



# Assessing research with AI support: Beneficial and impartial?

Mike Thelwall

Information School, University of Sheffield



# Overview

- Part 1: Machine learning to estimate post-publication journal article quality scores.
- Part 2: ChatGPT-4 for detecting research quality.
- Artificial intelligence (AI): computer-based human-like problem solving.
- Machine learning (ML): a type of AI involving learning a task from examples.
- Current generative AI tools like ChatGPT are pre-built with machine learning but can be applied to novel problems.



Part 1:

# Can machine learning AI estimate post-publication journal article quality scores?

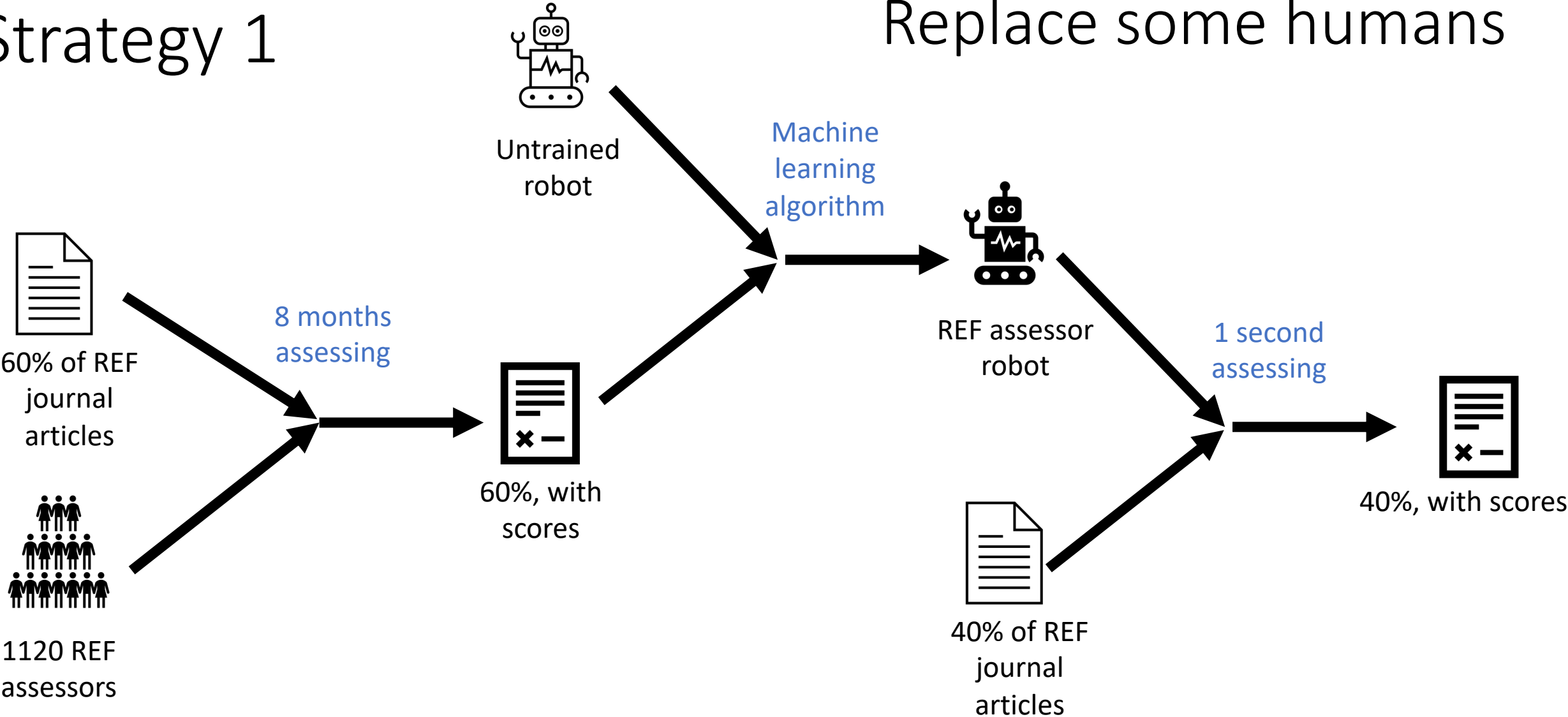


# Context: Expert review for the UK Research Excellence Framework (REF) 2021

- REF2021 expert review examined 185,594 research outputs (mainly journal articles) from 157 UK Higher Education Institutions (HEIs):
- 1120 experts (mainly senior professors) in 34 field-based Units of Assessment (UoAs) scored the work for originality, significance and rigour.
  - 4\*=world leading;
  - 3\*=internationally excellent;
  - 2\*=recognized internationally;
  - 1\*=recognized nationally.

