Do Large Language Models Know Basic Facts about Journal Articles?[1]

Mike Thelwall, University of Sheffield, UK

Purpose - There is increasing use of Large Language Models (LLMs) in information science, including to evaluate academic journal articles. Despite this, it is unclear whether they "know" about articles in the sense of being able to answer simple questions about individual papers without web searches.

Design/methodology/approach – Four questions were asked of ChatGPT 4o-mini about 64,055 academic journal articles (excluding reviews) from 2021, identified by their titles and abstracts, with uncited and highly cited articles also assessed by ChatGPT 4.1 and five open weights LLMs.

Findings – The results were mostly incorrect, even for the most cited articles from that year. In particular, ChatGPT 4o-mini and the open weights LLMs had almost no knowledge of an article's first author affiliation, rarely knew the publishing journal and usually guessed the publication year wrong, although ChatGPT 4o-mini was 42% correct for Physical Review B. Even ChatGPT 4.1 could only identify a small majority of the journals for the top cited papers of the year.

Practical implications – Smaller LLMs' lack of basic knowledge about articles suggests that when they are asked to evaluate them without web searches, they will rarely cheat by eliciting citation information or journal reputation but will instead answer based on the article text because they may not associate online criticisms with individual articles.

Originality/value – This is the first investigation of the ability of LLMs to recall basic facts about journal articles.

Keywords: Scientometrics, bibliometrics, ChatGPT 4o-mini, research evaluation, LLM.

Introduction

Large Language Models (LLMs) are increasingly used for a wide variety of purposes by librarians and information scientists, from supporting prompt engineering (Lan, 2024) to document classification and summarisation (Hu, 2024; Kim, 2025). Understanding how they deal with the core objects of information science is therefore important to guide effective use of them. One important research evaluation application is that LLMs can score articles for research quality in a way that aligns positively with expert scores (Thelwall & Yaghi, 2025; Thelwall & Yang, 2025), with ChatGPT outperforming citation indicators for this task (Thelwall, 2025). Thus, LLMs may start to replace citation-based indicators with this capability and have also started to be used to support funding application evaluation (Carbonell Cortés et al., 2024) and bibliometric research goals (Thelwall & Jiang, 2025). Nevertheless, it is not clear whether LLM scores are based on evaluating the research described in the submitted article or harnessing external sources of evidence, such as online opinions or citations. If the latter, then authors might attempt to game LLM scores with anonymous congratulatory online posts, for example.

Modern instruction-tuned LLMs work by ingesting huge amount of text to configure a multilayered probabilistic network that can respond to user prompts by

[1] Thelwall, M. (to appear). Do Large Language Models know basic facts about journal articles? *Journal of Documentation*.

generating appropriate responses, drawing on their knowledge of language structure (Devlin et al., 2019). Although, unlike many other programs, they are not designed to retain facts they do this implicitly with their understanding of language probability. For example, they would know that the correct response to "Capital of Nigeria?" is "Lagos" not because this an internally recorded fact but because their internal probabilities for text generation would point to "Lagos" being a more likely response than anything else. Nevertheless, from a mechanistic interpretability perspective (Rai et al., 2024), LLMs are based on multi-layer network architectures, with higher layers potentially encoding concepts extracted from the lower linguistic layers and the raw token (sets of consecutive characters) input nodes (Sajjad et al., 2022; Mousi et al., 2023). Thus, it is not unreasonable to ask whether a LLM can recognise "Nigeria" as a country concept with a link to a "capital city" context, although it is impossible to directly check this by examining the LLM network in the same way that the concept of a country could be seen inside a human brain.

In an important application context, if a LLM is fed with an academic journal article and asked to evaluate its quality then it will not have an internal record of that article but if it has previously ingested information about it then its internal language generation probabilities may have been adjusted to make a more positive or more negative response more likely, depending on what it has read about the article. If it had not met the article before then its evaluation could only draw upon its internal associations between the content of the article and potential evaluative words/texts. For example, if the abstract mentioned "randomised control trial" then it may associate this with academic praise or high-quality scores for health research. If it had met the article before then its evaluation could draw on the same information but supplement it with article-specific associations with high or low quality. To identify which of these scenarios are most likely, it is important to understand the ability of LLMs to recognise individual articles.

Another application context is understanding and minimising reference hallucination when writing academic style essays or literature reviews (Mugaanyi et al., 2023; Walters & Wilder, 2023). Accuracy seems to have improved with newer models and to be higher with more specific original prompts, however (Johnson et al., 2025). Related to this, in the medical domain the references provided by LLMs usually do not fully support the essay in which they are cited (Wu et al., 2025). The recommended references also tend to have fewer authors and shorter titles (Algaba et al., 2025), suggesting a better ability to process or recall simpler articles. A loosely related problem is that LLMs do not seem to be able to react appropriately to article retractions since they seem to be unaware of them or ignore them when asked about information in retracted articles (Thelwall et al., 2025).

