

# When the Models Become Kingmakers: Assessing Textual Similarity and Variation in AI LLM Academic Reference Selection

Jason Miklian

Research Professor, Centre for Global Sustainability, University of Oslo

Jason.miklian@globe.uio.no

— Early Discussion Draft —

— 05.10.26 —

## Abstract

Delegating academic reference searches to large language models (LLMs) gives automated systems a role in deciding which scholarship informs academic debate. But what is that role exactly, what gets prioritized, and what does that mean for forward scholarship? To answer, a controlled experiment was run across six scholarly fields to examine how these decisions distribute attention across authors and disciplinary boundaries. This paper presents four main findings. First, all tested LLM models strongly converged on the same overlapping selections when asked for references, with textual similarity the strongest measured predictor. Second, rewriting a paper to use fewer of the same words as the manuscript made the models less likely to recommend it. They also recommended fewer papers from other disciplines. Third, a paper's visibility depends on how its contribution is presented in its abstract. In short, LLM's rarely read past the abstract so whatever is written there matters more than the paper itself. Fourth, (and more encouragingly), regional differences were small and the models had less—but still statistically significant—gender bias than earlier LLM model tests showed. We argue that repeated reliance on recommendations which ultimately favor articles that closely resemble the manuscript being drafted risks narrowing the sources from which scholars develop theories and concretizing prevailing explanations. For researchers, this creates a risk of seeing more of the same while missing useful ideas developed elsewhere. Academic AI thus narrows fields and ideas even while under the guise of broadening useful scholarship through its recommendations.

**Keywords:** Large language models; reference recommendations; citation bias; textual similarity; scholarly visibility; knowledge diversity.

## Introduction

Academics are increasingly using Artificial Intelligence-based large language models (AI LLMs) to assist in academic paper and literature review development. These LLMs typically make recommendations on which papers to read and reference based from a search of titles and abstracts alone, shaping the literature researchers consider and the ideas available for developing an argument. It is in many ways the latest iteration of the longstanding role that academic citation has played to distribute scholarly recognition (or not), affecting whose contributions become visible, e.g. as established scientists often receive more credit and attention than less well-known researchers for comparable contributions (Merton 1968).

However, citation mechanics can be problematic. Existing research shows that men self-cite more often than women, recognition varies across demographic groups, and citation patterns differ by geography (King et al. 2017; Kozlowski et al. 2022; Pan et al. 2012). Emerging studies of LLM models and citations echo or even amplify several of these trends, depending on what models are asked to recommend and how the task is designed (Tian et al. 2025; He 2025).

The question for this paper is therefore twofold: whose scholarship do AI LLMs recommend, and what kinds of ideas do their recommendations bring into view? To answer, six models are compared across six scholarly fields, asking them to choose from the same candidate papers while varying the author information shown. Next, their choices are examined related to the language of the papers and whether their selections include work from other disciplines.

We find substantial agreement across models about which papers to recommend. The models favor papers whose language resembles the manuscript: changing shared wording lowers a reference's chance of selection. They also recommend fewer papers from other disciplines. Differences between geographic regions (e.g. Global North versus Global South scholars) are smaller than they used to be in earlier AI LLM models, and gender makes less difference to LLM outputs than it did in earlier models when male and female names are equally represented. Together, these findings direct attention toward how automated recommendations shape the range of ideas researchers encounter.

For scholarship, the central concern is that repeatedly recommending similar work will reinforce familiar explanations and disciplinary boundaries, especially as the overall volume of scholarly outputs amplifies. This sort of hidden echo chamber challenges the fundamental TKTKT that research develops through encounters with arguments that challenge its assumptions and through connections to work in other fields. Evaluating which useful ideas recommendations leave out addresses the risk of intellectual uniformity identified in wider debates about AI in science (Messerli & Crockett 2024). Therefore, academic AI should be judged partly by whether it helps researchers make those connections.

The societal stakes grow when scholarship informs public decisions. Governments and other institutions draw on research to define problems and decide which responses to fund. Repeatedly overlooking unfamiliar approaches (and using AI LLMs for their own research and analysis) risks limiting and simplifying the evidence available for those choices, especially when relevant contextual knowledge comes from places with less academic visibility. Such inequalities already affect citation: leading scientific countries receive more recognition than other countries doing similar research (Gomez et al. 2022). The design of academic AI therefore matters for whose knowledge helps shape public priorities.

## Literature and conceptual framework

Citation has, for better or for worse, largely become *the* empirical recognition of scholarly quality. It compounds over time to build a perceived evidence about which scholarship matters, by which scholars. We know this system has flaws. Work associated with prominent scientists attracts disproportionate recognition, creating an advantage that subsequent recognition can reinforce (Merton 1968; DiPrete & Eirich 2006). In a sample covering natural, medical and health, and agricultural sciences, the top one percent of scientists' share of citations rose from fourteen to twenty-one percent between 2000 and 2015 (Nielsen & Andersen 2021).

These inequalities involve different processes. Self-citation concerns authors' choices about their own work, whereas citation disparities concern the recognition a contribution receives from others (King et al. 2017; Kozlowski et al. 2022). Geographic citation patterns are associated with distance and research resources (Pan et al. 2012). Previous citation impact and international collaboration were associated with impact across countries, while specialization was more strongly associated with impact in the Global South (Confraria et al. 2017).

The distribution of recognition also affects ideas. As any scholar trying to get an “out there” new idea published can attest, novel combinations can be less likely to become highly cited within short assessment periods, and high publication volume can concentrate attention on established work (Wang et al. 2017; Chu & Evans 2021), and underrepresented scholars introduce novel conceptual combinations more often but receive less recognition for them (Hofstra et al. 2020). These findings make selection across intellectual boundaries relevant to the study of science, alongside selection across author groups (Fortunato et al. 2018).

In an ideal scholarly world, adding a reference contributes to an argument by showing the reader what has come before, how this new argument recognizes that grounding, and possibly in what knowledge gaps remain. Citation-function research examines reasons for citing and their expression in surrounding text. Its classification schemes distinguish relationships that citation counts cannot recover, using different categories and annotation evidence (Zhang et al. 2025). Therefore, relevance is itself a constitutive relationship between candidate and manuscript: a source can supply a method, challenge an explanation, or establish a finding's limits.

These contributions need not always share a topic: A theory developed elsewhere can offer an explanation absent from the focal literature (and is a common way to add new field insights), but its inclusion still requires a judgment about its application, and the information we work with in this process can constrain that judgment. A title and abstract aim to summarize a study's question and principal findings, while a methods section or discussion of competing explanations can reveal additional reasons to cite a reference. A paper's abstract should therefore describe the manuscript's most useful contribution to the argument, but such alignment is not always guaranteed. This creates a distinction between recognizing a relevant paper based on its presentation and recognizing a useful scholarly argument.