# Strategy 1

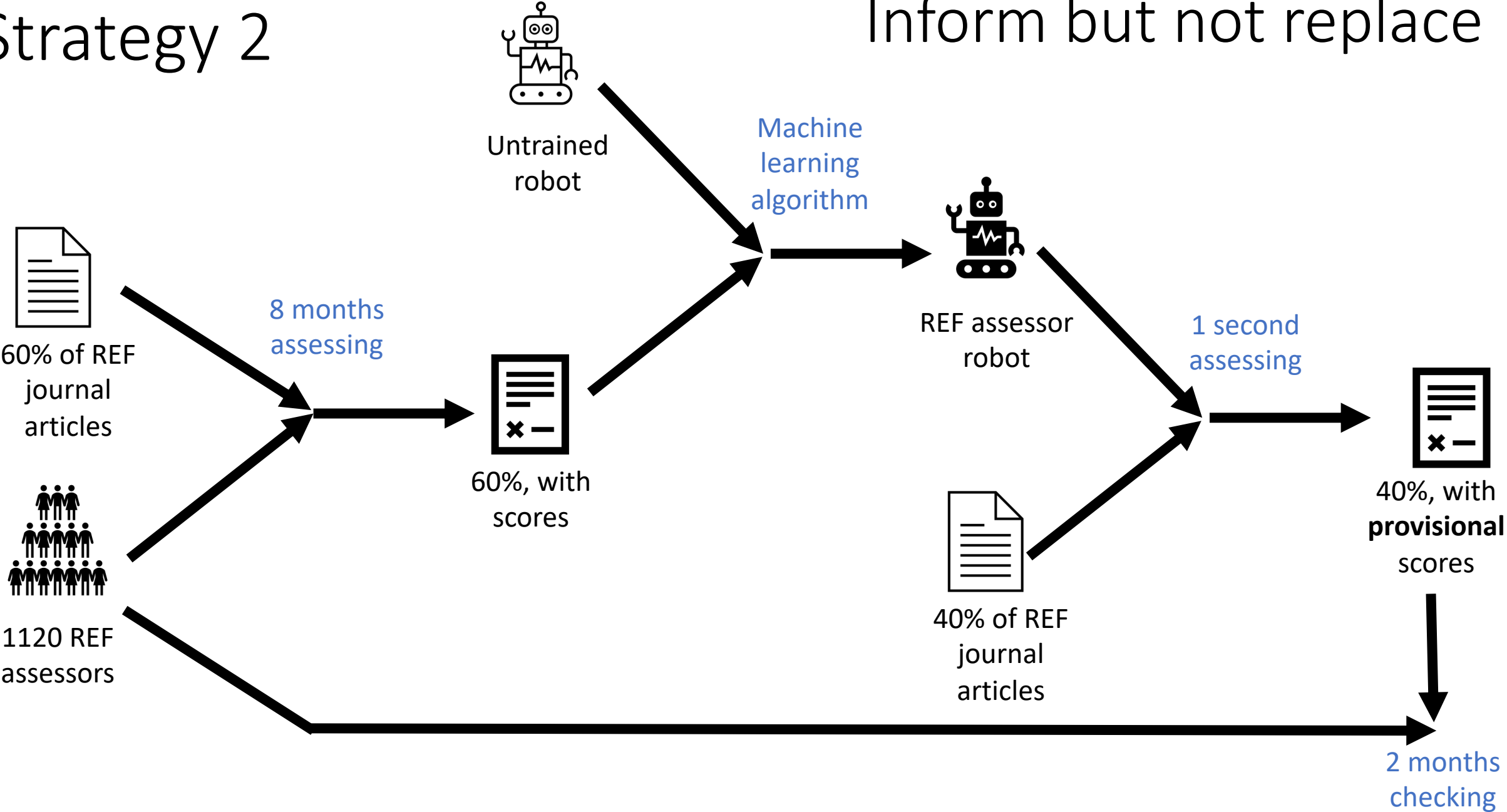
## Replace some humans



All this is done separately for each UoA

# Strategy 2

## Inform but not replace



# Machine learning bibliometric inputs

**Article impact** • **Citation count** (field and year normalized, log transform.)

**Research team size** • **Number of authors, institutions, countries**

**Research team bibliometric record** [ • **Number of Scopus-indexed journal articles of the first author** during the REF period  
• **Average citation rate of Scopus-indexed journal articles by (a) the first author and (b) any author** during the REF period (field and year normalized).

**Article length** • **number of pages (& abstract readability).**

**Journal impact/prestige** • **Journal citation rate** (field and year normalized, log transformed); **journal names.**

**Research description** • **Title and abstract words and phrases.**







# Most powerful **high** and **low** quality predictors

- Journal and article citation rates
- Team size and citation-based achievements
- Some topic/methods terms, e.g.: “**primary outcome**”, “**randomly assigned**” vs. **narrative, education**.
- Many stylistic terms, e.g., “**we**” and “**we show that**” vs. **study, “this article”**.

# Were **departmental-level** predictions accurate enough for the top 8 UoAs?

- Correlations between departmental **total** REF assessor scores and departmental **total** AI scores varied between 0.945 and 0.998, depending on the UoA.
  - So in the best case, the REF money changes little.
- Correlations between departmental average REF assessor scores and departmental **average** AI scores varied between 0.664 and 0.906, depending on the UoA.
- Even in the best case, departmental average score **rankings changed** a bit, which UK academics strongly opposed.

# Biases and error size

- AI predictions are **not** systematically biased:
  - By first author gender (male/female)
  - By submitting author Early Career Researcher status
- AI predictions **are** systematically biased (can compensate):
