative studies (e.g., Smith & Stewart, 2017; Wilkinson, 2019), so the evidence provided in ICSs might often be inconclusive. Assessing these claims is likely to be time-consuming and complex

(Derrick, 2018), with REF2014 evaluators from the social sciences and humanities (Watermeyer & Chubb, 2019) and biomedical sciences (Samuel & Derrick, 2015) struggling to interpret non-academic impact claims. Hence, any automated support could be useful for the evaluations as well as for the departmental process of writing and selecting the best ICSs.

Although there has been over half a century of research into the development and evaluation of citation-based indicators to help assess the scholarly impact of academic journal articles (de Bellis, 2009), there seems to have been only one previous attempt to support the evaluation of all ICSs or similar impact narratives with indicators or automated processing. The automated and semi-automated investigations so far have examined the nature of the impact claims and evidence used, mainly from a descriptive perspective. For example, the references and sources of non-academic impacts in ICSs have been analyzed (Digital Science, 2016; Kousha et al., 2021; Reddick et al., 2022), as have the "nature scale and beneficiaries of research impact" (King's College London and Digital Science, 2015).

The one attempt to assess whether information could be automatically extracted from ICSs to help expert judgments in evaluating individual ICSs used traditional machine learning (e.g., Random Forest) on REF 2014 ICSs. The inputs were an ad-hoc range of curated features covering, "discipline, institution, explicit text, implicit text, bibliometric indicators and policy indicators," including citation data for the references, author affiliation data and text properties, such as sentiment and readability. This study had an accuracy of up to 90% at distinguishing between ICSs from the top 20% with those from the bottom 20% in terms of submitting department ICS average score (Williams et al., 2023). This gives evidence that ICSs contain information that could be leveraged for a prediction but the conclusions of this study are limited by the lack of a development corpus given the curated set of features used, the inclusion of non-standard neural networks (which had the highest accuracy) and the ability to tweak input parameters for the machine learning models. Moreover, the experiment only processes extreme scoring ICSs and is therefore not realistic (this was not its intention) or comprehensive in the sense of attempting to score all ICSs.

In response to the lack of direct evidence that automated methods can support the evaluation of narrative impact claims by providing score estimates, this article investigates whether ChatGPT can do so. The UK ICSs are used as an example because they have indirect impact scores available and form a large corpus of careful narrative impact claims in a standard format. They are lengthy documents (5 pages) and previous research into detecting the quality of academic journal articles found that ChatGPT gave better scores when fed the title and abstract than when fed the full text (Thelwall, 2024, 2025a). Thus, it is logical to investigate whether feeding a summary or other parts of an ICS would be better than feeding it all to ChatGPT. In addition, it would be useful to know if it is possible to vary the ChatGPT system instructions to improve its performance, the default instructions being the same as those given to the human experts assessing ICSs in the UK. The goals are scientometric rather than computer science: not focusing on getting the best estimates (in any case LLMs are a rapidly evolving) but on understanding if the task is possible, the field variations in its value, and the possible reasons for positive results. More subtly, the focus is on ChatGPT scores as indicators rather than estimates. Previous research has found ChatGPT to be poor at estimating research quality scores but much better at ranking them. The ranks are useful as indicators to help human evaluators to score ICSs or make decisions about the relative merits of similar ICSs. Whilst ChatGPT estimates can be rescaled to give more accurate scores (Thelwall, 2025a), this is probably not the most useful way to employ them.

RQ1. For which text inputs does ChatGPT produce the most useful research quality estimates for ICSs?

RQ2. For which system prompts does ChatGPT produce the most useful research quality estimates for ICSs?

RQ3. Are there disciplinary differences in the ChatGPT quality estimates?

2 | REF IMPACT CASE STUDIES

Impact Case Studies (see Appendix C for a brief glossary) are structured narratives that provide evidence of the societal impacts of academic research beyond academia, as part of the UK national research evaluation exercises. Introduced in REF2014, ICSs provide evidence of the positive "effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment, or quality of life, beyond academia" (REF, 2021b, p. 68). This includes impacts on activities, attitudes, awareness, behaviors, performance, policies, or practices, influencing a variety of audiences, communities and organizations across any geographic location. In REF2021, ICSs accounted for 25% of the overall assessment and contained the following structured sections (REF, 2021a, pp. 96–98):

- **General information** about case study such as institution, Unit of Assessment (UoA), title, and names and roles of researchers involved in the study.
- **Summary of the impact** (100 words) describing the nature and extent of the impact.
- **Underpinning research** (500 words) explaining the key research findings related to the impact, the research produced (e.g., outputs or projects) and contextual information about the research area.
- **References to the research** (six references) citing the underpinning research
- **Details of the impact** (750 words) an evidence-backed narrative explaining how the research contributed to the non-academic impacts and the nature, extent, and beneficiaries of the impacts.
- **Sources to corroborate the impact** (10 sources) a list of sources that can support the impact claims, such as testimonials, policy documents, reports, news stories, or websites.

In REF2021, ICSs were evaluated for the reach and significance of their impacts on the economy, society, culture, public policy, health, the environment, or quality of life during the REF period (1 August 2013–31 December 2020). Reach refers to the extent and diversity of the beneficiaries of impacts, and significance measures how deeply the research has influenced performance, policies, practices, or services (REF, 2021b, p. 52).

The REF2021 impact assessment used primarily senior academic experts grouped into 34 UoAs (REF subjects), scoring from 1* to 4* based on the level of impact achieved (REF, 2021a, p. 85):

- 4*: Outstanding impacts in terms of their reach and significance.
- 3*: Very considerable impacts in terms of their reach and significance.
- 2*: Considerable impacts in terms of their reach and significance.
- 1*: Recognized but modest impacts in terms of their reach and significance.
- Unclassified: The impact is of little or no reach and significance; or the impact was not eligible; or the impact was not underpinned by excellent research produced by the submitted unit.

A single ICS could be assigned multiple scores for different aspects (e.g., reach, significance), but these scores are not published. The only public scores are the percentages of all ICSs achieving each of the star levels for each submission. Each "submission" is (almost always) from a single university and is made to a single UoA. For example, there was one "submission" from the University of Aberdeen to UoA 8: Chemistry. This included two ICSs and the public scores are 4*: 50%; 3*: 50%. This submission was presumably mainly from people employed by the Department of Chemistry at the University of Aberdeen, but some or all of the work may have derived from other University of Aberdeen departments.

3 | ASSESSING THE SOCIETAL IMPACTS OF ICSs

3.1 | Text mining to capture societal impacts

Text mining has been widely used to identify the societal benefits of REF ICSs (e.g., Adams et al., 2015; Bonaccorsi et al., 2021; Chowdhury et al., 2016; Zheng et al., 2021). Two large-scale studies applied text mining and topic modeling to REF2014 and REF2021 ICSs, reporting the diversity of impact pathways across subjects. The King's College London study (2015) identified 60 impact topics in REF2014, while a similar analysis of REF2021 found 79 topics (Stevenson et al., 2023). Text mining has also identified multiple broad societal impacts from REF2014 including "Education," "Environmental Energy Solutions" (Terämä et al., 2016), "People," and "Economy" (Parks et al., 2018). In REF2021, health and social work, education, and public administration were the most common impacts in Welsh ICSs (Pollitt et al., 2023).

Keyword searches of ICSs have also been used to report the prevalence of terms related to pre-defined impact topics such as educational technology (Jordan, 2020), social media (Jordan & Carrigan, 2018), gender and sexuality (Vanlee, 2024), leadership and management (Morrow et al., 2017), economic and social impacts (Koya & Chowdhury, 2020), and research data (Jensen et al., 2022). An analysis of the websites cited as evidence of non-academic impacts found that news stories government publications parliamentary records online videos and social media were all commonly used although with substantial disciplinary differences between UoAs (Kousha et al., 2021).

3.2 | Citations and altmetrics for impact assessment

Citations and alternative indicators can be used to assess the impacts of the research cited in ICSs. One study took advantage of the fact that the REF assesses both research outputs and ICSs, and that the scores can be independent. It extracted 921,254 submitted outputs (mostly journal articles) from the Outputs component of REF2014 (and its

predecessor RAE2008) and 36,244 ICS references from the Impact component of REF2014. It found that 42% of ICS references were also separately submitted for assessment as RAE/REF outputs. Thus, high-quality research submitted by UK academics for assessment is often used to support societal impact as well (Digital Science, 2016).

Publications referenced in REF2014 ICSs have significantly higher altmetric scores compared to publications submitted to REF as research outputs, suggesting that publications referenced in ICSs are more likely to attract social media attention (Bornmann et al., 2019). Moreover, grants linked to ICSs are generally longer, higher in value, and result in more publications and greater collaboration (Reddick et al., 2022). It would also be interesting to assess whether systematically gathered policy citations (Szomszor & Adie, 2022) are good evidence of societal impacts.