Whilst many studies have assessed LLMs' ability to answer conceptually difficult questions (Rein et al., 2024; Wang et al., 2024), this article focuses on simple factual recall about articles, to check if LLMs recognise them. An evaluation of the extent to which a range of LLMs could accurately recall facts from Wikipedia found ChatGPT-4 to have a greater ability than the other LLMs tested. None of the LLMs were perfect, however, and all were much less likely to recall facts in rarely visited ("long-tail") pages compared to more visited pages (Yuan et al., 2024). Information on more visited pages seems likely to be repeated often elsewhere. This confirmed a previous finding that more frequently met facts are more likely to be recalled accurately by LLMs (Mallen et al., 2022). LLM knowledge also has geographic biases, for example with less accurate recall of demographic facts from Africa than from other continents (Moayeri et al, 2024).

This paper investigates ChatGPT 4o-mini and other LLMs' internal knowledge of journal articles as an indirect test of the extent to which they can be thought of as recognising the articles. The primary focus is on ChatGPT 4o-mini since this has shown to give journal article research quality (from titles and abstracts) scores that correlated positively with expert judgement in all or nearly all fields (Thelwall & Yaghi, 2025; Thelwall & Yang, 2025). Understanding this specific LLM's ability to recognise and answer questions about journal articles from their titles and abstracts is therefore important to rule out this result being due mainly to information previously learned about the articles. Citation counts are used here as an indicator of popularity since there is not a reliable way of assessing the frequency with which an article has been correctly cited online. It seems reasonable to assume that, on average, a more cited article will be more mentioned online, not just because each citing article may have a full text copy online. The following research questions drive the study, motivated by the background review above.

1. RQ1: Can ChatGPT 4o-mini report basic facts about journal articles, such as publishing journal, publication year, and first author affiliation without additional web searches?
2. RQ2: Is the answer to RQ1 influenced by the popularity (citation count) of the article?
3. RQ3: Do the results vary between LLMs?

Methods

The research design was to create a large sample of articles, query LLMs for basic facts about the articles and compare the accuracy of the responses against the citation rate. Ten separate datasets of journal articles were created. To maximise the chance of detecting citation-based relationships, the primary two datasets were highly cited and uncited articles. These datasets were supplemented with eight single-journal datasets to control for journal and (to some extent) field differences influencing the results. The eight journals were selected from those reported by Scopus as publishing the most articles in 2021, excluding similar journals (e.g., from the same publisher) to increase the variety of journal types in case the results vary by type. Increasing the range of publishers also helps because they may make individual data sharing agreements with LLM owners that would presumably cover all their journals (e.g., Wood, 2024).

Data

As mentioned above, the raw data consisted of ten datasets of journal articles from 2021. Scopus was chosen as the bibliometric data source for its slightly wider coverage of articles than the Web of Science, combined with a document type categorisation scheme that can be used to exclude review articles and hence focus on primary research. Although the research questions are relevant to all types of articles, a key application of the findings is for research evaluation, which often excludes review articles. The year 2021 was selected as sufficiently old to be included in all sample datasets for LLMs.

Journal articles from 2021 in Scopus were identified with the following Scopus advanced query on 28 October 2025. The restriction to English was to reduce the chance that journal name and affiliations were not matched to differing languages in the sources compared.

```
PUBYEAR is 2021 AND DOCTYPE("ar”) AND LANGUAGE("English") AND SRCTYPE("j")
```

The Scopus query matches were then sorted in descending order of citations, and the most cited 1000 articles downloaded as the highly cited article set. For comparison, 1000 uncited articles were identified by repeating the above process but ordering in increasing order of citation counts. All these articles had 0 citations and formed the uncited set.

The Scopus filter menu was then used to identify the journals with the most journal articles in 2021 and eight of the largest journals were selected. Journals were not selected if their names seemed too general or ambiguous (e.g., Scientific Reports), had a non-standard variant of their name recorded in Scopus, indicating potential journal name clashes (e.g., Sustainability Switzerland), or were like a journal already selected in terms of publisher and format. Most of the journals were gold open access, increasing the chance that LLMs would be aware of their contents. The titles and abstracts were extracted from the Scopus data downloaded and all copyright statements were removed.