A reference list also reflects decisions about how candidates complement one another. Several individually relevant papers can repeat the same argument, whereas a less similar paper can contribute a needed comparison, and the value of that comparison depends on what the other selections already provide. One must therefore distinguish selection of individual references from coverage of the scholarly relationships needed by the focal manuscript.

Research on language models provides grounds for testing geographic cues. Training data and deployment practices can reproduce unequal treatment, although the behavior must be established in the task under study (Bender et al. 2021). In evaluations of places, models assign lower subjective ratings to locations with poorer socioeconomic conditions, with differences between

models (Manvi et al. 2024). Evidence from citation tasks also varies; references generated from model memory favor highly cited work, and fabricated references can reproduce features of human citation patterns (Algaba et al. 2025; Tian et al. 2025). Perhaps due to constructed bias, or perhaps in mimicking the scholarly world in practice, LLMs can favor male-coded references (He 2025).

Experiments that change venue prestige, citation counts, and author status while preserving a paper's title and abstract show that authority information can redirect the choice of a single recommended paper (Jinadu et al. 2026). Read alongside the gender findings, this evidence raises a more specific problem: sensitivity to established reputation and sensitivity to geographic or gender cues need not move together, so do location or gender cues themselves affect selection across scholarly fields? Establishing that distinction matters because a contribution can gain or lose visibility through the LLM's coding before a researcher has even seen the result of their prompt. Therefore we ask **RQ1:** *How does changing displayed institutional location or gender-coded names affect a reference's chance of selection by leading LLMs?*

A separate question concerns agreement among LLM models. Correlated decisions can reduce the value of having several decision-makers, even when a common algorithm improves individual accuracy (Kleinberg & Raghavan 2021). Shared training data can produce more similar outcomes across deployments; effects of model sharing vary with adaptation (Bommasani et al. 2022). Repeated use of common selection systems can concentrate exclusion on the same people (Creel & Hellman 2022), and agreement across models has been documented. Errors correlate across providers (Kim et al. 2025); synthetic survey responses echo their deviations when compared with human responses (Miklian et al. 2026); studies of assisted writing and ideation found greater similarity across many experiments among LLM users (Doshi & Hauser 2024; Anderson et al. 2024; Padmakumar & He 2024).

For academic work, the implications are clear. Models select the same papers because those papers are seen to be relevant, thus making them more relevant, thus making them more cited, thus making them more selected by future LLMs. Evaluating concentration therefore requires more than a comparison with random choice; it also requires evidence about the references selected less often. Shared representations offer one possible explanation for similar behavior across models (Huh et al. 2024), but the consequences of agreement depend on the level of evaluation. An individual researcher can benefit from a familiar source, while the community also needs competing approaches. A conceptual account of AI in science suggests that confidence in tools' apparent objectivity and productivity can obscure the dominance of particular questions and methods (Messeri & Crockett 2024). Research on developer worldviews connects design decisions with inequalities in information quality (Miklian & Hoelscher 2025), as repeated agreement can be mistaken for adequate coverage or a saturation point.

Models from different providers can converge on closely similar answers even when prompts allow many plausible responses (Jiang et al. 2025). Evidence about similar answers and single-paper choices leaves open how far agreement extends across a proposed bibliography. Models might share a small core of references while selecting substantially different additional sources, or repeatedly reproduce most of the same list. The difference determines how much opportunity consulting another model creates to encounter an alternative explanation. Therefore we ask **RQ2:** *How much do models from different developers agree on which references to recommend from the same candidates?*

Recent evidence reinforces the need to separate individual outcomes from collective coverage. An analysis of 41.3 million natural-science papers associated AI-augmented research with greater publication and citation activity, but a much narrower collective range of topics (Hao et al. 2026). In short, gains for the users of a technology can coexist with a contraction in the questions receiving attention. LLMs have also been shown to perpetuate conventional wisdoms as opposed to

delivering new insights (Miklian 2026b). It is also pervasive: one study estimated that LLMs had modified 22.5% of sentences in computer science abstracts by 2024 (Liang et al. 2025). The issue has also broken through to the public; accusations of AI use function as social gatekeeping around perceived authenticity (Miklian & Katsos 2026).

Reference selection also differs from citation fabrication. While LLM bibliographies used to contain substantial fabrication and errors (Walters & Wilder 2023), LLM capabilities have improved to the point that leading models are much less likely to hallucinate. Related work on conversational search associates LLM assistance with selective exposure under some conditions (Sharma et al. 2024). And geographic inequality remains relevant even when a task uses genuine references. Technologies developed around Western interests can reproduce relations of dependence when deployed elsewhere (Birhane 2020).

Geographic recognition and similarity are already connected in research on human citation practices. Comparing international citation flows with the similarity of published abstracts across nearly twenty million papers, leading scientific countries are much more cited relative to other countries doing similar work (Gomez et al. 2022). Geographic categories also require a choice about whose experience an average represents. Grouping institutions as Global South does not imply equivalent resources, research systems, or model treatment. Pooling places can combine positive and negative selection differences; averaging models can cancel opposing responses to the same subregion. Broad comparisons describe average responses under the chosen grouping and weights, while disaggregation identifies where that summary fails to describe individual models.

In scientific literature retrieval, keyword search outperformed two LLM retrievers on questions requiring a specific paper, while its advantage narrowed on questions seeking broader sets of evidence (Hu et al. 2026). The relation between language and discovery is consequential, but leave open whether textual resemblance continues to shape choices once candidate papers are already available to a model, especially as they can struggle to distinguish interdisciplinary combinations from feasible combinations within one field (Shen et al. 2025). Success under an explicit request to cross disciplines leaves unresolved whether such connections enter ordinary reference recommendations. If shared wording favors selection, scholars working across distinct vocabularies may need to write their abstracts in the focal field's terms to gain attention. Therefore we ask **RQ3:** *How are title and abstract similarity and disciplinary proximity associated with selection, and how does selection change when shared wording is reduced?*

## Data and design

We compared six LLMs' reference recommendations across six scholarly fields. Models were asked to select and rank the ten most relevant references from thirty papers in a focal paper's published bibliography, using titles and abstracts. We adapted an attribute-swap procedure to test responses to displayed institutional affiliations and gender-coded names while holding the texts fixed (He 2025). We also examined agreement across models and how textual similarity and disciplinary proximity relate to selection. A separate rewriting experiment tested whether reducing shared vocabulary changed selection.

We sampled 396 papers from OpenAlex, sixty-six in each of six OECD fields: Social Sciences, Engineering and Technology, Natural Sciences, Humanities, Medical and Health Sciences, and Agricultural Sciences. The sampling window was 1 February to 31 May 2026. Eligible English-language articles had a title and abstract, at least fifty references, and thirty candidate articles with titles and abstracts. We retrieved cited works in batches, retained up to the first thirty-four eligible returns, and used the first thirty in the saved order. Candidate selection therefore followed retrieval order rather than a random draw from each bibliography.

We retained the earliest valid response in the saved file for each arm, prompt and model. Valid responses contained exactly ten distinct candidate identifiers from the supplied pool; invalid responses were excluded without repair. Repeated valid responses were not treated as independent replicates. This yielded 10,455 region calls, 9,333 gender calls and 404 calls from the partial imbalanced arm. The region analysis covers 328 focal pools; the combined arms cover 330.

Displayed affiliations were drawn from an OpenAlex institution bank with thirty entries for the United States, Europe and each of four Global South subregions. Institutions were classified as United States, Europe, or Global South and assigned a prestige proxy based on two-year mean citedness. Each thirty-reference pool contained ten affiliations from each region. Six Latin-square-style rotations balanced regions across list-position bands, and affiliations were resampled across rotations. Prompts displayed institution names and countries; numerical citedness remained an analysis variable. Region-arm names were drawn independently from a common bank. Latin America and Africa each appeared in two rotations; South Asia and East Asia each appeared in one, at positions 1-10 and 21-30 respectively; candidate order stayed fixed. This allowed comparisons of the same reference across displayed regions and prestige values. Reassignment does not eliminate the association between an institution's region and its citedness.

The six models were gpt-5.4-mini (OpenAI), claude-haiku-4-5 (Anthropic), gemini-3.5-flash (Google), grok-4.20-0309-non-reasoning (xAI), deepseek-v4-flash (DeepSeek), and qwen-flash (Alibaba). Four developers are based in the United States and two in China. We set temperature to zero for five providers; the Gemini request omitted temperature and specified low reasoning effort. Qwen thinking was disabled. GPT received a completion-token cap without an explicit reasoning-effort setting. Provider-specific token limits and the model identifiers are included in the replication materials. Prompts supplied the focal paper's title and abstract, with each candidate's author names, title and abstract, plus affiliation in the region arm. The models returned ranked candidate identifiers as a JSON list.

The gender arm used names from the supplied gender-coded bank, with fifteen male-coded and fifteen female-coded candidates per pool. Names were reassigned across six rotations, and affiliation was omitted from the displayed prompt. The resulting within-reference comparison estimates the response to gender cues in the selected names under a balanced-pool design. It addresses the balanced condition of the earlier experiment; majority-gender amplification in imbalanced pools was not tested in this arm (He 2025).

Selection-rate ratios (SRR) and normalized selection differences (NSD) compare selection with availability in the pool. Selecting ten of thirty candidates gives an availability-proportional rate of one third. The reference-level logistic regression includes region, institutional prestige, list position, and field, with the United States as the regional reference category and standard errors clustered by focal paper. The principal regional contrast adjusts for institutional prestige. The study was not preregistered. We also compare the unadjusted regional difference and its 90% interval with a  $\pm 1$ -percentage-point margin as a practical-size diagnostic.

Cross-model agreement is measured by the mean pairwise Jaccard overlap of selected sets and compared with the expected overlap of 0.207 under independent uniform ten-of-thirty selection. The predictor analysis relates mean selection frequency across models and rotations to TF-IDF cosine similarity between candidate and focal title-plus-abstract text, with focal-paper fixed effects. A held-out comparison fits the model on half the pools and predicts membership in the approximate top third of selection rates in the other half, using average ranks for ties. An exploratory linear probability model includes fixed effects for each candidate within its focal paper and for each prompt, controls for log institutional prestige, and clusters standard errors by focal paper.

Exploratory tests of differences among models use Wald contrasts with the joint covariance estimated by resampling focal papers, preserving dependence from shared pools.

We tested sensitivity to shared vocabulary by rewriting top-consensus references in three conditions. The shared-word edit replaced content words also present in the focal manuscript with synonyms intended to preserve meaning. The placebo changed grammar while retaining vocabulary. For the third condition, we requested replacement of an equal number of non-shared words while retaining overlap. The edits did not receive independent human assessment of meaning or fluency. The comparison measures the effect of the edits; attributing it to vocabulary changes depends on comparable meaning and fluency.

The paired analysis matches valid responses across all three edited conditions within each target and model. We excluded one target whose shared-word rewrite ended before the results. The analysis includes twenty-nine targets from twenty-eight focal papers and 143 matched model triplets: twenty-seven targets have five models and two have four. Target means receive equal weight, and 95% intervals resample focal papers in 5,000 bootstrap draws using NumPy default\_rng with seed zero. Gemini was excluded from this run because of JSON-truncation errors. The original selection rate is a historical descriptive comparison, outside the matched contrasts.

The observational analysis uses saved candidate metadata derived from OpenAlex for 9,374 candidate-pool observations across 330 focal papers. These observations cover 94.7% of the 9,900 candidate-pool pairs represented in the combined region, gender and partial imbalanced arms. The saved fields identify first-author affiliation country, team-level affiliation indicators, publication language, citation count and primary field. We use the first listed country for the first author, falling back to the first available author-country. Team indicators record any affiliation in China, Hong Kong or Macao, any Global South affiliation, and affiliations exclusively outside the United States, United Kingdom, Ireland, Australia, New Zealand and Canada. These indicators describe institutional locations, not authors' nationality or language proficiency. A cross-field indicator compares candidate and focal OpenAlex primary fields; 54.2% differ.

Selection frequency is the proportion of valid retained calls selecting a candidate, among calls offering that candidate. Regressions give each candidate-pool observation equal weight and include focal-paper fixed effects, standard errors clustered by focal paper, and controls for citation count, age, author count, combined title-and-abstract length, and publication language. Unknown language is retained as a separate category and excluded from the non-English contrast. We reconstruct global TF-IDF similarity from the saved texts and standardize it within pools for the metadata regressions.

A separate saved dataset supplies citation-context counts and influential-citation flags for 3,635 observations in 133 focal pools. Those counts serve as a prominence proxy. Primary estimates use the saved July measurements; separate OpenAlex and Semantic Scholar records obtained on 7 September 2026 provide a source comparison. The replication materials include both sets of measurements and the September DOI links.

