Scientometrics
https://doi.org/10.1007/s11192-024-05039-7

The use of ChatGPT to find similar institutions for institutional benchmarking

Lutz Bornmann[1] · **Benedetto Lepori**[2]

Received: 3 April 2024 / Accepted: 19 April 2024
© The Author(s) 2024

Abstract

In evaluative bibliometrics and higher education studies, one is frequently confronted with the task of comparing institutions with similar institutions. In this Letter to the Editor, a simple approach is discussed which applies ChatGPT. Although the approach seems to produce promising results (tested with an example at the level of research institute and of a university), it is necessary to investigate it systematically based on a sample including many institutions before it should be applied in research evaluation. The challenge in systematic investigations is that ChatGPT provides the user with different answers on the sane request (missing reliability).

Keywords Bibliometrics · ChatGPT · Benchmarking

In evaluative bibliometrics, one is frequently confronted with the task of comparing focal institutions with similar institutions (e.g., Frietsch et al., 2022). Institutional performance cannot be absolutely assessed but is meaningful only relative to other, similar institutions: by the comparison with similar institutions, one knows the standing of the focal institution. Although these comparisons are central to research evaluation, only a few attempts have been published to identify and select similar institutions based on bibliometric data. The most recent paper by Wang and Jeppsson (2022) has been published in *Scientometrics*: the authors "define an appropriate benchmark as a well-connected research environment, in which researchers investigate similar topics and publish a similar number of publications compared to a given research organization during the same period. Four essential attributes for the evaluation of benchmarks are research topics, output, connectedness, and scientific impact" (p. 7557). The approach by Wang and Jeppsson (2022) has been developed based on preliminary studies by Noyons et al. (1999), Carayol et al. (2012), and Andersen et al.

✉ Lutz Bornmann
bornmann@gv.mpg.de

Benedetto Lepori
blepori@usi.ch

[1] Science Policy and Strategy Department, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany

[2] Institute for Communication and Public Policy, Università Della Svizzera Italiana, Via Buffi 13, 6904 Lugano, Switzerland

Published online: 29 April 2024

(2017). Complementarily, higher education scholars extensively discussed how to identify and compare similar Higher Education Institutions (HEIs) using multicriteria approaches (Daraio et al., 2014), since it has been shown that comparisons of university performance are systematically affected by institutional characteristics such as subject profiles (Bornmann et al., 2013). A tool to identify similar HEIs combining multiple criteria has been developed by the U-MULTIRANK project (van Vught et al., 2012).

Although these approaches may lead to valid approaches for identifying and selecting benchmark institutions, the approaches are complex and need the access to well prepared (bibliometric and institutional) data (Lepori et al., 2022). In this Letter to the Editor, another approach is introduced which may be more simple and produce helpful results by giving hints to possible benchmark institutions. van Eck and Waltman (2024) used the GPT 3.5 Turbo LLM (OpenAI) for the task to label the research areas of a set of papers based on their titles. Since GPT produced promising results in their study, we requested ChatGPT 3.5 (the freely available variant) to provide a list of similar institutions for a focal institution. We asked ChatGPT to explain itself ("*Please explain what ChatGPT is*") and received the following answer: "*ChatGPT is a variant of the Generative Pre-trained Transformer (GPT) model developed by OpenAI, specifically tailored for conversational tasks. It's designed to engage in dialogue with users in a manner that feels natural and human-like. Like other GPT models, ChatGPT is trained on a vast amount of text data to understand language patterns and generate coherent responses*" (date of reply: March 19, 2024). The potential of ChatGPT for science studies has been the object of several studies in the informetric community (see Farhat et al., 2024; Lund & Wang, 2023; Ray, 2023; for a review) and a debate has emerged on its advantages and problems (Gordijn & ten Have, 2023; Korzynski et al., 2023; Rahman & Watanobe, 2023; Sandnes, 2024).

The goal of this letter is to open the debate on the potential of ChatGPT in identifying benchmark institutions at the level of institute, research center, and HEIs. As examples in this study, we selected as institute the Centre for Science and Technology Studies (CWTS) at Leiden University and as HEI the Leiden University. In both cases, we can assess the delivered results from ChatGPT for their appropriateness by direct knowledge of the field and of the institution. The authors work in quantitative science studies with a focus on research evaluation since around 20 years.