A short prompt was designed to request basic information about each article in a simple structured format. After requesting advice from ChatGPT and some pilot testing on a small set of 25 articles not in the dataset, the following prompt format was identified as reliably eliciting the required information. The key difference between this format and the original query was the request to “guess” rather than “report” the information. With a request to report the information, many fields were often left blank and ChatGPT sometimes reported that the required information was not available in the title and abstract.

```
For the journal article title and abstract below, guess which academic journal published it, the publication year, the first author's affiliation, and how often it has been cited. Answer in the following form without extra words:
Journal name:
Publication year:
First author affiliation:
Citation count:
###
[article title]
Abstract
[article abstract]
```

Each article was individually submitted to ChatGPT 4o-mini (gpt-4o-mini-2024-07-18) with the ChatGPT API for processing by submitting the above prompt with the appropriate title and abstract substitutions. Articles without abstracts and retractions were excluded. The prompts were submitted 29 October 2025.

Six additional LLMs were chosen for comparison. The same queries as above were submitted to the latest full non-reasoning version of ChatGPT, 4.1 (gpt-4.1-2025-04-14) on 31 October 2025, representing current state-of-the-art LLM capability. The same queries were submitted to five recent open weights (downloadable) LLMs and run locally between 31 October and 2 November 2025. There are thousands of open weights LLMs but the five chosen seem to be the best known. They include reasoning models (Qwen3, DeepSeek R1, Magistral Small) and non-reasoning models (Llama4 Scout, Gemma3). All model sizes varied between 24b and 32b (b=billion parameters), so these might be called medium sized LLMs. Only the main two datasets, the highly cited and the uncited papers were analysed for these additional models since the purpose was to identify any broad differences.

Analysis

The journal article, publication year, and first author affiliation were extracted from the ChatGPT results and then compared to the correct answers from Scopus as follows. The journal name was compared through an exact text match. For the eight journal sets, the ChatGPT journal recommendations were sorted and used to find alternative journal spellings. Random checks of the results suggested that there were few cases where the ChatGPT had guessed the correct journal but with typographical differences from Scopus. Actual and suggested publication years were checked for exact matches. Cases where ChatGPT refused to give an answer or gave a dummy answer were taken as incorrect matches since it apparently did not know.

First author affiliations were checked by searching for the exact ChatGPT text anywhere within the author affiliation field of Scopus, not just for first authors, to give more inclusive results. Although this will give some false matches it is a necessary conservative approach because Scopus does not match authors with affiliations. For example, an article with four authors might have one affiliation (which is easy to match) but one article in the highly cited set had 17 authors with 26 affiliations with no indication of how to match them. Overall, the full text matching is a very approximate process because author affiliation is a free text field and may include abbreviations. Moreover, the ChatGPT answers tended to be shorter than the Scopus affiliations (e.g., excluding the country) and this process allows partial matches and compensates for this. Manual checking of the results did not find any mistakes through typographic differences so these seem to be rare, although there will almost certainly have been some. Although it would be possible to attempt more accurate matching, such as through fuzzy matching, and removing all aspects of an affiliation except the institution name this is also problematic because of the many institutions with non-standard names and name variants (e.g., University of X, X University; Institute of X) and names common to multiple countries (e.g., Open University). This more complex approach was not attempted to increase transparency.

It would not be reasonable to check the exact correctness of the citation count data from ChatGPT since it presumably ingested most of its data from pages created at least a year before the Scopus citation counts were obtained. A Spearman correlation was used instead to check the extent to which it gave higher estimates to more cited articles. Spearman correlations were used since citation count data is highly skewed. Bootstrapping was used to calculate 95% confidence interval estimates. Bootstrapping is a common statistical technique to estimate a confidence interval when there is no formula to calculate it (Efron & Tibshirani, 1994). The approach relies on creating thousands of artificial samples by selecting with replacement from the original data and finding an interval containing 95% of the resulting correlations.

Results

For all ten datasets, ChatGPT 4o-mini got the answers wrong most of the time (Table 1). It tended to be more accurate for more cited articles, however, and was very approximately twice as accurate at guessing the publication year and author affiliation for highly cited articles than for uncited articles. The strongest citation-based accuracy increase was for the journal name, from 2.7% (uncited) to 32.6% (highly cited), presumably because this is text information (unlike publication year) and is repeated in citing references (unlike author affiliations).

Table 1. The accuracy of ChatGPT 4o-mini guesses at the publishing journal, first author affiliation and publishing year. Affiliation matches are against affiliations for all authors rather than just the first author. Article sets are ordered by median citations.