3.3 | Content analyses of ICSs: The nature of the impacts claimed

Since understanding the impacts claimed in ICSs can be complex and multifaceted, text mining or keyword frequency may not fully capture all aspects of societal impacts, and content analysis may be more useful. In health-related fields, around two-thirds of ICSs influenced clinical guidelines and over half benefited clinical policies and practices (Greenhalgh & Fahy, 2015). Similarly, 93% of cancer ICSs cited clinical trials and claimed national or international health policy impacts (Hanna et al., 2020; see also Rivera et al., 2019). In social sciences, ICSs commonly reported policy-related impacts. For example, anthropology research claimed diverse impacts on UK, EU, or UN policies (Jarman & Bryan, 2015) and education researchers influenced parliamentary committees and policymakers (Cain & Allan, 2017; Laing et al., 2018). Social Work and Social Policy research also impacted policymaking through policy documents and consultations (Smith & Stewart, 2017) but business impact was mostly evidenced through testimonials (Hughes et al., 2019). In arts and humanities, museums, galleries exhibitions were often used to support impact claims (Brook, 2018; Kousha et al., 2024). In STEM fields, instrumental, environmental, or technological impacts were commonly claimed (Meagher & Martin, 2017; Midmore, 2017; Robbins et al., 2017).

4 | FACTORS ASSOCIATING WITH HIGHER SCORING ICSs

As mentioned in the introduction, traditional machine learning can be used to distinguish between ICSs from high- and low-scoring departments (Williams et al., 2023). Another study used regression to predict departmental average ICS scores. It analyzed DOIs from underpinning research in 1469 REF2014 ICSs in the nine physical sciences, engineering, and mathematics UoAs (grouped together as Panel B in the REF organizational structure) and found that the proportion of underpinning research articles with non-zero altmetric scores (i.e., some kind of online attention) had a statistically insignificant influence in two of the six regressions, with the citation rates of the underpinning research being much stronger predictors. The most accurate model had an R^2 of 0.283, which corresponds to a correlation of 0.532 (Wooldridge & King, 2019). This covered a minority of UoAs, used early and incomplete altmetric data, and concerned departmental averages rather than individual ICS scores. Moreover, it seemed to primarily leverage the quality of the department's research (using citation rates as a proxy) for its predictions so does not seem to be a helpful approach because the purpose of the ICS REF element is not to assess research quality. Another study found no significant correlations between mean normalized citation counts or altmetric scores and average institutional REF2014 ICS scores (Ravenscroft et al., 2017).

Other studies have investigated factors associated with higher scores without trying to make predictions. A moderate correlation ($R^2 = 0.37$) was found between REF2014 departmental average output scores (i.e., the average of the scores given to the department's assessed outputs—mainly journal articles) and departmental average ICS scores for UoA 19 Business and Management Studies, suggesting that high-quality research outputs do not necessarily generate high-impact research (Kellard & Śliwa, 2016). This difference might be due to the selection of datasets (Panel B vs. multiple subjects), which can influence the strength of correlation between societal impact and peer-review scores (Thelwall et al., 2023). A linguistic analysis of 124 high-scoring (3* or 4*) and 93 low-scoring (1* or 2*) REF2014 case studies found that high-scoring case studies were easier to read, providing "specific and high-magnitude articulations of significance and reach." However, low-scoring case studies often focused more on the pathways to impact rather than on the claimed impacts (Reichard et al., 2020). Positive correlations have also been found between submission size and average scores in panels A ($R^2 = 0.242$), B ($R^2 = 0.389$), C ($R^2 = 0.359$), and D ($R^2 = 0.120$), suggesting that bigger departments tended to generate disproportionately more societal impact (Pinar & Unlu, 2020), although the relationship is not strong.

5 | METHODS

The research design for RQ1 was to submit all REF2021 ICSs to ChatGPT to request a quality score; then, within

each Unit of Assessment, correlate the ChatGPT score with the mean departmental ICS score. The departmental mean is used as a proxy for the actual score for each ICS (following: Williams et al., 2023) because the former is not public (and has been destroyed as a policy decision) but the latter is. Departmental means are suitable proxies because there is considerable variation between departments in their average ICS scores. The results were then compared for five different inputs.

For RQ2, the research design was to repeat the above with the most promising input but varying the system instructions. Although less promising inputs could also have been checked with different system instructions, this parsimonious approach was taken due to cost considerations. As described below, the main problem with the default system instructions was that they allowed ChatGPT to be too generous, so adjustments were made to make it stricter (see below for details).

For RQ3, average scores and correlations were compared between UoAs.

5.1 | Data

All UK ICSs were downloaded in a spreadsheet source from the official website (results2021.ref.ac.uk/impact). This includes the text of the five sections (1. Summary of the impact, 2. Underpinning research, 3. References to the research, 4. Details of the impact, 5. Sources to corroborate the impact), the UoA and submitting institution, and institutional split, if any (i.e., two submissions from the same institution to the same UoA that are flagged to be analyzed separately—typically this might be for two departments/schools within the same UoA). Here, the combination of UoA and institution will be termed a department for convenience. Thus, for example, the Sociology ICS from the University of Sheffield will be assumed to be from a sociology department even though it might be from a combination of departments and institutes. Thus, the final data consisted of the five ICS sections, along with the identity of the submitting "department."

To clarify the REF terminology (results2021.ref.ac.uk), REF2021 is split into four disciplinary Panels (A, B, C, D), each containing at least six UoAs. Higher Education Institutions (HEIs) can submit sets of ICSs (and other assessed work) to any or all the 34 UoAs, with the number of ICSs related to the number of research-active staff within the scope of the UoA. For example, Sheffield University submitted 10 ICSs as part of its "submission" to UoA 12 Engineering in Panel B. The people who wrote and contributed to these ICSs would normally be in a department or school within the scope of UoA 12 (e.g., School of Mechanical, Aerospace and Civil Engineering), but do not have to be. They may instead be in an institute or faculty within UoA 12 (e.g., Faculty of Engineering) or in an unrelated department but contributing to an interdisciplinary project (e.g., as a statistician or artist). The term "department" is used in this article because it is more intuitive than "submission" even though it is less accurate.

The ICS spreadsheet does not contain the individual ICS scores because these have been destroyed. Instead, the ICS average for each submitted department was obtained from a second spreadsheet of all REF results (results2021.ref.ac.uk/filters/unit-of-assessment), which reports (in the rows labeled "impact") the percentage of ICS scores achieving each of 0, 1*, 2*, 3*, and 4*. The weighted average of these scores was used as the estimated ICS score for each submission. In the few cases where an institution had multiple submissions to a single UoA, these were treated as separate (using the letter identifiers in both spreadsheets).

5.2 | ChatGPT setup

ChatGPT 4o-mini was used, which was the latest model at the time of the study. It seems to be marginally less powerful than ChatGPT 4o (Thelwall, 2024, 2025a), but was 10 times cheaper at the time of writing, so it was a more practical choice for this paper and for applications. The API was used rather than the web interface because the API does not retain the submitted information to train the model, so it is possible to run repeated tests with the same data without inducing bias.

Each ChatGPT session can contain system instructions as well as a specific request. This system information can be used to configure ChatGPT for a task. The REF guidelines were used for the assessment of ICSs as the system instructions. Small stylistic changes were made to adapt them to the way in which the ChatGPT documentation reports examples of system instructions. This style change may improve ChatGPT's ability to ingest the information.

The user prompt for ChatGPT was the phrase "Score the following impact case study": then the ICS title followed by a newline, and the five headings, each followed by a newline, the contents and another newline (see Appendix B example). Five input variants were compared: The ICS title alone, the ICS title and summary, the ICS title, summary, and description of impacts, all descriptive fields (title, summary, underpinning research, details of the impact) and all sections (title, summary, underpinning research, references to the research, details of the impact, sources to corroborate the impact). These

combinations were chosen to be internally coherent, expanding from little information (just the title) to complete information.

The ChatGPT response to the above system information and user prompt should be a report on the ICS and a score: 1*, 2*, 3*, or 4*. The system information did not define the Unclassified score, so this was not an option for ChatGPT. Unclassified scores were very rare and the official reasons for them include "the impact was not underpinned by excellent research produced by the submitted unit" (REF, 2021a, p. 85). This is a complex condition that is not easily checkable by the system, because it would require assessing the referenced underpinning research for research excellence based on the citation alone. Thus, excluding a complex description of a rare category seems like a useful simplifying step. While rare categories are a recognized problem in machine learning research (Fernández et al., 2018) this is not directly relevant here because there is no learning stage in the "zero-shot" method used. More relevant is the observation that for human classification studies, rare classes cause problems because, if errors randomly occur, higher proportions of allocations to rarer classes are likely to occur through random errors (e.g., see Zec et al., 2017). Similarly in the current study, adding a rare category with a complex description seems likely to increase random errors more than accurate classifications. Scores were extracted from the reports automatically by pattern matching (see Webometric Analyst, AI menu; github.com/MikeThelwall/Webometric_Analyst) and, when this did not work, with the second author reading the ChatGPT output to extract the score.

Previous research suggests that more accurate results can be gained for complex scoring tasks by repeating the ChatGPT request multiple times (e.g., 5 or 30) and taking the mean (Thelwall, 2024, 2025a). To identify the optimal inputs and system instructions for RQ1, each ICS was submitted 5 times and the mean was used as the final score. Five times has been sufficient previously to establish the pattern (Thelwall, 2025a, 2025b; Thelwall et al., 2025; Thelwall & Yaghi, 2024) and additional iterations were not necessary since the goal was not to definitively establish the maximum predictive power. Nevertheless, 30 iterations were used (i.e., an additional 25) for the optimal prompting strategy to give the most precise data for RQ2 and RQ3.

The mean was used rather than the mode or median, despite the (usually) ordinal nature of the scores, because previous research has found this, followed by scaling, to be an effective prediction mechanism (Thelwall, 2025a). As the results below show, almost all medians would be 4*, and the mean has the advantage of capturing a degree of uncertainty about the 4* score in a simple way. More fundamentally, REF scores are inherently scales, and an ICS could theoretically have a score of 1.5*, for example, but the REF process forces human evaluators to round their scores (initially made on a nine-point scale) before reporting them.

5.3 | Analyses

For RQ1, the mean departmental scores were correlated with the ChatGPT predictions from the default system prompts separately for each UoA. Aggregation by UoA is appropriate because the evaluators assigning the original grades are organized by UoA and the types of impact can vary substantially between UoAs. Pearson correlations were used because the data were not highly skewed.

The correlations calculated in this way may be underestimates of the underlying correlation because the ChatGPT scores are not correlated against the true REF scores but against the departmental means, adding a degree of indirectness.

For RQ2, system prompt variations were designed and assessed as above. They were designed after RQ1 had been evaluated and in response to the observation that the ICS scores given tended to be the maximum, so improved scores might be obtained by encouraging ChatGPT to be stricter. This was achieved in two ways: making the descriptions of each score level more stringent and explicitly telling ChatGPT to be "very strict." ChatGPT was also suggested to use half scores to encourage it to give a 3.5* score to articles that might otherwise have attracted a 4*. The system prompt was also modified to ask ChatGPT not to explain its score in case that helped. These modifications were made as follows.

- **Strict**: Replacing the score level descriptions with the descriptors used for REF publications ("World leading" instead of "Outstanding" impacts; "Internationally excellent" instead of "Very considerable"; "Internationally recognized" instead of "Considerable"; "Nationally recognized" instead of "Recognised but modest").
- **Very strict**: As above and "You are an academic expert" replaced with "You are a very strict academic expert."
- **Very strict with half scores**: As above and, "Use half points if a case study is between two scores" added after the score descriptions.
- **Very strict with half scores, score only**: As above but, "You will provide a score of 1* to 4* alongside a detailed justification" changed to "You will provide a score of 1* to 4* without any explanation."

For RQ3, average scores were compared between UoAs and against average REF scores to assess whether ChatGPT favors some disciplines over others.

6 | RESULTS

6.1 | RQ1: Accuracy of the default ChatGPT for different inputs

The correlations between ChatGPT scores and departmental profiles were positive overall and statistically significantly different from 0 with $p < 0.001$, but the average scores were high and almost always 4* if enough information from the ICSs was entered (Table 1). The highest correlation is for the partial information from the title and summary. There is clearly a problem with ChatGPT overestimating the quality of the submitted ICS. For example, for 99% of the entire ICSs entered, ChatGPT gave a score of 4* all five times. The average of the institutional ICS REF scores was 3.242, which is substantially lower than the average ChatGPT scores.

For the input with the highest correlation overall, Title + Summary, the correlations between departmental average ICS scores and ChatGPT scores were positive for all UoAs, statistically significantly different from zero in nearly all (31 out of 34) and strong in some (Figure 1). There was a slight tendency for UoAs receiving higher average ChatGPT scores to have weaker correlations in Figure 1 (Pearson $r = -0.266$, $n = 34$ for average ChatGPT score against average correlation by UoA), suggesting that the high scores may be a problem. For all input sets, ChatGPT is therefore too generous in frequently giving top scores to ICSs that should not get them.

6.2 | RQ2: Comparison of different ChatGPT system prompts

Using stricter system prompts lowered the average ChatGPT scores slightly and slightly increased the correlation between the ChatGPT scores and the departmental average REF scores (Table 2). The differences are not statistically significant, despite the large sample sizes. Nevertheless, the *Very strict with half scores* system instructions seem to be optimal.

Repeating an additional 25 times for the most accurate system configuration (very strict with half scores) gives a higher correlation of 0.356, with each individual round tending to increase the correlation slightly (Figure 2).

The increased overall correlation is reflected in higher correlations for individual UoAs, on average, and the correlations are statistically significantly different from 0 in all UoAs except one (Figure 3).

If the ChatGPT scores for all ICSs associated with a department are averaged (with the arithmetic mean), then this gives a figure that is directly comparable to the departmental REF scores because both are averaged across all ICSs (although some redacted ICSs are missing from the current data). As expected, after this averaging the correlations between ChatGPT and the REF tend to be higher, reaching 0.7 in three UoAs (maximum: 0.711) and with a higher minimum correlation (Figure 4).

6.3 | RQ3: Disciplinary differences in ChatGPT scores

ChatGPT tends to give the highest scores for Main Panel A (health and life sciences, UoAs 1-6), the second highest for B (engineering and physical sciences, UoAs 7-12), and the third highest for C (social sciences, UoAs 13-24) (Figure 5). All Main Panel D UoAs (25-34) had lower ChatGPT scores than all other UoAs. Thus, there are clear disciplinary biases in ChatGPT for this task, at least relative to REF assessors. At the UoA level, average REF scores correlate moderately with average ChatGPT scores (Pearson correlation: 0.469, $n = 34$), so ChatGPT tends to give higher scores in UoAs where REF experts also give higher scores.

7 | DISCUSSION

7.1 | Relationship with prior research

The results confirm the evidence derived from REF2014 that ICSs contain score-relevant information that can be

TABLE 1 Average ChatGPT scores and Pearson correlations between average ChatGPT scores per article ($n = 5$ scores each) with the default system prompts and departmental average REF scores.

Input sets for ChatGPT	Correlation (95% CI)	R^2	Mean GPT score[a] (SD)
Title	0.235 (0.212, 0.259)	0.055	3.455 (0.413)
Title + Summary	0.337 (0.315, 0.359)	0.114	3.852 (0.304)
Title + Summary + Details	0.140 (0.116, 0.164)	0.020	3.993 (0.073)
Title + Summary + Details + Underpinning	0.147 (0.123, 0.172)	0.022	3.996 (0.050)
Entire ICS	0.175 (0.151, 0.199)	0.031	3.996 (0.048)

Note: The data is all 6220 public ICS associated with a department with a public ICS score profile. The columns are for different ChatGPT inputs.
[a]The mean score for ICSs (derived from department mean scores) was 3.242.