At the institute level, the request to ChatGPT was as follows:

> *The Centre for Science and Technology Studies at the Leiden University is active in science of science studies using bibliometric data in many research projects. In a benchmarking study, I would like to compare the research performance of the Centre for Science and Technology Studies with similar institutions. These institutions should be similar with respect to two characteristics: (1) They should be active in the same research areas in recent years. (2) They should have similar resources as the CWTS such as the number of researchers. Which institutions would you recommend for this benchmarking task? Please consider the following restrictions in your answer: (1) Only point to institutions other than the Centre for Science and Technology Studies. (2) List only the five most similar institutions. (3) The institutions should be active in 2022. (4) The institutions should be located in Europe. (5) The list should not contain programs at universities.*

ChatGPT provided the following results (March 2024):

Based on the criteria provided, here are five institutions in Europe that could be suitable for benchmarking the Centre for Science and Technology Studies at Leiden University:

1. *INGENIO (CSIC-UPV)—Institute for Innovation and Knowledge Management, Spain*
2. *Institute for Advanced Studies (IHS), Austria*
3. *Institute for Managing Sustainability (IMS), Vienna University of Economics and Business, Austria*
4. *Institute for Future Studies (IFF), Sweden*
5. *Fraunhofer Institute for Systems and Innovation Research ISI, Germany.*

Against the backdrop of our experiences in the science of science field, the provided institutions from ChatGPT seem reasonable. Depending on the focus of research in the evaluation study, the institutions are more or less suitable. We miss, however, institutions such as the German Center For Higher Education Studies (DZHW) or the Expertise Centre for Research and Development monitoring (ECOOM) in Belgium.

At the HEI level, the request to ChatGPT was as follows:

> *The University of Leiden is a medium-size research university covering most subject domains and with a very good position in international rankings. It also includes a large medical center. In a benchmarking study, I would like to compare the research performance of the University of Leiden with similar institutions. These should be similar with respect of two characteristics: (1) they should be active in similar research domains, (2) they should have similar resources at the University of Leiden such as the number of professors or the budget. Please consider the following restrictions in your answer: (1) Only point to institutions other than the University of Leiden. (2) List only the five most similar institutions. (3) The institutions should be active in 2022. (4) The institutions should be located in Europe.*

ChatGPT provided the following results (March 2024):

Certainly, considering your restrictions, here are five European universities similar to the University of Leiden in terms of research domains and resources:

1. *University of Amsterdam (Netherlands): The University of Amsterdam is also a leading research university in the Netherlands with a wide range of subject domains covered. It has notable strengths in fields such as social sciences, humanities, and natural sciences. It has a comparable size and budget to the University of Leiden.*
2. *University of Utrecht (Netherlands): Utrecht University is another prominent research university in the Netherlands with strengths across various disciplines, including social sciences, natural sciences, and humanities. It boasts a similar research profile and resources to the University of Leiden.*
3. *University of Groningen (Netherlands): The University of Groningen is known for its strong research programs across different domains, including science, engineering, social sciences, and humanities. It is comparable to the University of Leiden in terms of research output and resources.*
4. *University of Copenhagen (Denmark): The University of Copenhagen is one of the top universities in Europe and offers a broad spectrum of research areas, including natural sciences, social sciences, and humanities. It is similar to the University of Leiden in terms of research diversity and resources.*
5. *University of Edinburgh (United Kingdom): The University of Edinburgh is a prestigious research-intensive institution in the UK, with strengths across various disciplines, including medicine, sciences, social sciences, and humanities. While not geographically*

> *in mainland Europe, it matches the criteria of being similar in research domains and resources.*

These institutions are renowned for their research activities, cover a wide range of subject domains, and have resources comparable to the University of Leiden.

A cursory look and a comparison with data from the European Tertiary Education Register (Lepori et al., 2023) suggests that these are indeed good peers of the University of Leiden, with a similar position in international rankings, similar size and subject concentration, and having a medical school (like the focal university). ChatGPT was also effective in excluding highly-ranked universities which do not match some of the criteria indicated, such as technical universities or universities without medical schools. Of course, we could point to some additional peers such as University of Ghent or Gothenburg.

Our analysis of using ChatGPT has been purely illustrative, but it shows that, if a sensible querying strategy is adopted, ChatGPT might be able to suggest reasonable peers. We made the experience that one cannot expect valid results from ChatGPT in all cases. Another experience is missing reliability: The same query leads to different results at different time points (Thelwall, 2024). These experiences make clear that it is necessary to involve an expert in the search who inspect and assess the results. We strongly recommend that the institutions from ChatGPT should only be used as hints to possible benchmark institutions. ChatGPT opens several interesting questions, such as how to best design querying strategies, for which types of institutions results are (un)reliable, and whether approaches could be developed to combine data-based benchmarking tools with generative AI by exploiting the respective complementarities.

