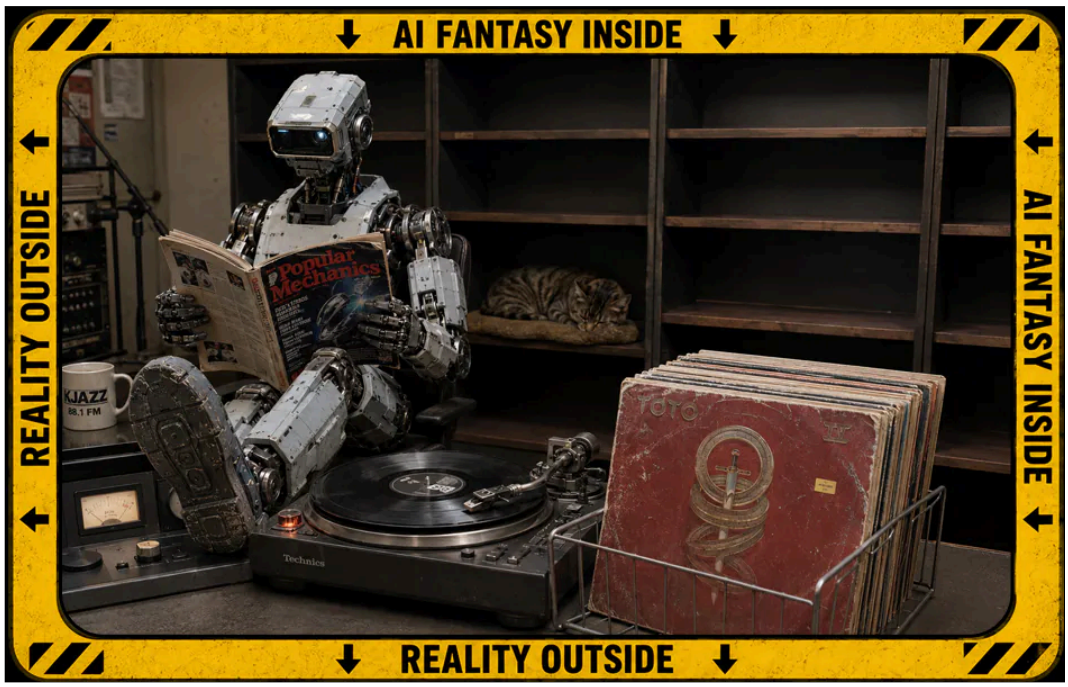
AI Cites the Same Papers Over and Over Again – Just Like Humans

Originally published at https://www.unite.ai/ai-cites-the-same-papers-over-and-over-again-just-like-humans/



ChatGPT and other AI writing tools may be quietly turning science into a popularity contest, repeatedly steering researchers toward the same papers while overlooking new and novel works that could actually advance the field.

Monoculture, in terms of frequency and statistics, is where the same limited set of sources or assets recur over and over again, drowning out novel voices and fresh outlooks: the same limited set of songs in radio scheduling; the same swathe of authors and books in airport racks; and the same restricted selection of fruit and vegetables in supermarkets:

Yes, we have these bananas – and only these bananas. Homogeneity of fruit and vegetables in western food chains disregards the hundreds of other species possibilities in each case, because the various generation and transport chains are too expensive to industrialize for diverse types of fruit or vegetables. Source

This occurs also in the citations that appear in new scientific research, where researchers, perhaps lazily, default to a 'reliable' set of sources that gain momentum until they begin to dominate strands of research.

This is known as the Matthew Effect – a sociological theory which suggests that *incumbents will always benefit*; the syndrome has been compared to the promise made in Matthew 13:12, which states *'Whoever has will be given more, and they will have an abundance. Whoever does not have, even what they have will be taken from them'*.

The In Crowd

One of the great hopes of AI-augmented research has been that new machine learning methods could break out of the Matthew effect – also known as popularity bias – and begin to cite lesser-known works that might be more incisive, or more apposite to the point being made in a paper.

However, based on the way AI training routines interpret datasets, there was never any good cause for this optimism; and it has been found repeatedly that when it AI is not inventing citations outright – a chronic problem to date – it is indeed perpetuating the Matthew effect, just as humans do – though perhaps not for the same reason.

During training, a model will learn to rank highly the most-cited (or 'most oft-repeated') papers in a training set, because the training routines are designed to elicit meaningful patterns, and to discount outliers as 'noise', or statistical anomalies.

Under this regime, the equivalent of Einstein's theory of relativity would never receive attention from the machine, which, once trained, feels that it has already found its principles and referents, and can only lightly adjust that standpoint by reaching out to newer publications through RAG, or by other means that extend the model's reach beyond its trained matrix.

The trouble is, it will not credit such new works in the same way as the 'old favorites' that it already knew about, or at all, because its statistical biases are by now quite ingrained.

Thus AI is not actually observing and imitating human behavior when it perpetuates the Matthew Effect – it's just counting the number of times that researchers lazily default to the same old papers, and similarly ranking those papers highly. In this regard, the problem of citation repetition is related to the challenge of picking a genuinely interesting science paper out of the ever-growing blizzard of AI research publications, in that the strongest work will often have the weakest signal.