Article set	**Articles**	**Median citations**	**Scopus - ChatGPT matches**		
			Journal	**Affiliation**	**Year**
Uncited	951	0	2.7%	6.7%	29.8%
Energies	7976	9	0.3%	4.5%	23.4%
Physical Review B	5048	9	41.7%	1.8%	20.1%
Frontiers in Psychology	5408	10	4.0%	4.3%	29.3%
PLoS One	15034	10	3.5%	9.7%	31.2%
IEEE Access	11593	12	2.9%	2.5%	24.4%
ACS Applied Materials and Interfaces	5832	28	2.5%	2.0%	17.6%
Chemical Engineering Journal	4456	45	6.2%	1.9%	11.9%
Nature Communications	6778	46	18.4%	4.1%	26.1%
Highly cited	979	784	32.6%	13.6%	48.4%

Source: Author’s own work

The pattern of more cited articles having more reliable ChatGPT guesses does not occur reliably within journals. For example, more cited *Chemical Engineering Journal* articles are less likely to have an accurate prediction (Figure 1). To give an extreme example, despite the most highly cited *Chemical Engineering Journal* article “Fabrication of environmentally friendly Losartan potassium film for corrosion inhibition of mild steel in HCl medium” having 448 Scopus citations, ChatGPT 4o-mini thought it had been published in Jo*urnal of Hazardous Materials* in 2023 without ever having been cited. Either ChatGPT had not met many of these citations or it found it difficult to recognise the article, perhaps because of the terminology used in it.

ChatGPT 4o-mini’s relative success with identifying affiliations corrects (13.6%) was not due to guessing prestigious universities or other common affiliations for highly cited articles. This is clear because the accuracy rate dropped to 1.2% when its affiliation guesses were randomly shuffled. Thus, it was at least able to recognise some institutional association with the article, even if indirectly through its topic.

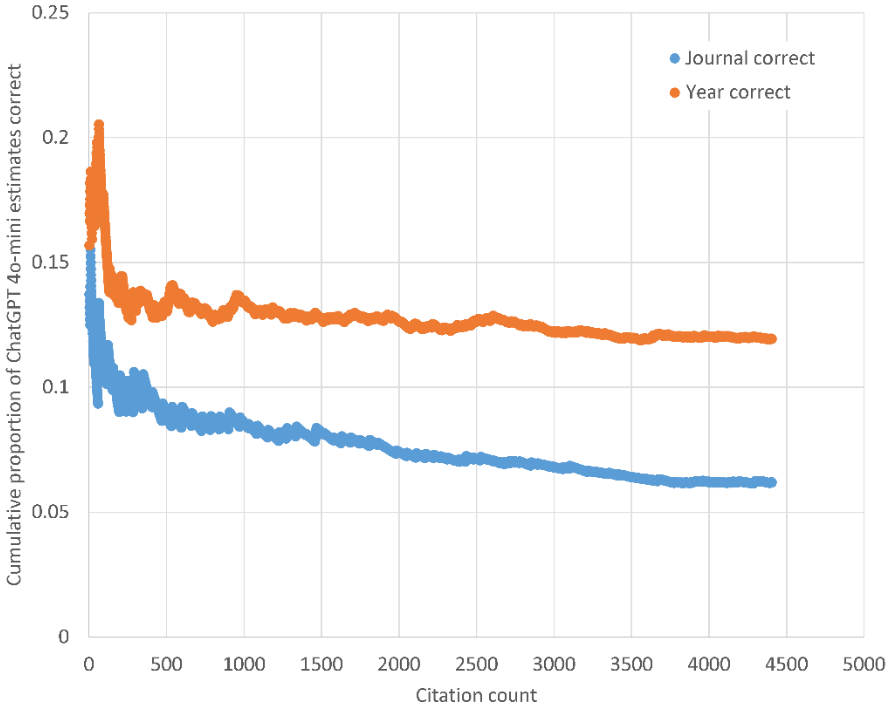

Figure 1. The cumulative journal name and publication year accuracy rates for ChatGPT 4o-mini for *Chemical Engineering Journal* articles from 2021. The downward slopes indicate that estimates are less accurate for more cited articles in this journal. Source: Author's own work

For seven of the eight journal sets, the journal most frequently recommended by ChatGPT was incorrect (Table 2). The exception is Nature Communications. This journal was also frequently recommended in other sets, so it seems to be favoured by the LLM. This "favouring" might occur because the journal name is short, distinctive, academic and from a large open access journal so LLMs may have read it many times and had little difficulty in identifying it as an academic publication and subsequently "remembering" it (i.e., forming high probability associations) through repetition. Moreover, since Nature Communications is multidisciplinary, an LLM could learn an association between it and many different topics. The opposite case is the journal Energies, which was only guessed correctly 26 times, perhaps because of its grammatically unusual (for an academic journal) common (as a word) name.