Some informetric scholars have expressed concerns on the use of ChatGPT and highlighted its limitations, arguing that it will never reach the precision of the quantitative approaches developed by our field (Gordijn & ten Have, 2023; Sandnes, 2024). While we respect this argument, it should be considered, first, that generative AI tools are developing rapidly and might provide much better results in the future; as of identifying peers, their results are already now reasonably good. The output of ChatGPT can be improved by giving (human) feedback “to produce more consistently useful or correct results. Feedback can also help it learn to avoid controversial or illegal responses” (Thelwall, 2024). Second, generative AI tools are much easier to use than data-based benchmarking tools and have attractive interactive features, where the user can easily dialogue with the tool; hence, they might turn to be a serious competitor for evaluative purposes than the tools developed by our community, particularly for students, institutional managers, and policy analysts.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Competing interest Lutz Bornmann is Editorial Board Member of *Scientometrics*. No funding was received for writing this paper. The authors declare they have no financial interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Andersen, J. P., Didegah, F., & Schneider, J. W. (2017). The necessity of comparing like with like in evaluative scientometrics: A first attempt to produce and test a generic approach to identifying relevant benchmark units. In *Proceedings of the science, technology, & innovation indicators conference "Open indicators: Innovation, participation and actor-based STI indicators"*. Paris, France.

Bornmann, L., de Moya Anegón, F., & Mutz, R. (2013). Do universities or research institutions with a specific subject profile have an advantage or a disadvantage in institutional rankings? A latent class analysis with data from the SCImago ranking. *Journal of the American Society for Information Science and Technology, 64*(11), 2310–2316.

Carayol, N., Filliatreau, G., & Lahatte, A. (2012). Reference classes: A tool for benchmarking universities' research. *Scientometrics, 93*(2), 351–371. https://doi.org/10.1007/s11192-012-0672-2

Daraio, C., Bonaccorsi, A., & Simar, L. (2014). *Rankings and university performance: A conditional multidimensional approach*. University of Rome.

Farhat, F., Silva, E. S., Hassani, H., Madsen, D. O., Sohail, S. S., Himeur, Y., & Zafar, A. (2024). The scholarly footprint of ChatGPT: A bibliometric analysis of the early outbreak phase. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2023.1270749

Frietsch, R., Gruber, S., & Rothengatter, O. (2022). *Normierter Vergleich der im Pakt für Forschung und Innovation geförderten Organisationen mit internationalen Vergleichsorganisationen anhand ausgewählter Indikatoren*. Fraunhofer ISI.

Gordijn, B., & ten Have, H. (2023). ChatGPT: Evolution or revolution? *Medicine Health Care and Philosophy, 26*(1), 1–2. https://doi.org/10.1007/s11019-023-10136-0

Korzynski, P., Mazurek, G., Altmann, A., Ejdys, J., Kazlauskaite, R., Paliszkiewicz, J., Wach, K., & Ziemba, E. (2023). Generative artificial intelligence as a new context for management theories: Analysis of ChatGPT. *Central European Management Journal, 31*(1), 3–13. https://doi.org/10.1108/Cemj-02-2023-0091

Lepori, B., Borden, V. M. H., & Coates, H. (2022). Opportunities and challenges for international institutional data comparisons. *European Journal of Higher Education, 12*, 373–390. https://doi.org/10.1080/21568235.2022.2094817

Lepori, B., Lambrechts, A. A., Wagner-Schuster, D., & Zahradnik, G. (2023). The European Tertiary Education Register, the reference dataset on european higher education institutions. *Scientific Data*. https://doi.org/10.1038/s41597-023-02353-2

Lund, B., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News*. https://doi.org/10.1108/LHTN-01-2023-0009

Noyons, E. C. M., Moed, H. F., & Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science, 50*(2), 115–131. https://doi.org/10.1002/(sici)1097-4571(1999)50:2%3c115::Aid-asi3%3e3.3.Co;2-a

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences, 13*(9), 5783. https://doi.org/10.3390/app13095783

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Sandnes, F. E. (2024). Can we identify prominent scholars using ChatGPT? *Scientometrics, 129*(1), 713–718. https://doi.org/10.1007/s11192-023-04882-4

Thelwall, M. (2024). Can ChatGPT evaluate research quality? Retrieved 3 Apr 2024, from https://arxiv.org/abs/2402.05519

van Eck, N. J., & Waltman, L. (2024). An open approach for classifying research publications. Retrieved 22 Feb 2024, from https://www.leidenmadtrics.nl/articles/an-open-approach-for-classifying-research-publications

van Vught, F. A., & Ziegele, F. (Eds.). (2012). *Multidimensional ranking: The design and development of U-Multirank*. Springer.

Wang, Q., & Jeppsson, T. (2022). Identifying benchmark units for research management and evaluation. *Scientometrics, 127*(12), 7557–7574. https://doi.org/10.1007/s11192-022-04413-7

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.