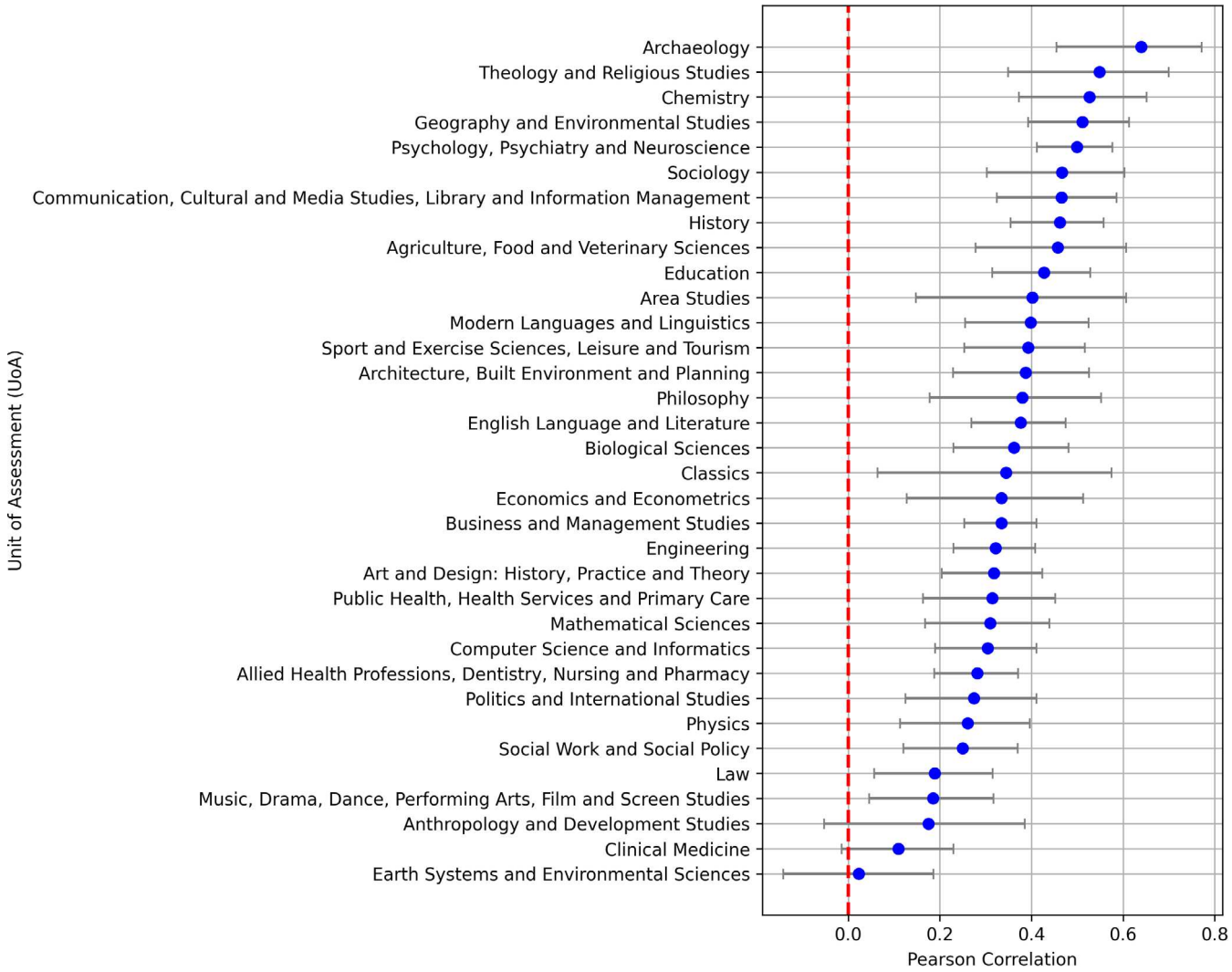

FIGURE 1 Pearson correlations between average ChatGPT score with default system prompts for an ICS and the departmental ICS score, by UoA for the average of five iterations of the default prompt applied to each ICS Title + Summary. Error bars indicate 95% confidence intervals for the UoA.

TABLE 2 Average ChatGPT scores and Pearson correlations between average ChatGPT scores per article ($n = 5$ scores each) for the title and summary with various system prompts and departmental average REF scores for all 6220 public ICSs associated with a department with a public ICS score profile.

System prompt	Correlation (95% CI)	R^2	Mean GPT score[a] (SD)
Default	0.337 (0.315, 0.395)	0.114	3.852 (0.304)
Strict	0.348 (0.326, 0.370)	0.121	3.744 (0.384)
Very strict	0.346 (0.324, 0.368)	0.120	3.641 (0.434)
Very strict with half scores	0.349 (0.327, 0.371)	0.122	3.686 (0.381)
Very strict with half scores, score only	0.344 (0.322, 0.366)	0.118	3.713 (0.358)

[a]The mean score for ICSs (derived from department mean scores) was 3.242.

extracted by automated methods with machine learning (Williams et al., 2023) and go further by finding that departmental average scores can be predicted—admittedly inaccurately—for most ICSs. This suggests, but does not prove, a similar or higher ability to predict individual ICS-level scores. Many of the departmental

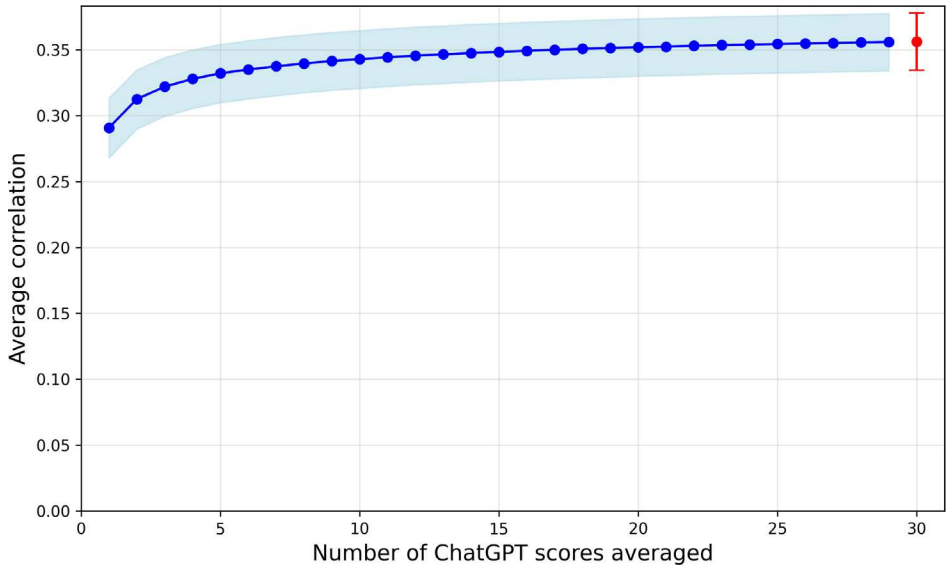

FIGURE 2 Average Pearson correlations between average ChatGPT score ($n = 30$ iterations) with **the very strict with half scores** system prompts for an ICS and the departmental ICS score applied to each ICS Title + Summary. The ribbon from $n = 1$ to 29 indicates 95% confidence intervals within the samples. The error bar for $n = 30$ indicates a population 95% confidence interval for the single $n = 30$ correlation.

level correlations are also higher than the maximum previously found (0.532) from a regression approach that primarily leveraged citation rates for Main Panel B (Wooldridge & King, 2019). From a different perspective, the findings confirm evidence from previous studies that ChatGPT can extract meaningful scores or predictions from various types of academic text (Saad et al., 2024; Thelwall, 2024, 2025a; Thelwall & Yaghi, 2024).

7.2 | Reasons given by ChatGPT for scores below 4*

To investigate why ChatGPT allocated its scores, 100 of its reports (default prompt, Title + Summary input) were selected randomly and the stated reasons for scores identified, if any. Reports recommending 4* did not tend to be informative. For reports recommending a 3*, there was usually a statement about the limitations of the ICS: either that it was limited in scope or that the impact was not transformative.

For limitations in scope, geography was mentioned (e.g., the impact was localized, regional, purely national, or not fully international, such as, "it is contained within a national context (England) without extended or explicit acknowledgment of broader international implications"), or that there was a single or few applicable contexts (e.g., "it primarily influences surgical practices rather than undergoing a transformational effect on a broader scale across multiple sectors beyond health"). These reasons are not convincing because breadth does not have to equate to international reach and expecting health research to transform non-health sectors seems unreasonable.

For limitations in the depth of impact, a lack of evidence and shallow impact were both mentioned. In terms of lack of evidence, "it falls short of being classified as world-leading (4*) primarily because the assessment lacks specific metrics or evidence that convincingly demonstrate a transformative effect on heritage management at a global level" and "while it has shaped policies and practices, the case lacks detail on the measurable outcomes of these changes." For the lack of a transformative impact, "a perceived lack of remarkable transformative effects that would position it at the forefront of international research developments within the biomedical landscape," "primarily due to the lack of extraordinary transformative results," and "while the direct employment figures at the start-ups may seem modest." These reasons seem more convincing than those for scope, especially those claiming a lack of evidence. It is difficult to assess the accuracy of these claims, however, without subject expertise and norm referencing.