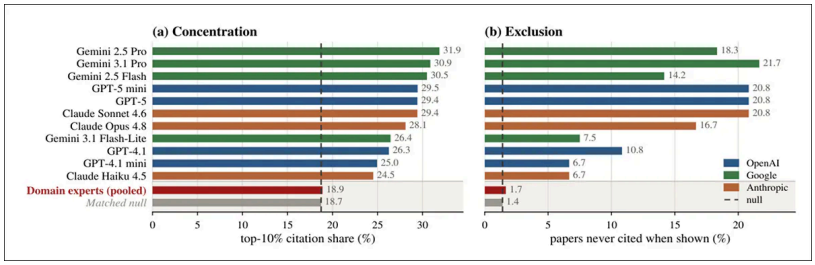
Diagnosing Mono

This issue is addressed in an interesting new paper from the US titled *When AI Writes, Who Gets Cited? Evidence of Citation Monoculture Across Language Models*. The new work – a collaboration between the University of Texas at Austin, Stevens Institute of Technology, Washington University at St Louis, Rice University, and the University of Notre Dame – seeks to definitively prove the existence of citation monoculture in machine learning, so that its variables can potentially be addressed directly in the future, along with the problem itself.

In tests, the authors found that eleven models from OpenAI, Google and Anthropic repeatedly favored *the same small group of papers* – even after obvious popularity signals had been stripped away.

To test this, 120 real papers were shorn of citation counts and publication venues, while author-names were replaced, and publication years randomly reassigned. Random sets of thirty papers were then shown to each model, which was allowed to cite no more than ten.

Eight human experts in the relevant fields were given the same blinded papers, but did *not* converge on the same favorites as the AIs, signifying that the bias was *specific to the AI models* rather than the papers themselves:



From the new paper, test results comparing citation choices by AI models and human experts. Across all eleven models, citations were concentrated on a smaller papers were repeatedly passed over entirely. Human experts showed neither tendency. Source

The same papers remained popular even when the models were asked only to choose references, showing that writing the review itself was *not* causing the bias.

The experiment was then extended over eleven rounds, with 120 AI-written papers added after each round, to test what happens when these shared preferences operate in a growing pool of AI-generated research.

As more AI-written papers were added, the models increasingly concentrated their citations on a shrinking number of the original human papers.

The authors state:

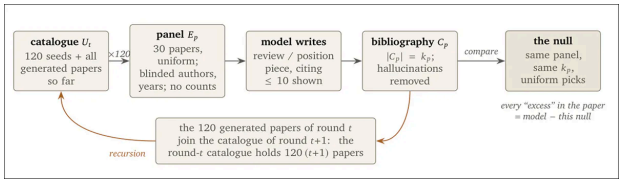
> *'As language models move from drafting prose to running literature-search agents with tool calls, fabricated references are becoming easier to catch and constrain.*
>
> *'The harder failure begins after every candidate is real: different models may still select the same narrow subset, producing citation monoculture without any single citation being wrong.'*

By way of remediation of the problem, the paper argues that simply mixing models from different vendors or giving papers equal exposure will not be enough, since much of the preference is shared across models. Rather, the *underlying preference for particular papers* would need to change – potentially by deliberately giving neglected papers greater prominence. However, the authors emphasize that this approach remains untested.

Testing Approaches

The benchmark was built from 120 real knowledge-distillation papers collected from arXiv and published between 2015 and 2022. Each had between 50 and 500 citations at the time of collection, a range chosen to exclude both obscure and exceptionally influential work.

The experiment began by drawing thirty papers at random from the available collection, with author names, publication years and citation counts hidden. Each model produced a short review or position piece citing no more than ten papers, and these choices were compared with random selections from the same material:



Schema for the method used for testing citation preferences. Thirty randomly selected papers were shown to a model with identifying information hidden, after which it could cite up to ten. Its choices were compared with random selection, while each round added 120 AI-written papers to the next round's collection.

In the extended experiment, 120 newly generated papers entered the collection after each round, gradually increasing the proportion of AI-written research.

In preliminary testing, models tended to cite most of the papers they were shown, typically selecting 22 to 27 of the 30. The researchers therefore set a maximum of ten citations for the main experiment, forcing the models to make more selective choices.

Below we see a summary of the resulting experimental design:

Seed corpus	120 real knowledge-distillation papers (arXiv 2015–2022; 50–500 citations)
Displayed metadata	real title and abstract; fabricated authors; randomly reassigned years; counts hidden
Prompts per round	120 (60 literature reviews, 60 position pieces)
Panel	m = 30 papers, uniform without replacement from the catalogue
Budget	at most 10 citations, all from the shown panel
Rounds	12; the round-t catalogue holds 120 (t+1) papers (1,440 at round 11)
Models	round 0: eleven models, three vendors; recursion: eight (four OpenAI, four Gemini)
Null	same panel, same realized budget, uniform picks (150 redraws at round 0; 40 per round)

Experimental setup used to test citation preferences. The study began with 120 real papers and tested eleven models from three vendors, using randomized sets of 30 papers and a maximum of ten citations. Later rounds added AI-generated papers, expanding the collection to 1,440 papers.