  - Against higher scoring submissions





ChatGPT

Part 2:

**Can Large Language  
Models like ChatGPT  
estimate research quality?**

REF D Assessor for Main Panel D ▾



## REF D Assessor for Main Panel D

By Mr M A Thelwall

Assists with evaluating journal articles for REF Main Panel D.

Evaluate this research article.

Suggest improvements for this ar...

How does this article meet REF cr...

Compare two articles for REF rele...

📎 Message REF D Assessor for Main Panel D... ⬆

# Strategy

- Configured GPT with the REF research quality guidelines.
- Entered some of my research documents.
- Entered a fake article to see whether it could assess genuine significance.

[ ]Based on the REF criteria for assessing research outputs, the article would be evaluated under the following categories:

1. **Originality**: The study demonstrates originality through its comprehensive, large-scale analysis of academic collaboration patterns across multiple countries and disciplines. It engages with the complex problem of understanding the dynamics of academic collaboration and its impact on citation rates, a topic of significant relevance in the context of global research practices.
2. **Significance**: The research has a significant potential to influence understanding in multiple fields, particularly in research policy and scholarly communication. It provides valuable insights that can inform national and international strategies for research collaboration and funding. The study's findings challenge the uniformity of collaboration practices across different regions and disciplines, highlighting the need for context-specific approaches in research management and policy-making.
3. **Rigour**: The study exhibits intellectual coherence and integrity. It adopts a robust and appropriate methodology, utilizing a large dataset and a well-justified metric for citation impact analysis. The research design is methodologically sound, and the authors provide a thorough analysis of their findings, acknowledging the limitations of their study.