## Results

Selection was close to proportional availability for United States, European and Global South affiliations across all six models (SRRs 0.978 to 1.019; Table 1). Relative to the United States, the adjusted Global South shortfall was small but statistically distinguishable from zero (odds ratio 0.978,  $p = 0.0003$ ); the European difference was not (odds ratio 0.993,  $p = 0.260$ ). The unadjusted Global South-minus-United States difference was -0.50 percentage points, with a 90% focal-bootstrap interval of -0.72 to -0.29 points, entirely within a margin of  $\pm 1$  percentage point.

**Table 1. Selection by displayed attributes.**

| Model    | SRR United States   | SRR Europe          | SRR Global South    | Female-name SRR     | Prestige (pp/SD)    |
|----------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Claude   | 1.000 [0.991,1.008] | 1.007 [0.997,1.016] | 0.994 [0.985,1.001] | 0.998 [0.987,1.009] | +0.42 [+0.24,+0.60] |
| DeepSeek | 1.005 [0.995,1.015] | 0.997 [0.987,1.007] | 0.999 [0.989,1.009] | 0.999 [0.988,1.009] | +0.19 [-0.04,+0.44] |
| Gemini   | 0.994 [0.986,1.002] | 1.001 [0.993,1.009] | 1.005 [0.997,1.013] | 1.001 [0.986,1.016] | +0.11 [-0.08,+0.28] |
| GPT      | 1.019 [1.007,1.032] | 0.986 [0.974,0.997] | 0.995 [0.983,1.008] | 0.995 [0.978,1.010] | +0.49 [+0.22,+0.72] |
| Grok     | 1.017 [1.006,1.029] | 1.005 [0.996,1.013] | 0.978 [0.968,0.989] | 0.999 [0.989,1.009] | +0.16 [-0.06,+0.36] |
| Qwen     | 1.006 [0.994,1.018] | 1.011 [0.999,1.023] | 0.982 [0.969,0.995] | 0.991 [0.980,1.001] | +0.38 [+0.17,+0.62] |

Notes: SRR = selection-rate ratio; 1.000 indicates proportional selection. Brackets show 95% confidence intervals from resampling focal papers. Prestige is the within-reference percentage-point change per standard deviation of log displayed institutional citedness.

In the partial imbalanced-pool run, majority-region selection-rate ratios clustered around 1.000, indicating selection broadly proportional to availability. Unadjusted country-level differences between models were larger and are reported with the subregional results in Table 3.

The balanced gender arm showed little pooled difference in selection. The within-reference female-minus-male estimate was -0.11 percentage points (95% CI -0.35 to +0.07). This is consistent with the earlier study's gender-even condition, which also found no significant gender difference. The present arm does not retest its imbalanced-pool findings (He 2025).

Prestige effects were small under the tested assignments. In the within-reference specification, a one-standard-deviation increase in log institutional citedness was associated with selection increases below half a percentage point for GPT, Claude and Qwen; their 95% intervals excluded zero. Estimates for Gemini, Grok and DeepSeek included zero (Table 1).

### *Selection overlap*

On 1,124 matched presentations completed by all six models, model pairs selected an average of 6.45 references in common. On these presentations, 36.7% of candidates were selected by none of the six models, compared with 8.8% under the random null. The ten most frequently selected references accounted for 77.6% of selections. The strongest pairwise agreement was between Gemini and DeepSeek, developed in different countries, with a Jaccard overlap of 0.620. Qwen had the lowest agreement with the other models.

Textual similarity was the strongest measured predictor of selection. TF-IDF similarity between candidate and focal-paper text correlated with mean selection frequency at +0.556. Similarity explained 30.9% of within-pool variation; adding citation count, publication year, author count, and text lengths increased this to 32.7%. Abstract similarity had a larger standardized coefficient than title similarity (+0.461 versus +0.152). In held-out pools, similarity alone predicted consensus membership with an AUC of 0.785.

Selection overlap varied little across the six sampled fields. Mean pairwise Jaccard values ranged from 0.483 in Social Sciences to 0.508 in Natural Sciences. On matched six-model presentations, the unselected share ranged from 36% to 38%, and the entropy-based effective number ranged from 15.7 to 16.2 (Table 2).

**Table 2. Selection concentration by OECD field.**

| Field                       | Matched prompts | Mean Jaccard | Similarity association | Effective number | Unselected (%) |
|-----------------------------|-----------------|--------------|------------------------|------------------|----------------|
| Natural Sciences            | 164             | 0.508        | +0.446                 | 15.8             | 37%            |
| Engineering and Technology  | 138             | 0.506        | +0.439                 | 15.7             | 38%            |
| Medical and Health Sciences | 193             | 0.497        | +0.388                 | 15.9             | 37%            |
| Agricultural Sciences       | 216             | 0.491        | +0.417                 | 16.1             | 36%            |
| Humanities                  | 207             | 0.485        | +0.364                 | 16.2             | 36%            |
| Social Sciences             | 206             | 0.483        | +0.382                 | 16.2             | 36%            |

Notes: Similarity association is the mean within-prompt correlation between selection and candidate-to-focal similarity. Effective number is exponentiated entropy.

A linear latent-semantic representation constructed from the same term space as TF-IDF performed similarly in predicting selection. The two measures correlated at 0.884 and each produced an average overlap of 4.92 references with a model's selected ten. In the matched wording comparison, selection under the shared-word edit was 4.8 percentage points lower than under the grammar-edit placebo (95% CI 0.6 to 9.7) and 5.7 points lower than under the non-shared-word control (95% CI 2.1 to 10.0). The interval for the difference between the two controls included zero. Even after the shared-word edit, targets were selected in 92.2% of cases.

#### *Subregional differences between models*

Unadjusted selection rates varied across Global South subregions (Table 3). Model-specific estimates differed in direction. Claude and Qwen selected African and South Asian affiliations above their proportional availability and East Asian and Latin American affiliations below it; Claude's South Asia selection rate was 24.1% above proportional availability. Gemini and GPT showed the opposite direction across these four subregions. Eighteen of the twenty-four model-by-subregion comparisons differed from parity after Benjamini-Hochberg correction across that set.

**Table 3. Unadjusted selection-rate ratios by model and Global South subregion.**

| Model    | Latin America | Africa | South Asia | East Asia | Chinese affiliation |
|----------|---------------|--------|------------|-----------|---------------------|
| GPT      | 1.054         | 0.963  | 0.902      | 1.036     | 1.059               |
| Claude   | 0.868         | 1.120  | 1.241      | 0.744     | 0.718               |
| Gemini   | 1.106         | 0.928  | 0.811      | 1.146     | 1.140               |
| Grok     | 0.999         | 1.022  | 0.955      | 0.874     | 0.868               |
| DeepSeek | 0.983         | 1.038  | 1.034      | 0.916     | 0.920               |
| Qwen     | 0.858         | 1.102  | 1.230      | 0.742     | 0.746               |

Notes: SRR = 1.000 indicates proportional selection. The Chinese-affiliation subset overlaps the regional groups.

Unadjusted differences among models were more pronounced within Southern subregions than in the broad United States and European categories. Wald tests indicated differences among models in each regional group. Within-reference regressions controlling for prompt and institutional prestige

gave Claude a South Asia-minus-United States estimate of +0.28 percentage points (95% CI -1.11 to +1.66). Across the twenty-four adjusted Global South contrasts, Benjamini-Hochberg q-values exceeded 0.09.

The Chinese-affiliation results did not align consistently with developer country: Claude and Qwen showed similar directions despite their different national origins, as did GPT and Gemini (Table 3). Mean candidate-level disagreement was also similar for Chinese and other affiliations on presentations completed by all six models.

The observational analysis used authors' recorded affiliations (Table 4A). Confidence intervals for differences by first-author affiliation region included zero for all groups except developing East Asia, whose estimate was +4.0 percentage points (95% CI +0.2 to +7.9). References with any Chinese-affiliated author had a +5.1-point selection difference (95% CI +2.0 to +8.2), declining to +2.8 after adjustment for similarity (Table 4B). Chinese-affiliated and Global South references had higher mean similarity by 0.13 and 0.09 standard deviations, respectively. These associations do not establish a mechanism of linguistic adaptation or equivalence across author populations.

**Table 4. Selection and textual similarity by observed author affiliations.**

| A. First-author affiliation region                               | Selection (pp)       | 95% CI            |                          |
|------------------------------------------------------------------|----------------------|-------------------|--------------------------|
| Europe                                                           | +0.6                 | [-1.8, +3.0]      |                          |
| Anglosphere (CA, AU, NZ)                                         | -1.8                 | [-5.0, +1.3]      |                          |
| High-income Asia and Israel                                      | +0.3                 | [-4.0, +4.6]      |                          |
| Latin America                                                    | +0.9                 | [-5.2, +7.0]      |                          |
| Africa                                                           | -1.2                 | [-7.0, +4.5]      |                          |
| South Asia                                                       | +0.6                 | [-5.5, +6.7]      |                          |
| East Asia (developing)                                           | +4.0                 | [+0.2, +7.9]      |                          |
| Middle East and selected neighbouring countries                  | -0.4                 | [-5.7, +4.9]      |                          |
| Other mapped countries                                           | -4.2                 | [-11.6, +3.3]     |                          |
| B. Team affiliation indicator                                    | Similarity (SD)      | Selection (pp)    | Similarity-adjusted (pp) |
| Any Chinese-affiliated author                                    | +0.13 [+0.04, +0.21] | +5.1 [+2.0, +8.2] | +2.8 [+0.3, +5.4]        |
| Any Global South author                                          | +0.09 [+0.03, +0.16] | +2.0 [-0.4, +4.3] | +0.3 [-1.6, +2.2]        |
| All observed affiliations outside the six-country Anglophone set | +0.05 [-0.00, +0.11] | +1.3 [-0.7, +3.3] | +0.4 [-1.3, +2.0]        |

Notes: The United States is the reference category in Panel A. Panel B's final column adds textual similarity. Estimates use the controls in Section 3; brackets show 95% confidence intervals clustered by focal paper. Middle East includes Turkey; Africa includes North Africa.

References classified outside the focal paper's primary field were selected less often (Table 5). Their mean selection rate was 29.7%, compared with 38.7% for same-field references. With focal-paper fixed effects and controls, the selection difference was -14.1 percentage points (95% CI -16.6 to -11.7). The adjusted selection difference appeared in all six fields ( $p < 0.001$  in each).

Cross-field references also had lower textual similarity. Adding similarity reduced the selection difference by 61%, to -5.5 percentage points (95% CI -7.7 to -3.4). This reduction reflects statistical adjustment and does not identify a mediated share.

**Table 5. Selection differences for references outside the focal paper's primary field.**

| Outcome                                   | Cross-field effect | 95% CI         |
|-------------------------------------------|--------------------|----------------|
| Similarity to manuscript (SD)             | -0.49              | [-0.56, -0.43] |
| Selection (pp)                            | -14.1              | [-16.6, -11.7] |
| Selection adjusted for similarity (pp)    | -5.5               | [-7.7, -3.4]   |
| Never selected (pp)                       | +11.3              | [+8.9, +13.7]  |
| Nonselection adjusted for similarity (pp) | +5.7               | [+3.5, +7.9]   |

Notes: Fields use OpenAlex classifications; estimates follow Table 4. Never selected means no selections across completed prompts, rather than within a single matched presentation.

The saved prominence subsample contains 3,635 candidate observations from 133 focal papers. Citation-context counts correlated with selection at 0.203 within pools, compared with 0.503 for textual similarity. In the cross-field specification with the other controls, adding raw context counts changed the selection difference from -17.6 to -16.5 percentage points. Adding similarity as well reduced it to -6.5 points (95% CI -10.0 to -3.0).

September metadata classifications agreed with the saved values on 98.77% to 100% of comparable references, depending on the field. In the restricted secondary sample across 236 papers, similarity again correlated more strongly with selection than citation-context counts (0.498 versus 0.157).

## Discussion

RQ1 asked whether changing displayed institutional location or gender-coded names affects reference selection. The broad regional differences were small, and the balanced gender comparison showed little difference. The gender finding agrees with the gender-even condition in earlier research, while majority-group preferences arose under different candidate compositions (He 2025).

The distinction between recognizing a scholar and evaluating an available paper helps explain how these results fit the wider literature. An evaluation of 100,000 physicists found that LLM recognition was associated with citation impact and uneven across gender and geography (Liu et al. 2025). Our experiment supplies the candidate texts directly, removing the need to recall their authors or discover their work. Small differences at this selection stage can therefore coexist with unequal visibility before a paper reaches the candidate list. Authority cues also merit separate attention: experiments varying author status, venue and citations redirected single-paper recommendations even with the title and abstract fixed (Jinadu et al. 2026).

Small broad regional differences can conceal variation in how individual models respond to subregions. The opposing unadjusted patterns are relevant to evaluation because a single Global South average can combine gains for one group with losses for another. Yet the largest differences shrink when the same reference is compared across displayed affiliations. The practical implication of RQ1 is to evaluate the same candidate under different cues and report results at more than one geographic scale, as the contrast between the pooled estimates and Table 3 illustrates.

RQ2 asked how much models from different developers agree when choosing from the same candidates. The overlap was substantial: model pairs shared an average of 6.45 of their ten selections, and more than a third of candidates were selected by none of the six models on matched presentations. The pattern varied little across fields. Evidence of similar open-ended answers and correlated errors thus extends to the composition of recommended bibliographies, where agreement affects which papers a researcher encounters (Jiang et al. 2025; Kim et al. 2025). The strongest agreeing pair, Gemini and DeepSeek, also shows that developer-country differences offer little assurance of an independent second reading.

Consulting several models can consequently reproduce the same omissions. This gives a concrete scholarly application to the concern that correlated selection systems concentrate exclusion, even when each system appears useful to its individual user (Kleinberg & Raghavan 2021; Creel & Hellman 2022). A researcher seeking familiar work can benefit from a stable set of recommendations; a field developing new explanations also needs routes to sources outside that set. The distinction between individual benefit and collective coverage is consistent with the association between AI-augmented research, greater individual scientific impact, and narrower collective topic coverage (Hao et al. 2026).

RQ3 asked how textual similarity and disciplinary proximity relate to selection, and what happens when shared wording is reduced. Textual similarity was the strongest measured predictor, including in held-out pools. This locates a consequential choice after retrieval: supplying a paper to a model does not ensure that its contribution will be recommended. Our findings extend the evaluation of scholarly search to how models rank an already available set, where finding additional papers alone cannot resolve a preference for familiar language. Recent retrieval benchmarks show a related dependence on research purpose: performance changes between questions seeking a specific paper and those seeking broader evidence (Hu et al. 2026).

The wording intervention identifies a practical vulnerability in how a recommendation is elicited. Reducing shared words lowered selection relative to both the grammar-edit placebo and the control that changed non-shared words. Their high selection rates make a general response to any edited text a less satisfactory explanation for the contrast. We interpret the intervention as evidence that the presentation of a paper affects its chance of recommendation. The effect was modest but consequential for references that had previously been selected almost universally: selection fell by 4.8 percentage points relative to the placebo, while remaining at 92.2% after the shared-word edit.

Disciplinary proximity adds a further concern for the range of scholarship brought into view. Cross-field references were selected less often in every sampled field, and the difference remained after adjustment for textual similarity (Table 5). Shared language accounts statistically for much of the difference, but disciplinary classification also distinguishes papers whose purposes and contributions can differ. The finding therefore motivates closer examination of which useful connections are lost when models assemble an ordinary bibliography. The comparison complements interdisciplinary benchmarks that explicitly ask models to recognize combinations across fields: such capabilities do not by themselves establish that a model will recommend those connections without being asked to cross a disciplinary boundary (Shen et al. 2025).

A source's contribution depends on how it relates to an argument. A method developed elsewhere can be useful despite different terminology, and a paper challenging the focal explanation can matter precisely because it reaches another conclusion. The selection patterns observed here make such relationships a priority for evaluating omitted references. Reaching unfamiliar authors and encountering unfamiliar arguments are different achievements. In a masked-citation task using NLP papers, LLMs drew on more socially distant authors while producing fewer critical citations and favoring established work; citation intent was assessed using LLM judges (Liu et al. 2026).

Repeated use could extend a selection difference beyond a single recommendation. If researchers adopt suggested references, those references gain citations; if subsequent search or training systems respond to that visibility, the initial advantage could recur. For cumulative-advantage theory, this suggests a pathway through the presentation of scholarship: a contribution can gain repeated exposure because its abstract resembles the manuscript being drafted. Researchers could then face incentives to express unfamiliar ideas in already familiar terms. Feedback through users' judgments and later model training are distinct processes examined in human-AI interaction and recursive-training research (Glickman & Sharot 2025; Shumailov et al. 2024). Applied to scholarly discovery, these processes give a testable basis for the concern that repeated reliance on AI can erode access to less familiar knowledge (Peterson 2025).

Whose scholarship becomes visible depends on access to the candidate list as well as selection within it. The positive similarity associations for Chinese-affiliated and Global South references describe work that had already been published, indexed, cited and supplied with an abstract (Table 4B). They cannot represent scholars whose work rarely reaches those stages. This distinction matters when interpreting small pooled differences: authors can receive comparable treatment once visible while facing unequal costs of becoming visible. Additional burdens of reading, writing and presenting in English are documented among non-native English-speaking scientists (Amano et al. 2023). Recognition inequalities also affect ideas, as underrepresented scholars' novel contributions receive less uptake in subsequent work (Hofstra et al. 2020).

The societal implication concerns which accounts of a problem become available for public use. Governments and other institutions draw on scholarship when deciding what merits intervention and which responses deserve funding. If familiar terminology repeatedly determines the evidence brought forward, locally developed explanations can lose visibility even when their authors' displayed affiliations receive comparable treatment. A policy review can then miss the contextual evidence that would qualify a proposed solution's application elsewhere. Evidence from conflict-information tasks also links thinner retrievable records to more fabricated and misattributed answers, showing why access to contextual knowledge matters for AI-assisted public analysis (Miklian 2026a). Scientific recognition is already concentrated in countries whose work receives disproportionate citation relative to other countries doing similar research (Gomez et al. 2022).

Evaluation should therefore assess references against a stated research purpose. Independent readers could identify each candidate's contribution and record what its omission would cost the selected set. Several reference sets may be useful. Recent literature-search research shows that recovering a human bibliography and scoring topical relevance with an LLM judge can reward different kinds of citations, supporting evaluation across several dimensions (Sahu et al. 2026). Our candidates' presence in published bibliographies establishes prior scholarly use, while leaving their specific value to the focal argument to be assessed. Citation-function classifications offer a starting vocabulary for distinguishing evidence, methods and contrasting explanations, with categories adapted to the fields being compared (Zhang et al. 2025).

For researchers and institutions adopting academic AI, the findings support preserving opportunities to encounter useful alternatives. Asking a model for a counterargument or a method from another field makes a different scholarly purpose explicit; whether that improves recommendations should be tested against independently assessed contributions. A direct comparison could give the same candidates alongside the focal abstract, a fuller account of the argument, or relevant full-text passages, while holding the recommendation budget fixed. Combinations of models should likewise be judged by the additional useful references they recover for a given reading effort. These evaluations would make intellectual coverage a criterion of success alongside efficiency, addressing the risk that AI-assisted science produces more while narrowing the questions and methods considered legitimate (Messeri & Crockett 2024).

## Limitations and Future Research

This study has several limitations. First, the findings come from supplied bibliographies, applying to eligible English-language focal papers across six fields, with candidates already cited and available with titles and abstracts, restricting database coverage and availability (Culbert et al. 2025). The design therefore does not establish representation across the wider literature or compare models' independent discovery of sources. Retrieval-order sampling and exclusion of invalid responses further condition the population, and generalization is limited to the six models and provider settings tested; changing prompts, candidate pools or access to full texts can change selection. The models' training exposure to these papers is unknown, and the available records do not support a controlled comparison across model generations.