In several cases, the report contained no specific criticisms despite allocating a 3* score. For example, two reports concluded, "In summary, the dual aspects of reach through extensive mediums and diverse beneficiaries, combined with significant changes to public engagement with heritage, justify a score of 3*. This reflects an international recognition of the impact, although it is not

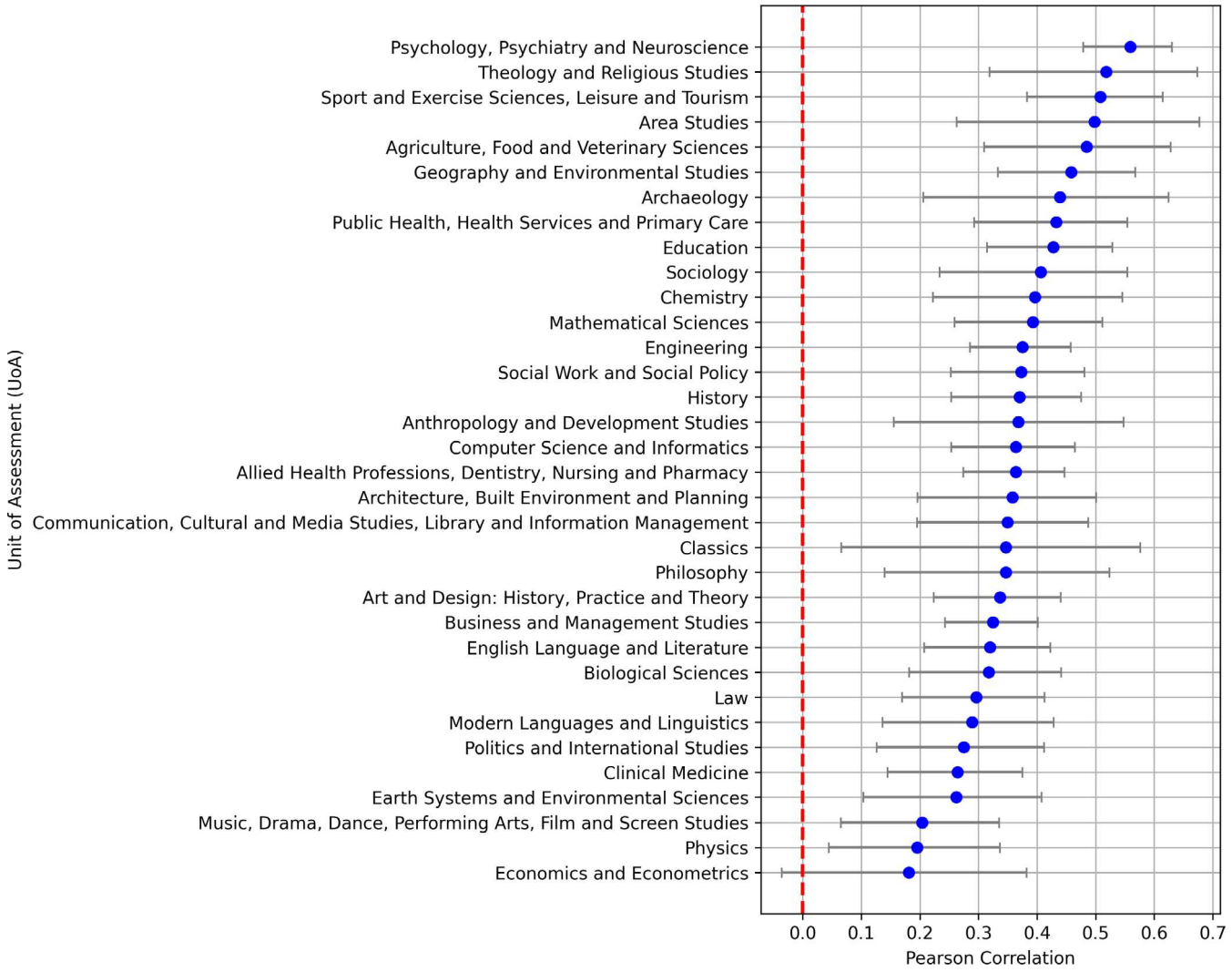

FIGURE 3 Pearson correlations between average ChatGPT score for an ICS and the departmental ICS score, by UoA for the average of 30 iterations of the very strict prompt with half scores applied to each ICS Title + Summary. Error bars indicate 95% confidence intervals for the UoA.

yet to the level of being world-leading or revolutionary" and "Overall, while the case study demonstrates an impressive impact with extensive reach (cross-sector and multi-generational stakeholders) and significant changes in awareness, practices, and personal development, it does not fully achieve the very high benchmarks required for a 4* rating. The impacts, while notable, may not exemplify the world-leading status that would warrant the highest score; they display excellence, particularly within international contexts, hence the 3* rating is appropriate."

In summary, the reasons given for scores were often weak or non-existent but, in some cases, seemed to point to genuine limitations. It is possible that such cases drive the positive correlations, or that the positive correlations are driven by weaker associations in the data that do not translate into specific reasons in the ChatGPT output.

7.3 | Disciplinary differences

There were substantial differences in average ChatGPT scores between UoAs. These would be problematic if ChatGPT were to be used to compare work from different disciplines, unless with norm referencing. For the REF, ChatGPT scores could be scaled separately for each UoA, although this would lessen their value because it would not consider the fact that there are genuine differences in average ICS scores between UoAs—potentially changing between REFs.

From a common-sense perspective, the ChatGPT average score differences between UoAs seem likely to reflect underlying differences in the depth of impact and possibly also the breadth of impact. For example, it seems likely that much Clinical Medicine impact (average ChatGPT score: 3.98) would tend to be more consequential

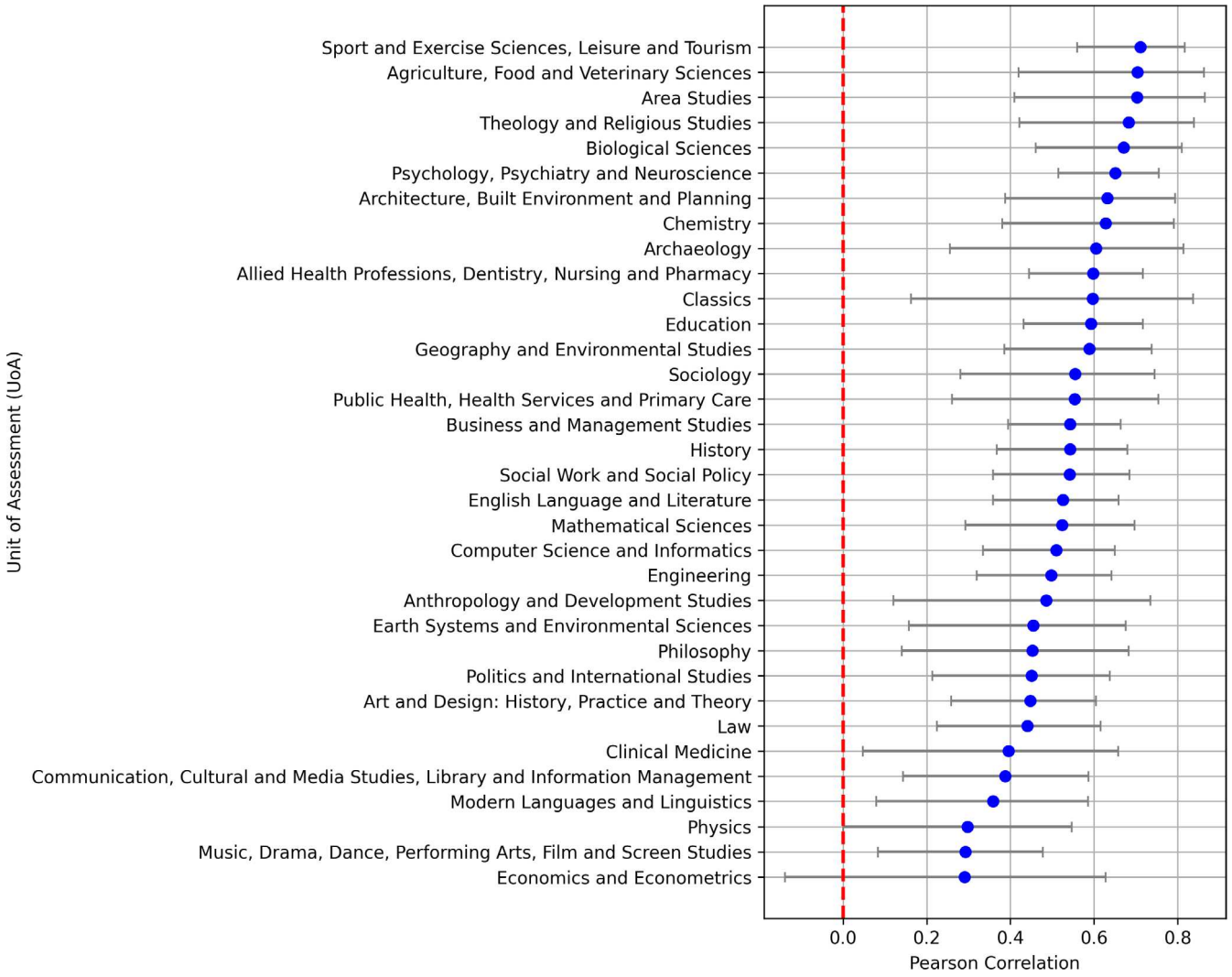

FIGURE 4 Pearson correlations between *departmental average ChatGPT scores for ICSs* and the departmental ICS score, by UoA for the average of 30 iterations of the very strict prompt with half scores applied to each ICS Title + Summary. Error bars indicate 95% confidence intervals for the UoA.

than impact from Music, Drama, Dance, Performing Arts, Film and Screen Studies (average ChatGPT score: 3.35), but this is perhaps a philosophical issue.

8 | LIMITATIONS

The results are limited to a single country (the UK), a single conceptualization of societal impact value (that of the REF) and a single format for describing research impact (the REF ICS template). They are also limited to a single iteration of the REF. It seems plausible that similar results would be obtained for other countries and other impact claim documents if they were standardized, had clear evaluation guidelines that could be used for system prompts, and included a relatively brief summary. Evaluating more ad hoc impact claims would be particularly difficult because of the apparent tendency of ChatGPT to give high scores to long documents.

The results may be different with other LLMs (although Google Gemini 1.5 Flash performs similarly to ChatGPT 4o-mini on a related research evaluation task: Thelwall, 2025b) and for updated and larger versions of ChatGPT (compared to 4o-mini). Better results may also have been obtained with other prompting strategies than those tried. The data also is restricted by the use of departmental average ICS scores rather than individual ICS scores, and the implicit simplifying assumption that the five REF scores (0 to 4*) are equidistant and can be averaged for a department. Importantly, it is not clear whether the positive correlations are at least partly due to "cheating" in the sense of leveraging properties irrelevant to research quality, such as institutional prestige, when

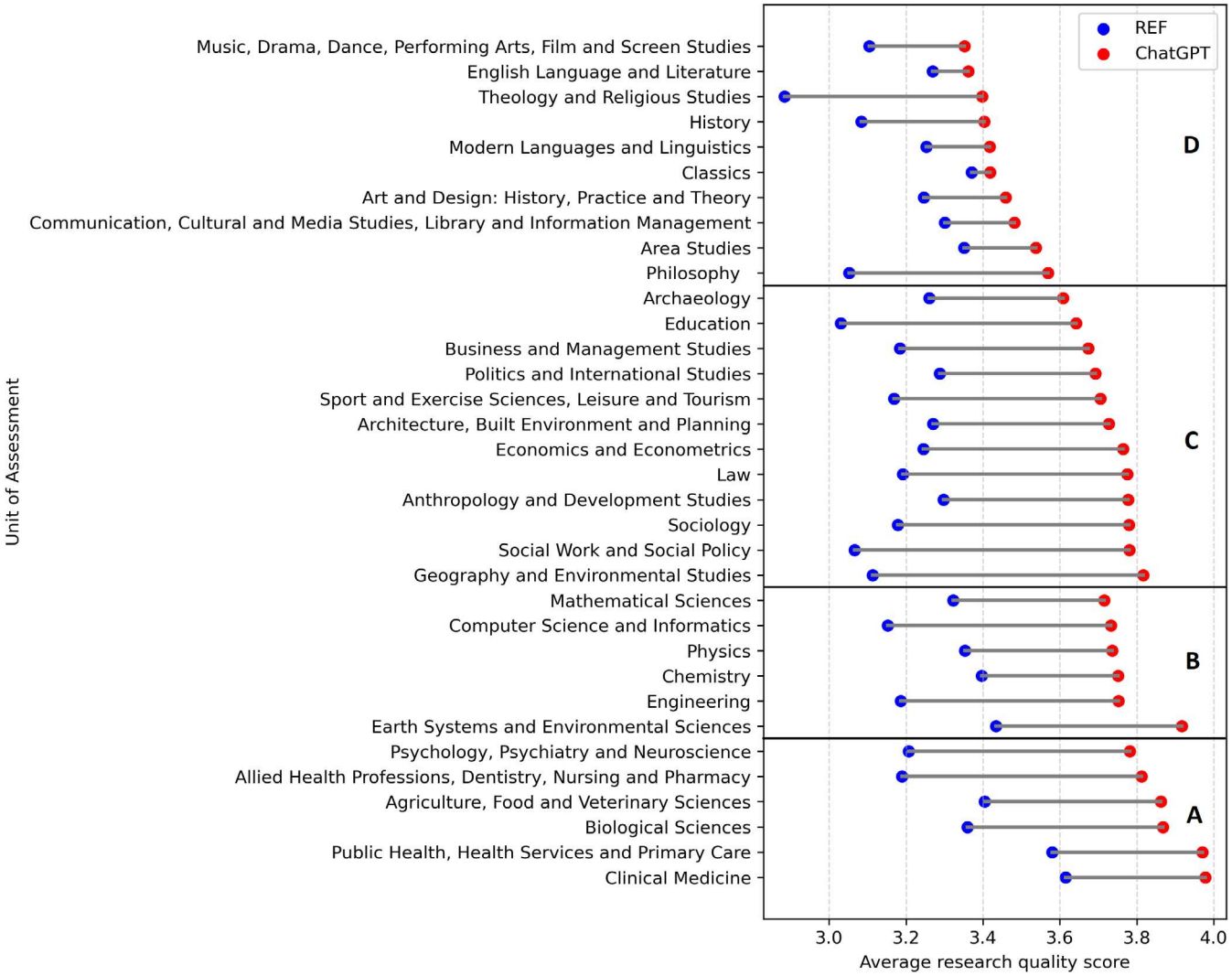

FIGURE 5 Average REF and ChatGPT scores for ICSs, by UoA for the average of 30 iterations of the very strict prompt with half scores applied to each ICS Title + Summary. UoAs are sorted first by main panel, then by average ChatGPT score. For comparability, average REF scores are per ICS rather than per department (i.e., the departmental REF average scores are weighted by the number of ICSs submitted).

making predictions (e.g., see Laurer et al., 2024). More detailed analyses in future research to understand the main influences on ChatGPT's scores and when it is inaccurate may help with this. Finally, the ICS-level correlations reported are indirect.

9 | CONCLUSIONS

The results show that ChatGPT 4o-mini has the ability to estimate impact case study claims for reach and significance for nearly all UoAs at a level above statistical significance, but still inaccurately. It tends to overestimate, and there are substantial differences between disciplines in both the average value of the scores and the extent to which they correlate with (departmental level) human expert scores. This is the first time a practical automated method has been found that has a non-trivial capability to predict ICS scores across the REF (rather than differentiating between top and bottom 20% ICSs: Williams et al., 2023) or, more generally, to quantify the extent of impact described by academics. Nevertheless, since the highest predictions are derived from the title and summary without the full details of an ICS, it is clear that ChatGPT is making an intelligent guess rather than fully assessing each ICS. Although the human REF2021 experts could, in theory, have also read only the title and summary, it seems inconceivable that many did not carefully read each of their allocated ICSs given the financial and reputational importance of their scores and the relatively short ICS length (e.g., see Samuel & Derrick, 2015; Watermeyer & Chubb, 2019).

The maximum within-UoA departmental-level correlation of 0.711 between departmental REF average and

GPT average score is not high enough to consider replacing expert evaluations of ICSs with AI evaluations, even without considering the systemic implications of such a change. The ICS-level correlations may be high enough for them to be useful in a supporting role, however, such as for cross-checking expert scores or as a second opinion or (together with feedback) to support internal university reviews of potential ICS submissions. For this, the ChatGPT scores should either be scaled to conform to the human expert scales (i.e., typically reducing them or using a lookup table to convert the ChatGPT predictions to corresponding likely human predictions) or used to rank sets of ICSs, depending on the task. Nevertheless, ChatGPT scores should not be used in mixed discipline areas where their estimates can be expected to vary substantially for disciplinary reasons.

ACKNOWLEDGMENTS

Mike Thelwall is funded by the Economic and Social Research Council (ESRC), UK (APP43146).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Kayvan Kousha https://orcid.org/0000-0003-4827-971X
Mike Thelwall https://orcid.org/0000-0001-6065-205X

REFERENCES

Adams, J., Loach, T., & Szomszor, M. (2015). The diversity of UK research and knowledge. Analyses from the REF impact case studies. Digital Research Reports. Retrieved from https://www.digital-science.com/resource/diversity-of-uk-research/

Bonaccorsi, A., Chiarello, F., & Fantoni, G. (2021). Impact for whom? Mapping the users of public research with lexicon-based text mining. *Scientometrics*, *126*(2), 1745–1774. https://doi.org/10.1007/s11192-020-03803-z

Bornmann, L., Haunschild, R., & Adams, J. (2019). Do altmetrics assess societal impact in a comparable way to case studies? An empirical test of the convergent validity of altmetrics based on data from the UK research excellence framework (REF). *Journal of Informetrics*, *13*(1), 325–340. https://doi.org/10.1016/j.joi.2019.01.008

Brook, L. (2018). Evidencing impact from art research: Analysis of impact case studies from the REF 2014. *Journal of Arts Management, Law, and Society*, *48*(1), 57–69. https://doi.org/10.1080/10632921.2017.1386148

Cain, T., & Allan, D. (2017). The invisible impact of educational research. *Oxford Review of Education*, *43*(6), 718–732. https://doi.org/10.1080/03054985.2017.1316252

Chowdhury, G., Koya, K., & Philipson, P. (2016). Measuring the impact of research: Lessons from the UK's research excellence framework 2014. *PLoS One*, *11*(6), e0156978. https://doi.org/10.1371/journal.pone.0156978

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Scarecrow Press.

Derrick, G. (2018). *The evaluators' eye: Impact assessment and academic peer review*. Springer.

Digital Science. (2016). *Publication patterns in research underpinning impact in REF2014: A report to HEFCE by digital science*. Digital Science. Retrieved from https://dera.ioe.ac.uk/26933/1/2016_refimpact.pdf

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905.

Greenhalgh, T., & Fahy, N. (2015). Research impact in the community-based health sciences: An analysis of 162 case studies from the 2014 UK research excellence framework. *BMC Medicine*, *13*(1), 1–12. https://doi.org/10.1186/s12916-015-0467-4

Hanna, C. R., Gatting, L. P., Boyd, K. A., Robb, K. A., & Jones, R. J. (2020). Evidencing the impact of cancer trials: Insights from the 2014 UK Research Excellence Framework. *Trials*, *21*(1), 1–13. https://doi.org/10.1186/s13063-020-04425-9

Hughes, T., Webber, D., & O'Regan, N. (2019). Achieving wider impact in business and management: Analysing the case studies from REF 2014. *Studies in Higher Education*, *44*(4), 628–642. https://doi.org/10.1080/03075079.2017.1393059

Jarman, N., & Bryan, D. (2015). Beyond the academy: Applying anthropological research, a case study of demonstrating impact in the UK 2014 REF. *Anthropology in Action*, *22*(2), 36–41. https://doi.org/10.3167/aia.2015.220205

Jensen, E. A., Wong, P., & Reed, M. S. (2022). How research data deliver non-academic impacts: A secondary analysis of UK Research Excellence Framework impact case studies. *PLoS One*, *17*(3), e0264914. https://doi.org/10.1371/journal.pone.0264914

Jordan, K. (2020). Examining educational technology and research impact: The two roles of E-learning and related terms in the 2014 REF impact case studies. *Research in Learning Technology*, *28*, 1–17. https://doi.org/10.25304/rlt.v28.2306

Jordan, K., & Carrigan, M. (2018). How was social media cited in 2014 REF impact case studies? Impact of Social Sciences Blog. Retrieved from https://blogs.lse.ac.uk/impactofsocialsciences/2018/06/06/how-was-social-media-cited-in-2014-ref-impact-case-studies/

Kellard, N. M., & Śliwa, M. (2016). Business and management impact assessment in research excellence framework 2014: Analysis and reflection. *British Journal of Management*, *27*(4), 693–711.

King's College London and Digital Science. (2015). *The nature, scale and beneficiaries of research impact: An initial analysis of Research Excellence Framework (REF) 2014 impact case studies*. HEFCE. Retrieved from https://www.kcl.ac.uk/policy-institute/assets/ref-impact.pdf

Kousha, K., Stuart, E., Abdoli, M., & Thelwall, M. (2024). How do museums and galleries help academics create societal impact? An analysis of the UK REF2021 impact case studies. *Scientometrics*, *129*, 7759–7782. https://doi.org/10.1007/s11192-024-05180-3

Kousha, K., Thelwall, M., & Abdoli, M. (2021). Which types of online evidence show the nonacademic benefits of research? Websites cited in UK impact case studies. *Quantitative Science Studies*, *2*(3), 864–881. https://doi.org/10.1162/qss_a_00145

Koya, K., & Chowdhury, G. (2020). Measuring impact of academic research in computer and information science on society. In *Proceedings of the 2020 2nd Asia Pacific information technology conference* (pp. 78–85). ACM. https://doi.org/10.1145/3379310.3379312

Laing, K., Mazzoli Smith, L., & Todd, L. (2018). The impact agenda and critical social research in education: Hitting the target but missing the spot? *Policy Futures in Education*, *16*(2), 169–184. https://doi.org/10.1177/1478210317742214

Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2024). On measurement validity and language models: Increasing validity and decreasing bias with instructions. *Communication Methods and Measures*, *19*, 1–17. https://doi.org/10.1080/19312458.2024.2378690

Meagher, L. R., & Martin, U. (2017). Slightly dirty maths: The richly textured mechanisms of impact. *Research Evaluation*, *26*(1), 15–27. https://doi.org/10.1093/reseval/rvw024

Midmore, P. (2017). The science of impact and the impact of agricultural science. *Journal of Agricultural Economics*, *68*(3), 611–631. https://doi.org/10.1111/1477-9552.12242

Morgan Jones, M., Manville, C., & Chataway, J. (2017). Learning from the UK's research impact assessment exercise: A case study of a retrospective impact assessment exercise and questions for the future. *Journal of Technology Transfer*, *47*, 722–746. https://doi.org/10.1007/s10961-017-9608-6

Morrow, E. M., Goreham, H., & Ross, F. (2017). Exploring research impact in the assessment of leadership, governance and management research. *Evaluation*, *23*(4), 407–431.

Parks, S., Ioppolo, B., Stepanek, M., & Gunashekar, S. (2018). *Guidance for standardising quantitative indicators of impact within REF case studies*. RAND Europe.

Pinar, M., & Unlu, E. (2020). Evaluating the potential effect of the increased importance of the impact component in the Research Excellence Framework of the UK. *British Educational Research Journal*, *46*(1), 140–160. https://doi.org/10.1002/berj.3572

Pollitt, A., Sreenan, N., Grant, J., Szomszor, M., Leeworthy, D., & Hughes, D. (2023). The impacts of research from Welsh universities: Full report. Retrieved from https://www.learnedsociety.wales/wp-content/uploads/2023/10/The-impacts-of-research-from-Welsh-universities-Final.pdf

Raaper, R., & Olssen, M. (2015). Mark Olssen on neoliberalisation of higher education and academic lives: An interview. *Policy Futures in Education*, *14*(2), 147–163. https://doi.org/10.1177/1478210315610992

Ravenscroft, J., Liakata, M., Clare, A., & Duma, D. (2017). Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. *PLoS One*, *12*(3), e0173152.

Reddick, G., Malkov, D., Sherbon, B., & Grant, J. (2022). Understanding the funding characteristics of research impact: A proof-of-concept study linking REF 2014 impact case studies with Researchfish grant agreements. *F1000Research*, *10*, 1291. https://doi.org/10.12688/f1000research.74374.2

REF. (2021a). Guidance on submissions to REF 2021. Retrieved from https://2021.ref.ac.uk/media/1447/ref-2019_01-guidance-on-submissions.pdf

REF. (2021b). Panel criteria and working methods on REF 2021. Retrieved from https://2021.ref.ac.uk/media/1450/ref-2019_02-panel-criteria-and-working-methods.pdf

REF. (2021c). Impact case study database: Securing the Legacy of the Late Spanish Filmmaker Bigas Luna. Retrieved from https://results2021.ref.ac.uk/impact/3c81f543-4e15-4a90-9fae-db49d6769336?page=1

Reichard, B., Reed, M. S., Chubb, J., Hall, G., Jowett, L., Peart, A., & Whittle, A. (2020). Writing impact case studies: A comparative study of high-scoring and low-scoring case studies from REF2014. *Palgrave Communications*, *6*(1), 1–17. https://doi.org/10.1057/s41599-020-0394-7

Rivera, S. C., Kyte, D. G., Aiyegbusi, O. L., Slade, A. L., McMullan, C., & Calvert, M. J. (2019). The impact of patient-reported outcome (PRO) data from clinical trials: A systematic review and critical analysis. *Health and Quality of Life Outcomes*, *17*(1), 156. https://doi.org/10.1186/s12955-019-1220-z

Robbins, P. T., Wield, D., & Wilson, G. (2017). Mapping engineering and development research excellence in the UK: An analysis of REF2014 impact case studies. *Journal of International Development*, *29*(1), 89–105. https://doi.org/10.1002/jid.3255

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

Samuel, G. N., & Derrick, G. E. (2015). Societal impact evaluation: Exploring evaluator perceptions of the characterization of impact under the REF2014. *Research Evaluation*, *24*(3), 229–241. https://doi.org/10.1093/reseval/rvv007

Smith, K. E., & Stewart, E. (2017). We need to talk about impact: Why social policy academics need to engage with the UK's research impact agenda. *Journal of Social Policy*, *46*(1), 109–127. https://doi.org/10.1017/S0047279416000283

Stevenson, C., Grant, J., Szomszor, M., Ang, C., Kapoor, D., Gunashekar, S., & Guthrie, S. (2023). *Data enhancement and analysis of the REF 2021 Impact Case Studies*. RAND. Retrieved from https://www.rand.org/content/dam/rand/pubs/research_reports/RRA2100/RRA2162-1/RAND_RRA2162-1.pdf

Szomszor, M., & Adie, E. (2022). Overton: A bibliometric database of policy document citations. *Quantitative Science Studies*, *3*(3), 624–650.