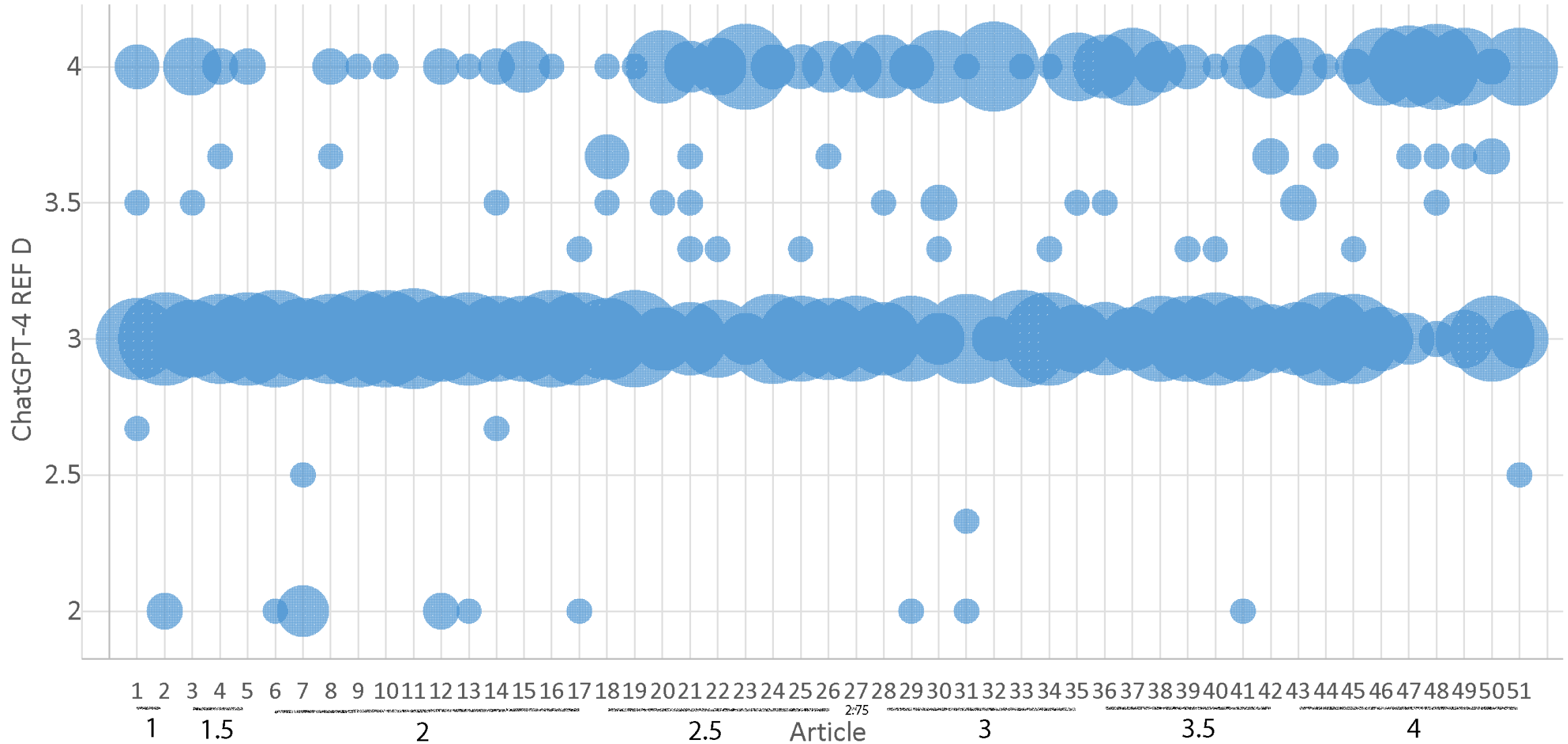
Based on these considerations, the article would be rated as follows:

- **4\* (World-leading)**: While the study is methodologically sound and offers new insights into the patterns of academic collaboration and citation impact, it may not reach the highest standards of world-leading work in terms of providing transformative theoretical advances or developing entirely new methods or paradigms. It offers substantial contributions to the field but may fall slightly short of a pioneering breakthrough that would characterize a definitive 4\* rating.
- **3\* (Internationally excellent)**: The article most likely fits within this category. It is of high quality, contributes important knowledge, and is likely to be influential at an international level. The study is well-designed, applies a robust methodology, and addresses a significant issue in academic research. It is likely to be an important point of reference for future research in this area.

In conclusion, the article is assessed to be of **3\*** quality, indicating that it is internationally excellent in terms of originality, significance, and rigour.

**ChatGPT writes plausible REF evaluations**

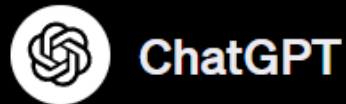
# ChatGPT is variable and poor at detecting REF quality



My estimated REF score for 51 of my papers

Me vs. ChatGPT correlation is 0.3; or 0.5 for Me vs. the average of 15 ChatGPTs.





“Do squirrel surgeons generate more citation impact?”



“[...] The research utilizes a robust methodology, analyzing the field-normalized citation impact of 94,001 US surgery articles through linear regression, **accounting for various variables like author species, publication year, and author numbers.**

In terms of originality, **the work is highly innovative, engaging with a novel approach in citation analysis and addressing a unique aspect of species differences in academic research.**

[...] Given these considerations, the study could be rated as **\*\*3\* (internationally excellent)\*\***. [...]” ChatGPT

(sometimes 4\*)



ChatGPT

# But GPT “knows” that squirrels can’t write research



Me> Can squirrels write academic research journal articles?

ChatGPT>No, squirrels cannot write academic research journal articles. Squirrels are animals without the cognitive capabilities necessary for complex tasks like academic writing. Writing research articles requires a deep understanding of language, the ability to conduct research, and an understanding of academic structures and standards, which are beyond the abilities of squirrels or any other non-human animals. [].

# Summary and lessons for other reviewing

- **Summary:** Machine learning AI might play a minor role in supporting post-publication expert research quality judgements when only *total* scores are important, or to inform but not replace assessors.
- **Grants:** Citations and journal citation rates are the strongest quality predictors, so traditional machine learning (ML) accuracy likely to be very low for grant submission reviewing.
- **Grants:** ChatGPT/LLMs may perform slightly better than traditional ML (since citations irrelevant) but may introduce biases, such as for writing quality. Must avoid using public LLMs due to copyright.
- **Grants:** Might be possible to reliably identify very weak submissions with AI.
- **Narrative CVs:** ChatGPT/LLMs the best option, but high accuracy unlikely.
- **Other issues**
  - AI must be fully evaluated before use.
  - Plausible fake reports are a threat to reviewer integrity.

# References and bibliography

1. Thelwall, M., Kousha, K., Wilson, P. Makita, M., Abdoli, M., Stuart, E., Levitt, J., Knoth, P., & Cancellieri, M. (2023). Predicting article quality scores with machine learning: The UK Research Excellence Framework. *Quantitative Science Studies*, 4(2), 547-573. [https://doi.org/10.1162/qss\\_a\\_00258](https://doi.org/10.1162/qss_a_00258)
2. Kousha, K. & Thelwall, M. (2024). 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>
3. Thelwall, M. (2024). Can ChatGPT evaluate research quality? <https://arxiv.org/abs/2402.05519>