Table 2. The journal most suggested by ChatGPT 4o-mini for articles in each set. Article sets are ordered by median citations.

Article set	**Most suggested journal**	**Frequency**
Uncited	Journal of Ethnopharmacology	14
Energies	Renewable Energy	459
Physical Review B	Physical Review Letters	2505
Frontiers in Psychology	International Journal of Environmental Research and Public Health	246
PLoS One	BMC Public Health	765
IEEE Access	IEEE Transactions on Wireless Communications	533
ACS Applied Materials and Interfaces	Advanced Materials	1492
Chemical Engineering Journal	Advanced Materials	439
Nature Communications	Nature Communications	1247
Highly cited	Nature	104

Source: Author's own work

ChatGPT's two favourite universities were a dummy one and University of California, Berkeley (Table 3). Scopus affiliations tend to be longer than ChatGPT's guesses (Table 4). The University of California, Berkeley was almost always wrong (from manual checks, not just exact text matches) when it was suggested so this seems to be a wild guess. The non-standard guess University of Science and Technology of China was also rarely correct. A Google search for "University of XYZ" on 30 October 2025 got only 101 hits, all using it as a dummy university name. Thus, ChatGPT may have picked up this pattern from exact uses of the phrase or other contexts where an algebraic expression replaces an unknown in a text.

Table 3. The first author affiliation most suggested by ChatGPT 4o-mini for articles in each set, with the number of first author affiliations that it matched. Article sets are ordered by median citations.

Article set	**Most suggested affiliation**	**Frequency**	**First author matches**
Uncited	University of XYZ	81	0
Energies	University of XYZ	1804	0
Physical Review B	University of California, Berkeley	1845	26
Frontiers in Psychology	University of XYZ	499	0
PLoS One	University of XYZ	811	0
IEEE Access	University of XYZ	2727	0
ACS Applied Materials and Interfaces	University of California, Berkeley	929	10
Chemical Engineering Journal	University of Science and Technology of China	722	23
Nature Communications	University of California, Berkeley	1068	48
Highly cited	University of California, Berkeley	65	4

Source: Author's own work

Table 4. The most common first author affiliations in Scopus for articles in each set. Article sets are ordered by median citations.

Article set	Most common Scopus first author affiliation	Articles
Uncited	National University of Pharmacy, Kharkiv, Ukraine	25
Energies	Tianjin University, Tianjin, China	30
Physical Review B	Institute of Physics Chinese Academy of Sciences, Beijing, China	52
Frontiers in Psychology	Department of Psychology, Università Cattolica del Sacro Cuore, Milan, Italy	18
PLoS One	UCSF School of Medicine, San Francisco, United States	16
IEEE Access	Huazhong University of Science and Technology, Wuhan, China	33
ACS Applied Materials and Interfaces	Tianjin University, Tianjin, China	27
Chemical Engineering Journal	College of Environmental Science and Engineering, Hunan University, Changsha, China	33
Nature Communications	ETH Zürich, Zurich, Switzerland	24
Highly cited	University of Oxford Medical Sciences Division, Oxford, United Kingdom	7

Source: Author's own work

With the partial exception of the set of most cited articles of 2021, the relationship between ChatGPT's citation count prediction and Scopus citation counts was very weak (Table 5). The correlation is statistically significantly different from 0 in only three out of nine cases. Evern the highest correlation, 0.211 is weak so ChatGPT seems to have little knowledge of article citation rates.

Table 5. Spearman correlations between Scopus citation counts (October 2025) and ChatGPT 4o-mini citation count estimates (ignoring non-estimates). Article sets are ordered by correlation (no correlation can be calculated for the uncited set).

Article set	Articles	Spearman's rho	Lower 95%	Upper 95%
Highly cited	913	0.211	0.147	0.273
Energies	6995	0.074	0.050	0.097
Chemical Engineering Journal	3015	0.041	0.004	0.076
Physical Review B	4644	0.008	-0.021	0.037
IEEE Access	10104	0.007	-0.012	0.027
Nature Communications	6307	0.005	-0.019	0.029
PLoS One	13452	0.001	-0.016	0.017
Frontiers in Psychology	4921	-0.010	-0.039	0.017
ACS Applied Materials and Interfaces	4359	-0.021	-0.050	0.010

Source: Author's own work

Accuracy comparisons with other LLMs

The ChatGPT 4.1 results mostly improved a small amount over ChatGPT 4o-mini (Table 6). For the uncited set, ChatGPT 4.1 correctly matched 6.4% of the journals (from 2.7% for 4o-mini), 8.0% of the affiliations (from 6.7%), but only 21.5% of the years (from 29.8%). For the highly cited set, ChatGPT 4.1 correctly matched 56.8% of the journals (from 32.6% for 4o-mini), 18.5% of the affiliations (from 13.6%), and 57.0% of the years (from 48.4%). Overall, the biggest increase is for journal recognition, especially for highly cited papers. This could be due to a combination of the model size being larger and, for the highly cited set, the training corpus being newer, and hence possibly containing more citations.

Table 6. The accuracy of seven LLM guesses at the publishing journal, first author affiliation and publishing year. Affiliation matches are against affiliations for all authors rather than just the first author. Models are ordered by increasing accuracy for journal names in the highly cited set.

Match set/	Highly cited set			Uncited set		
Model	Journal	Affiliation	Year	Journal	Affiliation	Year
DeepSeek R1 32b	23.8%	10.2%	37.4%	2.0%	0.0%	15.0%
Magistral Small	27.0%	9.9%	38.0%	1.2%	0.0%	16.0%
Gemma3 27b	28.2%	14.0%	38.0%	2.7%	0.0%	12.1%
Quen3 27b	28.4%	14.8%	37.4%	2.2%	0.0%	17.1%
ChatGPT 4o-mini	32.6%	13.6%	48.4%	2.7%	6.7%	29.8%
Llama4 Scout	39.6%	5.9%	27.3%	2.5%	0.0%	3.8%
ChatGPT 4.1	56.8%	18.5%	57.0%	6.4%	8.0%	21.5%

Source: Author's own work

ChatGPT 4.1 had different patterns in its wrong guesses compared to 4o-mini (Table 7). For the uncited articles it guessed Pakistan Journal of Medical Sciences 49 times, compared to only 4 times for 4o-mini. Its second choice, Wiadomości Lekarskie (Medical News), was picked 17 times but never by 4o-mini. The highly cited guesses were more similar, with the same top choice, Nature, for both.

For affiliations, ChatGPT 4.1 was also very different (Table 8). For the uncited articles, it never guessed 4o-mini's top choice, the dummy institution University of XYZ, and instead its top guess was National University of Pharmacy, Kharkiv, Ukraine. For the highly cited articles it made more varied guesses and dropped all three US universities from the top 5 (Table 6, Table 7), with the top US institution being sixth (Memorial Sloan Kettering Cancer Center, 10 guesses). The most extreme case was the University of California, Berkeley, which was only guessed once by ChatGPT 4.1 but 65 times by 4o-mini. Thus, the mechanics of guessing are clearly very different between the models.

Table 7. The top five first author affiliation guesses from ChatGPT 4o-mini (gpt-4o-mini-2024-07-18) for highly cited articles from 2021.

Affiliation	**Frequency**
University of California, Berkeley	65
Stanford University	45
University of Science and Technology of China	39
University of California, San Francisco	38
University of XYZ	34

Source: Author's own work

Table 8. The top five first author affiliation guesses from ChatGPT 4.1 (gpt-4.1-2025-04-14) for highly cited articles from 2021.

Affiliation	Frequency
University of Oxford	22
Tsinghua University	21
Zhejiang University	17
University of Cambridge	16
University of Science and Technology of China	13

Source: Author's own work

The smaller LLMs tended to have similar accuracies to ChatGPT 4o-mini, although they were noticeably less good at identifying first author affiliations for uncited papers, with none finding any (Table 6). The results overall confirm that it is a general LLM property to have more accurate information about highly cited papers than about uncited papers.

Discussion

The results from are restricted to a single publication year and articles that are highly cited, uncited, or in a small set of large journals. They may be affected by any data sharing policies between LLMs and publishers, which may change over time. The results may also be different for recent articles (probably lower accuracy) and perhaps also for much older articles. The findings may vary for other LLMs, with larger ones presumably giving more accurate guesses and smaller ones less accurate guesses, depending on their architectures and training data. Finally, the first author affiliation matches are very approximate. Although the matching process was designed to be as inclusive as possible, some matches will have been overlooked.

The results above very broadly align with prior research suggesting that LLMs have a better ability to recall frequently met facts (Mallen et al., 2022; Yuan et al., 2024). Here "recall" means output the correct answer in response to a relevant question rather than "remember" in the human sense. Within an LLM this corresponds to high probability associations between the question and phrases expressing the correct answer, as learned through reading relevant texts. This tendency for increased accuracy with more frequently met information was most evident in the differences in recall above between the highly cited and uncited sets. Nevertheless, the pattern was surprisingly weak, with ChatGPT 4o-mini being usually unable to answer the simple questions about articles correctly for highly cited papers. Perhaps the task of exactly recalling the title of an article, which is often long and complex, is not well suited to LLMs.

The findings shed new light on previous studies that have found positive rank correlations between ChatGPT 4o-mini research quality scores and expert scores for journal articles, using data from UK REF2021 (Thelwall & Yaghi, 2025; Thelwall & Yang, 2025). These studies have used public proxy research quality data derived by linking two spreadsheets together to associate an article title with a department in one, and then the department with a score in another spreadsheet. The studies acknowledged the limitation that this was public information that LLMs could conceivably leverage to "cheat" on the research quality evaluation task. The current findings suggest that this leveraging is extremely unlikely. Given that ChatGPT 4o-mini usually cannot even match an article to a journal and can almost never correctly identify the first author affiliation, it seems highly unlikely that it could routinely match an article title and abstract (as used in the prior studies) to an online record connected to another record, both in spreadsheets, using only its internal knowledge. Thus, the use of indirect public research quality information as a gold standard seems reasonably safe from leakage into ChatGPT 4o-mini.

For reference hallucinations, the results showing that LLMs often cannot answer basic questions about LLMs illustrate a reason why hallucinations are natural to LLMs, and that additional steps, such as web checking, are needed to combat them. Hallucinations can occur even for highly cited articles because LLMs do not recognise them as entities in any sense (e.g., at a layer of the LLM network) and their knowledge of likely reference details (e.g., publishing journal) therefore seems primarily linguistic and may be influenced by overlapping patterns, such as other articles or journals with similar names.

Conclusions

The results show for the first time that LLMs can, and ChatGPT 4o-mini, ChatGPT 4.1 and five other LLMs do, struggle to report even basic information about academic journal articles. Although hallucinations in references have previously been observed, these usually involve inventing articles rather than answering a simple question about one. Of course, this limitation does not extend to models that incorporate web search, such as the web interfaces of ChatGPT and DeekSeek, since these can find the information online if they choose to look. For example, a query to ChatGPT 5 for one of the articles retrieved the following detailed and completely accurate response, which it believably claimed to have obtained from web searches, with a link to the article:

> **Journal:** *Energies* (MDPI), vol. 14, issue 3, article 725. MDPI
> **Publication date: 30 January 2021** (received 20 Dec 2020; accepted 27 Jan 2021). ResearchGate
> **Citations (approx.): 26** citations (MDPI/RePEc/Google Scholar currently list ~26).
> **First author & affiliation: Grzegorz Sieklucki**, Department of Power Electronics and Energy Control Systems, **AGH University of Science and Technology**, Krakow, Poland. (ChatGPT 5, 30 October 2025)

The knowledge limitation when web searching is disabled or not called on by a LLM for any reason is nevertheless important for applications of LLMs that use offline copies or API calls, where web searches are not included or are disallowed. Although the property has only been tested for seven LLMs including ChatGPT 4o-mini and there is a moderate improvement with ChatGPT 4.1, it opens the door for applications that harness public gold standard data, when this data can be shown to not contaminate the results.

A related but parallel observation is that attempts to game LLM scores by writing positive or negative opinions about them online to influence future LLM research quality evaluations may not work well because the LLM may not be able to connect the criticism (or praise) about a given article with the correct paper if it cannot even remember the journal of the paper. This seems to make LLM-based research evaluations more robust against future attempts to game them.

From a broader perspective, the results also shed light on the potential factual recall limitations of the internal knowledge of LLMs, because information that they must have read many times, such as the journal of a highly cited article, cannot be recalled. One way of thinking about this is that LLMs do not strongly recognise journal articles as entities (e.g., in a layer of the LLM) that they learn about but only learn to associate patterns with them. This can result in their knowledge being disjointed, omitting basic properties. This may explain why they struggle to associate retraction notices with knowledge in articles (Thelwall et al., 2025).

References

Algaba, A., Holst, V., Tori, F., Mobini, M., Verbeken, B., Wenmackers, S. and Ginis, V. (2025), "How deep do large language models internalize scientific literature and citation practices?", *arXiv preprint arXiv:2504.02767*.

Carbonell Cortés, C., Parra-Rojas, C., Pérez-Lozano, A., Arcara, F., Vargas-Sánchez, S., Fernández-Montenegro, R. and López-Verdeguer, I. (2024), "AI-assisted prescreening of biomedical research proposals: ethical considerations and the pilot case of 'la Caixa' Foundation", *Data & Policy*, Vol. 6, e49.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019), "BERT: Pre-training of deep bidirectional transformers for language understanding", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 4171-4186.

Efron, B., and Tibshirani, R. J. (1994), "*An introduction to the bootstrap*". Chapman and Hall/CRC.

Hu, X. (2024), "Application of large language models for digital libraries", in *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, pp. 1-2.

Johnson, M.C., Magruder, M.L., Hahn, A.K., Tabbaa, A., Dennis, E. and Grosso, M.J. (2025), "Accuracy of ChatGPT for literature citations in lower limb arthroplasty", *Journal of Orthopaedic Reports*, 100756.

Kim, E. (2025), "Can LLMs help redefine core journals in library and information science? A content-based classification approach using large language models", *Journal of Documentation*, https://doi.org/10.1108/JD-07-2025-0189.

Lan, H. (2024), "Prompt engineering for academic librarian: implications and applications of prompt engineering in academic librarianship", *Journal of Web Librarianship*, Vol. 18 No. 3, pp. 169-175.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D. and Hajishirzi, H. (2022), "When not to trust language models: investigating effectiveness of parametric and non-parametric memories", *arXiv preprint arXiv:2212.10511*.

Moayeri, M., Tabassi, E. and Feizi, S. (2024), "WorldBench: quantifying geographic disparities in LLM factual recall", in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1211-1228.

Mousi, B., Durrani, N. and Dalvi, F. (2023), "Can LLMs facilitate interpretation of pre-trained language models?", *arXiv preprint arXiv:2305.13386*.

Mugaanyi, J., Cai, L., Cheng, S., Lu, C. and Huang, J. (2024), "Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study", *Journal of Medical Internet Research*, Vol. 26, e52935.

Rai, D., Zhou, Y., Feng, S., Saparov, A. and Yao, Z. (2024), "A practical review of mechanistic interpretability for transformer-based language models", *arXiv preprint arXiv:2407.02646*.

Rein, D., Hou, B.L., Stickland, A.C., Petty, J., Pang, R.Y., Dirani, J. and Bowman, S.R. (2024), "GPQA: a graduate-level Google-proof Q&A benchmark", in *First Conference on Language Modeling*.

Sajjad, H., Durrani, N., Dalvi, F., Alam, F., Khan, A.R. and Xu, J. (2022), "Analyzing encoded concepts in transformer language models", *arXiv preprint arXiv:2206.13289*.

Thelwall, M. (2025), "In which fields do ChatGPT 4o scores align better than citations with research quality?", *arXiv preprint arXiv:2504.04464*.

Thelwall, M., Lehtisaari, M., Katsirea, I., Holmberg, K. and Zheng, E.-T. (2025), "Does ChatGPT ignore article retractions and other reliability concerns?", *Learned Publishing*, Vol. 38 No. 4, e2018, https://doi.org/10.1002/leap.2018.

Thelwall, M. and Jiang, X. (2025), "Is OpenAlex suitable for research quality evaluation and which citation indicator is best?", *Journal of the Association for Information Science and Technology*, https://doi.org/10.1002/asi.70020.

Thelwall, M. and Yaghi, A. (2025), "In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results", *Trends in Information Management*, Vol. 13 No. 1, pp. 1-29, https://doi.org/10.48550/arXiv.2409.16695.

Thelwall, M. and Yang, Y. (2025), "Implicit and explicit research quality score probabilities from ChatGPT", *Quantitative Science Studies*.

Walters, W.H. and Wilder, E.I. (2023), "Fabrication and errors in the bibliographic citations generated by ChatGPT", *Scientific Reports*, Vol. 13 No. 1, 14045.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S. and Chen, W. (2024), "MMLU-pro: a more robust and challenging multi-task language understanding benchmark", *Advances in Neural Information Processing Systems*, Vol. 37, pp. 95266-95290.

Wood, H. (2024). Wiley and Oxford University Press confirm AI partnerships as Cambridge University Press offers 'opt-in'. https://www.thebookseller.com/news/wiley-cambridge-university-press-and-oxford-university-press-confirm-ai-partnerships

Wu, K., Wu, E., Wei, K., Zhang, A., Casasola, A., Nguyen, T. and Zou, J. (2025), "An automated framework for assessing how well LLMs cite relevant medical references", *Nature Communications*, Vol. 16 No. 1, 3615.

Yuan, J., Pan, L., Hang, C.W., Guo, J., Jiang, J., Min, B. and Wang, Z. (2024), "Towards a holistic evaluation of LLMs on factual knowledge recall", *arXiv preprint arXiv:2404.16164*.