Terämä, E., Smallman, M., Lock, S. J., Johnson, C., & Austwick, M. Z. (2016). Beyond academia–Interrogating research impact in the research excellence framework. *PLoS One*, *11*(12), e0168533. https://doi.org/10.1371/journal.pone.0168533

Thelwall, M. (2024). Can ChatGPT evaluate research quality? *Journal of Data and Information Science*, *9*(2), 1–21. https://doi.org/10.2478/jdis-2024-0013

Thelwall, M. (2025a). Evaluating research quality with large language models: An analysis of ChatGPT's effectiveness with different settings and inputs, *10*(1), 1–25. https://doi.org/10.2478/jdis-2025-0011

Thelwall, M. (2025b). Is Google Gemini better than ChatGPT at evaluating research quality. *Journal of Data and Information Science*. https://doi.org/10.2478/jdis-2025-0014

Thelwall, M., Jiang, X., & Bath, P. (2025). Estimating the quality of published medical research with ChatGPT. *Information Processing & Management*, *62*(4), 104123. https://doi.org/10.1016/j.ipm.2025.104123

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., & Levitt, J. (2023). Do altmetric scores reflect article quality? Evidence from the UK Research Excellence Framework 2021. *Journal of the Association for Information Science and Technology*, *74*(5), 582–593. https://doi.org/10.1002/asi.24751

Thelwall, M., & Yaghi, A. (2024). In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results. Retrieved from https://arxiv.org/abs/2409.16695

Vanlee, F. (2024). Gender, sexuality and non-academic impact: Exploring the role of gender and sexuality (studies) in the 2021 "research excellence framework". *DiGeSt-Journal of Diversity and Gender Studies*, *11*, 1. https://doi.org/10.21825/digest.86011

Watermeyer, R., & Chubb, J. (2019). Evaluating "impact" in the UK's Research Excellence Framework (REF): Liminality, looseness and new modalities of scholarly distinction. *Studies in Higher Education*, *44*(9), 1554–1566. https://doi.org/10.1080/03075079.2018.1455082

Wilkinson, C. (2019). Evidencing impact: A case study of UK academic perspectives on evidencing research impact. *Studies in Higher Education*, *44*(1), 72–85. https://doi.org/10.1080/03075079.2017.1339028

Williams, K., Michalska, S., Cohen, E., Szomszor, M., & Grant, J. (2023). Exploring the application of machine learning to expert evaluation of research impact. *PLoS One*, *18*(8), e0288469. https://doi.org/10.1371/journal.pone.0288469

Wooldridge, J., & King, M. B. (2019). Altmetric scores: An early indicator of research impact. *Journal of the Association for Information Science and Technology*, *70*(3), 271–282. https://doi.org/10.1002/asi.24122

Zec, S., Soriani, N., Comoretto, R., & Baldi, I. (2017). High agreement and high prevalence: The paradox of Cohen's kappa. *Open Nursing Journal*, *11*, 211–218.

Zheng, H., Pee, L. G., & Zhang, D. (2021). Societal impact of research: A text mining study of impact types. *Scientometrics*, *126*(9), 7397–7417. https://doi.org/10.1007/s11192-021-04096-6

How to cite this article: Kousha, K., & Thelwall, M. (2025). Assessing the societal influence of academic research with ChatGPT: Impact case study evaluations. *Journal of the Association for Information Science and Technology*, 1–17. https://doi.org/10.1002/asi.25021

APPENDIX A: SYSTEM PROMPT

You are an academic expert, assessing impact case studies, which describe specific impacts that have occurred from academic research. You will provide a score of 1* to 4* alongside a detailed justification.

For the purposes of this assessment, impact is defined as an effect on, change, or benefit to the economy, society, culture, public policy or services, health, the environment, or quality of life, beyond academia.

Impact includes, but is not limited to, an effect on, change, or benefit to: the activity, attitude, awareness, behavior, capacity, opportunity, performance, policy, practice, process, or understanding of an audience, beneficiary, community, constituency, organization, or individuals in any geographic location, whether locally, regionally, nationally, or internationally.

Impact includes the reduction or prevention of harm, risk, cost, or other negative effects.

Academic impacts on research or the advancement of academic knowledge (whether in the UK or internationally) are excluded, but impacts on students, teaching, or other activities are included.

Impacts will be assessed in terms of their "reach and significance" regardless of the geographic location in which they occurred, whether locally, regionally, nationally, or internationally.

The scoring system used is 1*, 2*, 3*, or 4*, which are defined as follows.

4*: Outstanding impacts in terms of their reach and significance.

3*: Very considerable impacts in terms of their reach and significance.

2*: Considerable impacts in terms of their reach and significance.

1* Recognized but modest impacts in terms of their reach and significance.

You will understand reach as the extent and/or diversity of the beneficiaries of the impact, as relevant to the nature of the impact. Reach will be assessed in terms of the extent to which the potential constituencies, number or groups of beneficiaries have been reached; it will not be assessed in purely geographic terms, nor in terms of absolute numbers of beneficiaries. The criteria will be applied wherever the impact occurred, regardless of geography or location, and whether in the UK or abroad.

You will understand significance as the degree to which the impact has enabled, enriched, influenced, informed, or changed the performance, policies, practices, products, services, understanding, awareness, or wellbeing of the beneficiaries.

You will make an overall judgment about the reach and significance of impacts, rather than assessing each criterion separately. While case studies need to demonstrate both reach and significance, the balance between them may vary at all quality levels. You will exercise your judgment without privileging or disadvantaging either reach or significance.

APPENDIX B: EXAMPLE OF THE CONTENTS OF A QUERY, REDACTED FOR BREVITY, WITH BOLD FONT ADDED FOR CLARITY (THE ICS IS HERE: REF, 2021c)

Score the following impact case study: Securing the Legacy of the Late Spanish Filmmaker Bigas Luna

1. **Summary of the impact**

José Juan Bigas Luna (1946–2013) is one of Spain's most important filmmakers. [...] generated additional income and helped to attract new audiences, as well as transforming audiences' experience and understanding of the films themselves.

2. **Underpinning research**

Fouz Hernández's research on Bigas Luna dates back to a 1999 journal article about the "Iberian Trilogy" (showcased in all the events). [...]

3. **References to the research**

R1. *El legado cinematográfico de Bigas Luna*, edited by Santiago Fouz Hernández (Valencia: Tirant lo Blanch, 2020). ISBN 978-84-1815-595-6, p. 348. Edited book including single-authored introduction and two chapters. [...]

4. **Details of the impact**

The "Bigas Luna Tribute" (henceforth BLT) events organized by Fouz Hernández in close collaboration with Betty Bigas, a Barcelona-based artist and curator, have attracted audiences of approximately 3,000 people, and garnered extensive coverage in print, radio, and screen media (E1, E2), consolidating the global legacy of Bigas Luna's work and enhancing its value. [...]

5. **Sources to corroborate the impact**

E1 Printed media (25 pp.)—including *Diari Ara* (2016), *El País* (2017), *The Age* Arts Supplement (front page, 2017) or *La nación* (2018).

E2 [...]

*Items E1, E2, E3, E5, E7, and E8 contain material in Spanish and Catalan.

APPENDIX C: GLOSSARY

UoA: Unit of Assessment. This is one of 34 broadly field-based areas in which the evaluation and reporting of UK research was organized in REF2021. The number of UoAs sometimes changes between REFs.

HEI: Higher Education Institution. HEIs submit ICSs to a UoA in the REF. These are usually universities but include some other types, such as the Institute of Cancer Research.

ICS: Impact Case Study. This is a 7-page structured claim for non-academic impact from research, as submitted by a UK HEI to the REF for evaluation.

REF: Research Excellence Framework. This is the periodic (e.g., REF2014, REF2021, REF2029?) national research evaluation exercise that assesses research outputs, research environments and non-academic impacts and awards the UK's government block research grants based on the scores awarded.