In addition, the attribute swaps measure responses to displayed location and gender-coded names, while the recovered-affiliation comparisons are observational. Neither establishes treatment across all signals of identity, and affiliation country does not identify language proficiency. Regional comparisons also retain associations between institutions and prestige; the exploratory subregional results should not be interpreted as national preferences. Inference about sparsely represented author populations is limited by the precision of the observational estimates. Likewise, agreement among models does not prove shared training data or internal representations as its cause (but has been suspected more generally).

Also, selection is an outcome of the recommendation procedure, not an independent judgment of a reference's scholarly value. The wording intervention covers twenty-nine selected targets and five models; meaning and fluency were not independently assessed, so the contrast cannot isolate vocabulary from other consequences of editing. Textual similarity and primary-field labels also provide incomplete measures of the relationship between papers. The cross-field difference is observational, does not identify a causal mechanism or the share of useful scholarship omitted. Finally, the study observes recommendations rather than their adoption, subsequent citations or policy effects; claims about cumulative exclusion require evidence of human processing of model recommendations.

Future research should therefore establish whether these patterns persist when references offer comparable contributions to a manuscript. Independent specialists from the focal and neighbouring fields should assess the theoretical, methodological or empirical value of candidate references, allowing relevance to be evaluated separately from textual similarity. Researchers can then vary disciplinary wording while preserving meaning, testing whether the same reference becomes more likely to be selected when expressed in language closer to the focal paper. Repeating these comparisons across fields and languages would help identify when disciplinary differences in expression obstruct the recommendation of useful scholarship, and whether changes to recommendation systems can reduce that exclusion.

The substantial overlap between models also motivates research on whether shared recommendations produce cumulative concentration. Longitudinal studies could follow recommendations through researchers' reading decisions and into subsequent bibliographies, examining whether repeated exposure increases citations to a common set of papers. Randomised comparisons of literature searches with and without LLM assistance would help distinguish the effects of these tools from researchers' existing reading habits and citation preferences. Building on the variation observed within broad geographical categories, these studies should examine whose work gains visibility at more specific regional levels, alongside changes in the disciplinary and theoretical range of cited scholarship. Experiments can also test whether recommendations that explicitly include relevant work from neighbouring fields broaden researchers' reading and influence the arguments they develop.

**AI Use Disclosure Statement:** The authors used Claude Opus 4.8 (Anthropic) as an adversarial reviewer to assess a first draft of this document, to build tables and verify data, and to take a final copyediting / references pass. OpenAI Codex Astra was used to verify and audit the replication package against the study data, e.g. in coding and denominator corrections and ensuring consistency between the paper and replication materials. All AI-assisted output was carefully reviewed and accepted by the authors, who take full responsibility for all material herein. Neither tool was used to develop the primary argumentation or to assist in the authors' substantive interpretation.

## References

- Algaba, A., Mazijn, C., Holst, V., Tori, F., Wenmackers, S., & Ginis, V. (2025). Large language models reflect human citation patterns with a heightened citation bias. *Findings of the Association for Computational Linguistics: NAACL 2025*, 6844–6879. <https://doi.org/10.18653/v1/2025.findings-naacl.381>
- Amano, T., Ramírez-Castañeda, V., Berdejo-Espinola, V., Borokini, I., Chowdhury, S., Golivets, M., González-Trujillo, J. D., Montaña-Centellas, F., Paudel, K., White, R. L., & Verissimo, D. (2023). The manifold costs of being a non-native English speaker in science. *PLOS Biology*, 21(7), e3002184. <https://doi.org/10.1371/journal.pbio.3002184>
- Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. *Creativity and Cognition (C&C '24)*, 413–425. <https://doi.org/10.1145/3635636.3656204>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed*, 17(2), 389–409. <https://doi.org/10.2966/scrip.170220.389>
- Bol, T., de Vaan, M., & van de Rijdt, A. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences*, 115(19), 4887–4890. <https://doi.org/10.1073/pnas.1719557115>
- Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems*, 35, 3663–3678. <https://doi.org/10.52202/068431-0265>
- Chu, J. S. G., & Evans, J. A. (2021). Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41), e2021636118. <https://doi.org/10.1073/pnas.2021636118>
- Confraria, H., Mira Godinho, M., & Wang, L. (2017). Determinants of citation impact: A comparative analysis of the Global South versus the Global North. *Research Policy*, 46(1), 265–279. <https://doi.org/10.1016/j.respol.2016.11.004>
- Creel, K., & Hellman, D. (2022). The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1), 26–43. <https://doi.org/10.1017/can.2022.3>
- Culbert, J. H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics*, 130(4), 2475–2492. <https://doi.org/10.1007/s11192-025-05293-3>
- DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology*, 32, 271–297. <https://doi.org/10.1146/annurev.soc.32.061604.123127>
- Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), eadn5290. <https://doi.org/10.1126/sciadv.adn5290>
- Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, 359(6379), eaao0185. <https://doi.org/10.1126/science.aao0185>
- Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9(2), 345–359. <https://doi.org/10.1038/s41562-024-02077-2>
- Gomez, C. J., Herman, A. C., & Parigi, P. (2022). Leading countries in global science increasingly receive more citations than other countries doing similar research. *Nature Human Behaviour*, 6(7), 919–929. <https://doi.org/10.1038/s41562-022-01351-5>
- Hao, Q., Xu, F., Li, Y., & Evans, J. (2026). Artificial intelligence tools expand scientists’ impact but contract science’s focus. *Nature*, 649(8099), 1237–1243. <https://doi.org/10.1038/s41586-025-09922-y>
- He, J. (2025). Who gets cited? Gender- and majority-bias in LLM-driven reference selection. *arXiv preprint arXiv:2508.02740*. <https://doi.org/10.48550/arXiv.2508.02740>
- Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., & McFarland, D. A. (2020). The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17), 9284–9291. <https://doi.org/10.1073/pnas.1915378117>
- Hu, T., Zhao, Y., Zhang, C., Cohan, A., & Zhao, C. (2026). SAGE: Benchmarking and Improving Retrieval for Deep Research Agents. *arXiv preprint arXiv:2602.05975*. <https://doi.org/10.48550/arXiv.2602.05975>

- Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). Position: The Platonic representation hypothesis. In Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, 235, pp. 20617–20642). <https://proceedings.mlr.press/v235/huh24a.html>. Preprint version: <https://doi.org/10.48550/arXiv.2405.07987>
- Jiang, L., Chai, Y., Li, M., Liu, M., Fok, R., Dziri, N., Tsvetkov, Y., Sap, M., Albalak, A., & Choi, Y. (2025). Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). arXiv preprint arXiv:2510.22954. <https://doi.org/10.48550/arXiv.2510.22954>
- Jinadu, U., Ghazvinian, P., Budathoki, A., Ampel, B. M., Sunderraman, R., & Ding, Y. (2026). Authority Bias in Conversational Search Engines for Academic Paper Recommendation. arXiv preprint arXiv:2609.00248. <https://doi.org/10.48550/arXiv.2609.00248>
- Kim, E. M., Garg, A., Peng, K., & Garg, N. (2025). Correlated errors in large language models. In Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, 267, pp. 30038–30066). <https://proceedings.mlr.press/v267/kim25e.html>. Preprint version: <https://doi.org/10.48550/arXiv.2506.07962>
- King, M. M., Bergstrom, C. T., Correll, S. J., Jacquet, J., & West, J. D. (2017). Men set their own cites high: Gender and self-citation across fields and over time. *Socius*, 3, 2378023117738903. <https://doi.org/10.1177/2378023117738903>
- Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22), e2018340118. <https://doi.org/10.1073/pnas.2018340118>
- Kozłowski, D., Larivière, V., Sugimoto, C. R., & Monroe-White, T. (2022). Intersectional inequalities in science. *Proceedings of the National Academy of Sciences*, 119(2), e2113067119. <https://doi.org/10.1073/pnas.2113067119>
- Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou, J. (2025). Quantifying large language model usage in scientific papers. *Nature Human Behaviour*, 9(12), 2599–2609. <https://doi.org/10.1038/s41562-025-02273-8>
- Liu, Y., Elekes, Á., Lu, J., Dorantes-Gilardi, R., & Barabási, A.-L. (2025). Unequal Scientific Recognition in the Age of LLMs. Findings of the Association for Computational Linguistics: EMNLP 2025, 23558–23568. <https://doi.org/10.18653/v1/2025.findings-emnlp.1279>
- Liu, Y., Chen, L., Liu, Z., Lu, J., & Murray, D. (2026). Citing Less Critically: LLMs Reshape the Rhetoric and Reach of Scientific Citation. arXiv preprint arXiv:2609.01432. <https://doi.org/10.48550/arXiv.2609.01432>
- Manvi, R., Khanna, S., Burke, M., Lobell, D. B., & Ermon, S. (2024). Large language models are geographically biased. In Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, 235, pp. 34654–34669). <https://proceedings.mlr.press/v235/manvi24a.html>. Preprint version: <https://doi.org/10.48550/arXiv.2402.02680>
- Merton, R. K. (1968). The Matthew effect in science. *Science*, 159(3810), 56–63. <https://doi.org/10.1126/science.159.3810.56>
- Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002), 49–58. <https://doi.org/10.1038/s41586-024-07146-0>
- Miklian, J. (2026a). How artificial intelligence LLM engines shape the global conflict information environment. arXiv preprint arXiv:2607.14197. <https://doi.org/10.48550/arXiv.2607.14197>
- Miklian, J. (2026b). The quiet and the compliant: How regulation and polarization shape conventional wisdoms on corporate social engagement in high-risk settings. arXiv preprint arXiv:2604.06223. <https://doi.org/10.48550/arXiv.2604.06223>
- Miklian, J., & Hoelscher, K. (2025). A new digital divide? Coder worldviews, the slop economy, and democracy in the age of AI. arXiv preprint arXiv:2510.04755. <https://doi.org/10.48550/arXiv.2510.04755>
- Miklian, J., Hoelscher, K., & Katsos, J. E. (2026). Stochastic parrots or singing in harmony? Testing five leading LLMs for their ability to replicate a human survey with synthetic data. arXiv preprint arXiv:2603.00059. <https://doi.org/10.48550/arXiv.2603.00059>
- Miklian, J., & Katsos, J. E. (2026). "That's AI slop, you bot!" Studying accusations, evidence, and credibility in online discourse towards LLM-generated comments. arXiv preprint arXiv:2606.12073. <https://doi.org/10.48550/arXiv.2606.12073>
- Nielsen, M. W., & Andersen, J. P. (2021). Global citation inequality is on the rise. *Proceedings of the National Academy of Sciences*, 118(7), e2012208118. <https://doi.org/10.1073/pnas.2012208118>

- Padmakumar, V., & He, H. (2024). Does writing with language models reduce content diversity? *International Conference on Learning Representations (ICLR 2024)*. Preprint version: <https://doi.org/10.48550/arXiv.2309.05196>
- Pan, R. K., Kaski, K., & Fortunato, S. (2012). World citation and collaboration networks: Uncovering the role of geography in science. *Scientific Reports*, 2, 902. <https://doi.org/10.1038/srep00902>
- Peterson, A. J. (2025). AI and the problem of knowledge collapse. *AI & Society*, 40(5), 3249–3269. <https://doi.org/10.1007/s00146-024-02173-x>
- Sahu, G., Charlin, L., & Pal, C. (2026). Rethinking Literature Search Evaluation: Deep Research Helps, and Human Citation Lists Are Not a Ground Truth. *arXiv preprint arXiv:2605.29234*. <https://doi.org/10.48550/arXiv.2605.29234>
- Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762), 854–856. <https://doi.org/10.1126/science.1121066>
- Sharma, N., Liao, Q. V., & Xiao, Z. (2024). Generative echo chamber? Effect of LLM-powered search systems on diverse information seeking. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, Article 1033, 1–17. <https://doi.org/10.1145/3613904.3642459>
- Shen, Y., de Sousa, D. X., Marçal, R., Guo, H., & Zhu, X. (2025). IDRBench: Understanding the Capability of Large Language Models on Interdisciplinary Research. *arXiv preprint arXiv:2507.15736* (version 4, revised 2026). <https://doi.org/10.48550/arXiv.2507.15736>
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755–759. <https://doi.org/10.1038/s41586-024-07566-y>
- Tian, Y., Liu, Y., Bu, Y., & Liu, J. (2025). Who gets recommended? Investigating gender, race, and country disparities in paper recommendations from large language models. *arXiv preprint arXiv:2501.00367*. <https://doi.org/10.48550/arXiv.2501.00367>
- Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13, 14045. <https://doi.org/10.1038/s41598-023-41032-5>
- Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8), 1416–1436. <https://doi.org/10.1016/j.respol.2017.06.006>
- Zhang, Y., Wang, Y., Sheng, Q. Z., Yao, L., Chen, H., Wang, K., Mahmood, A., Zhang, W. E., Zaib, M., Sagar, S., & Zhao, R. (2025). Deep learning meets bibliometrics: A survey of citation function classification. *Journal of Informetrics*, 19(1), 101608. <https://doi.org/10.1016/j.joi.2024.101608>