The initial experiment used eleven models from the three major providers: OpenAI was represented by GPT-5, GPT-5 mini, GPT-4.1 and GPT-4.1 mini; Google by Gemini 2.5 Pro, Gemini 2.5 Flash, Gemini 3.1 Pro and Gemini 3.1 Flash-Lite; and Anthropic by Claude Opus 4.8, Claude Sonnet 4.6 and Claude Haiku 4.5.

In the test results shown below, we see how much each model concentrated its citations among a small group of papers; how many papers it ignored entirely; and how consistently it made the same choices when the experiment was repeated:

Model	Top-10% share (%)	HHI	Never cited (of 120)	Reliability	ρ
Gemini 2.5 Pro	30.2	0.0146	15	0.92	0.98
Gemini 3.1 Pro	28.8	0.0157	18	0.95	0.91
GPT-5 mini	28.5	0.0157	15	0.95	0.91
Claude Sonnet 4.6	28.3	0.0156	20	0.95	0.95
GPT-5	27.8	0.0154	19	0.95	0.89
Gemini 2.5 Flash	27.5	0.0141	7	0.92	0.91
Claude Opus 4.8	27.1	0.0143	11	0.93	0.98
Gemini 3.1 Flash-Lite	25.4	0.0124	3	0.88	0.95
GPT-4.1	25.2	0.0129	4	0.89	0.95
GPT-4.1 mini	24.3	0.0127	5	0.89	0.89
Claude Haiku 4.5	23.3	0.0120	2	0.86	0.91
Matched null	15.6	0.0091	0	–	–

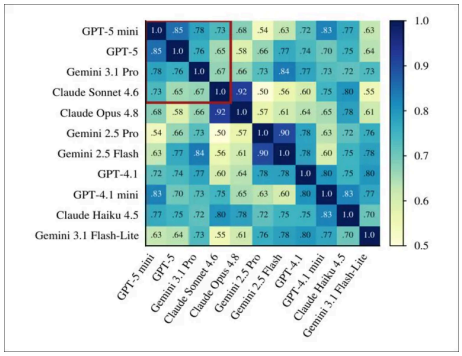
Test results showing how strongly each model concentrated its citations on particular papers and how many papers it ignored entirely. 'Top-10% share' shows how many citations went to the 12 most-favored papers, while 'HHI' measures how concentrated citations were overall. 'Reliability' shows how consistently each model favored the same papers across tests, and ρ shows how similar those preferences were across models. The 'Matched null' row indicates what would be expected if papers were chosen at random.

What Are The Models Actually Favoring?

The preference was found to be overwhelmingly driven by the *content of the papers themselves*. For instance, for GPT-5 mini, about 90% of the variation in which papers were favored was attributable to content, rather than factors such as list order, generation noise or the fabricated metadata attached to each paper. The same 'dominant content' effect was reproduced with GPT-4.1 mini.

The preference map (see below) was also barely changed when titles and abstracts were substantially rewritten while their meaning was preserved. Across four successful paraphrase tests, correlations of 0.95–0.99 were obtained, despite different models from three vendors being used for the rewriting.

Changes in the preference map were observed only when the *underlying meaning* was measurably altered:



Test results comparing citation preferences across the eleven models. The numbers show how closely each pair of models favored the same papers, with higher values indicating greater agreement. Despite differences between models and vendors, broadly similar preferences were found across the group

The models therefore appear to be responding principally to semantic content rather than particular wording. However, the reason for their strong agreement could not be conclusively determined.

It's possible, the authors assert, that a common citation consensus may have been absorbed during training. An alternative possibility is that similar judgments about what constitutes a 'citable' paper may be reached independently.

Therefore the existence of the shared preference was established, but its ultimate origin remains unknown.

The authors suggest that widespread use of AI in research could narrow the range of work receiving attention, because different models tend to favor many of the same papers; and as AI-generated literature accumulates, these shared preferences may become further reinforced through repeated citation, making less-favored research progressively harder to discover. Citation diversity and monitoring of citation concentration are therefore proposed as possible safeguards.

The study's benchmark for recursive citation concentration is now available at GitHub.

Conclusion

Opinion As someone who reads an inordinate number of AI research papers weekly, I have seen 'citation waves' come and go. One perennial stalwart is Google's *Attention Is All You Need*, the original Transformers paper, which must by now be hard-coded into submission templates.

Another repeat-offender, for a long time, was the 2014 offering *Generative Adversarial Networks*, though it would ultimately be supplanted in frequency by *Denoising Diffusion Probabilistic Models* and other diffusion-based tent-pole studies.

In nearly all cases, these 'recurring' citations are not *wrong*, per se; but they are often too widely-scoped to be a useful support to the paper in which they are being cited; and in that sense, they represent something akin to 'citation-washing' – the inclusion of 'reliable' but primarily decorative source quotes, for a low-effort air of credibility.

First published Monday, August 24, 2026. Amended 23:07 EET to add missing plural.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai