Supplementary Information for

LLM hallucinations in the wild: Large-scale evidence from non-existent citations

Zhenyue Zhao, Yihe Wang, Toby Stuart, Mathijs De Vaan, Paul Ginsparg, Yian Yin

S1 Data description

We collected publication data, including paper metadata and raw reference information, across four large-scale datasets.

S1.1 arXiv

arXiv is a major preprint repository, especially central in physics, math, computer science, and related fields. The metadata and full-text PDFs are available on Kaggle (https://www.kaggle.com/datasets/Cornell-University/arxiv). We downloaded the metadata, which contains information about the arXiv papers, including the arXiv id, field category, title, abstract, and author list. We also retrieved the submitter's country of registration from arXiv.

We compiled raw references data from two sources: (1) 30,786,632 references from 756,048 generated from LaTeX source files provided in the original submissions. Each file contains formatted information about the references, including the title, authors, publication year, venue, etc., which we further parse and extract using GROBID[1]. (2) For papers without LaTeX source files, we used GROBID to extract 13,320,897 references from their PDF files.

We also leveraged a unique dataset of 30,981 submissions rejected by arXiv from January 2022 to August 2025. We extracted and parsed the 1,061,660 references using AnyStyle (https://anystyle.io/), an open-source tool that extracts and parses references into structured information. Whereas GROBID is capable of parsing the references from bbl files and PDF files directly, AnyStyle performs well at identifying references in plain-text files and parsing reference strings into structured fields. To assess the consistency of these reference parsing methods, we tested the extraction pipelines on a sample of 2,201 papers where we have both PDF and BBL files, finding high consistency at the paper level (Jaccard similarity > 0.85).

S1.2 bioRxiv

bioRxiv is a leading preprint repository for the life sciences. We collected XML full-text data (including DOI, publication date, and reference list) of 261,928 preprints from Jan 2020 to August 2025. The total 21,183,111 references obtained through the API were already parsed into fields including the title, venue, publication year, author list, etc.

S1.3 SSRN

SSRN (Social Science Research Network) is a major repository for social science scholarship, especially in economics, law, and related fields. As there is no publicly available API or data dump from SSRN, we retrieved paper title, creation date, and reference metadata from Crossref. Among 738,379 SSRN manuscripts with registered DOIs from 2020 to 2025, 421,698 have reference

metadata registered on Crossref, totaling 26,815,043 references. We collected the title, authors, publication year, venue, etc. for the references.

S1.4 PubMed Central

Going beyond preprints, we also examine PubMed Central (PMC), one of the largest full-text corpora of life science research, maintained by the U.S. National Library of Medicine. We downloaded the XML files for papers indexed in the PMC Open Access Subset under either the "commercial use allowed" or "non-commercial use only" license and published between 2020 and 2025.

In more recent years, PMC has implemented the NIH Preprint Pilot and started indexing preprint papers[2]. To ensure that our analysis is not affected by this compositional change, we focus on papers published in peer-reviewed venues and exclude items from preprint repositories – arXiv, bioRxiv, medRxiv, and Research Square (identified from PMC metadata).

For computational efficiency, we focused on a random 10% sample of the data, covering 374,807 papers and 19,245,787 references cited. The metadata contains information about each paper, including its title, DOI, publication year, and reference list. Reference entries have been parsed by NLM's internal algorithm, and classified into different publication types (e.g., journal, confproc, book). The majority (92.4%) of the references were already parsed into fields including the title, venue, author list, publication year, etc. For the rest that were not classified as academic items, only raw reference strings were provided, which we parsed using AnyStyle.

S1.5 Author-level features from Semantic Scholar and OpenAlex

We downloaded the entire datasets of Semantic Scholar and OpenAlex, two of the largest publicly available citation corpora. The arXiv papers were mapped to Semantic Scholar using arXiv ID; the bioRxiv, SSRN and PMC papers were mapped to OpenAlex. We primarily relied on DOI for paper matching and used title for fallback cases in bioRxiv and SSRN.

S2 Methods

S2.1 Verifying the existence of references

We built a three-step title-based pipeline to verify the existence of references in our dataset, detailed as follows. References without an explicit title (18.2% of the dataset) were excluded.

Step 1: Matching with existing bibliographic databases. We then queried an Elasticsearch system built from the combined publication metadata of Semantic Scholar and OpenAlex using the extracted title of each reference. For each query title, the system returned the 20 most similar records. We considered a candidate record as a match if it satisfied one of the three criteria:

(1) a strict match, if the Levenshtein edit distance between the two titles is at most 3;
(2) a loose match, defined as the longer title beginning or ending with the shorter title, where the shorter title is at least 10 alphanumeric characters in length.
(3) a fuzzy match, if the indel distance divided by the total length of two titles is at most 0.1.

Step 2: Cleaning and refining reference information. For references that could not be matched, we used GPT-4o-mini to assess whether the reference was intended to cite an academic document – to exclude non-academic documents (e.g., news articles, technical reports, and policy documents) from our analysis. Because reference parsing was sometimes inaccurate, with venue titles or other components occasionally extracted as document titles, we used an LLM to re-extract titles from raw reference strings when those strings were available (Fig. S1). We then repeated the Elasticsearch search and matching procedure using the updated titles.

Step 3: Validation in Google Scholar. After the first two steps, 1.54% of references were flagged as potentially hallucinated. Some of these may still have corresponded to real academic documents missing from the open bibliographic databases we used. We therefore performed a final Google Scholar search using the full reference string when available, followed by the quoted title and then the unquoted title. Based on results from Google Scholar search, a reference is considered as non-hallucinated if it satisfied one of the three criteria:

(1) the document itself can be found as a paper record under the matching criteria in Step 1;
(2) the document can be found as a citation record (indicated by a [citation] badge), with at least three citations (shown as "Cited by");
(3) the quoted title can be found in at least three records (either as citation records or in papers that explicitly cited the title)[*].

[*] We use the threshold to avoid cases where a hallucinated reference is indexed only from the original citing article. Requiring at least three mentions reduces the risk of treating such self-indexed hallucinations as valid references. This concern is consistent with the post-2024 increase in citation-only entries documented in Fig. 3c.

Notably, the thresholds used here are necessarily imperfect. Our principle is to apply relatively conservative criteria, which may introduce misclassification but are more likely to underestimate than overestimate the magnitude of citation hallucinations. Robustness checks using alternative thresholds and validation rules yield qualitatively consistent temporal trends and cross-field patterns (Fig. S5).

S2.2 Estimating the hallucination rates and counts

Fig. S2 plots the monthly rate of references that could not be verified to exist using our approach (unmatched references), as a fraction of all academic references with a non-empty title that are classified by GPT as academic references. Across all four datasets, unmatched references follow a non-zero yet relatively stable rate before the availability of ChatGPT, which represents a baseline rate of false positives in identifying hallucinated references. Among the unmatched references in the PMC dataset, we spotted citations to documents published in non-English venues, where titles were either translated into English or remained in the original language. To account for these potential false positives, we counted the number of times each venue appeared before 2023 and excluded references to papers published in venues that appeared fewer than 10 times during that period.

Using reference-level records with publication year and venue parsed from the reference strings, we estimated the hallucination rate ϵ_t of references using the following regression model:

$$U_{ivt} = \alpha + \sum_{t \neq t_0} \epsilon_t \, \mathbb{1}\{T = t\} + \gamma_v + \eta_{ivt}$$

where U_{ivt} is an indicator equal to 1 if reference i is unmatched, and 0 otherwise; $\mathbb{1}\{T = t\}$ are month indicators; and γ_v are venue fixed effects that absorb time-invariant differences across venues. For the results displayed in Fig. 1 of the main text, we set November 2022, t_0, as the omitted baseline month, so each ϵ_t can be interpreted as the change in the probability of an unmatched reference in month t relative to November 2022, holding venue composition constant.

We estimate the number of hallucinated references in month t, defined as N_t, by multiplying the total number of references in that month, R_t, with the estimated hallucination ratio ϵ_t

$$N_t = R_t \cdot \epsilon_t$$

We then fit an exponential growth model to the monthly number of hallucinated citations and used the fitted curve to project the final months of 2025 (Fig. S3).

$$\widehat{N_t} = \hat{c} + \hat{a}\, e^{\hat{b}t}$$

S2.3 Defining scientific fields

To compare hallucination rate across different areas of study, we first defined paper-level categories using labels assigned in the datasets. For arXiv, we used the primary category in which

the paper was listed as its field of study. For SSRN and PMC papers, we used the field name of the paper's "primary topic" provided by OpenAlex. Table S1 details how we aggregated the fields into higher-level macro research areas.

To avoid duplicate records of the same paper across platforms, for each macro research area, we focus on data from one representative source. For example, for the area "Computer Science," we only use data from arXiv, instead of combining all records under the same category in arXiv, SSRN, and PMC. Similarly, for life sciences, our analysis focuses on PMC, although including bioRxiv preprints yields qualitatively similar results.

S2.4 Estimating LLM use in scientific writing

We consider two complementary approaches to estimate the use of LLMs in scientific writing:

First, we use Pangram AI-text scoring, which is one of the most widely used approaches to AI-writing detection[3,4] (Fig. 1k-l). To obtain a per-paper measure of AI-assisted writing, we score a 10% random sample of arXiv abstracts using the Pangram v3 detection API. For accuracy, we only focus on abstracts with at least 200 characters. The Pangram response returns one or more sliding text windows, each with a word count and an AI assistance score between 0 and 1. If an abstract contains multiple windows, we consider a paper-level Pangram score as the word-count-weighted mean of the per-window AI assistance score. After excluding the small number of requests that did not return HTTP 200, the final scored sample comprises 127,182 arXiv abstracts between 2020 and 2025.

Notably, text-based AI detections are inherently imperfect. As a cross-validation, we consider another open-source approach developed by Liang et al.[5], and apply it to all arXiv abstracts in our data (Fig. S11). The central modelling assumption is that every document in the target corpus is drawn independently from a population where α is the fraction of LLM-modified documents:

$$(1-\alpha)P + \alpha Q$$

Here, P and Q represent the probability distributions of human-written and LLM-modified documents, respectively. The single scalar α is the quantity of interest. Given n documents $\{x_1, x_2, \ldots, x_n\}$ drawn independently from the population, the log-likelihood of the entire corpus as a function of α follows:

$$L(\alpha) = \sum_{i=1}^{n} log\big((1-\alpha)P(x_i) + \alpha\, Q(x_i)\big)$$

We leveraged the trained reference distributions in Kusumegi et al.[6], which is based on a subset of randomly selected arXiv papers (2,000 papers per month, January–October 2022). To estimate the LLM token distribution, we used the original abstracts as the human-written corpus, and the LLM-rewritten version, generated using GPT-3.5-turbo-0125, as the LLM-modified corpus. For each paper, we estimate the α parameter by maximizing $L(\alpha)$ on its title and abstract.

To quantify LLM use at the subfield level, we averaged the LLM use scores of papers within each subfield. Because the pre-LLM baseline also varies across subfields—reflecting differences in the inherent stylistic affinity between human writing and LLM output—we estimated a subfield-specific baseline using papers from 2020 to November 2022. We then subtracted this baseline from the post-LLM values to obtain the adjusted LLM use for field-level comparison.

S2.5 Analytical framework for characterizing hallucinated references

A central challenge in characterizing hallucinated references is that the set of unmatched references contains substantial false positives—items that are not hallucinated but instead reflect irrelevant bibliographic noise. To address this, we developed a population-level statistical framework that treats pre-LLM citation matching errors as a longitudinal baseline, allowing us to decompose the mixture of unmatched references and isolate hallucinated citations from this background noise.

Before the release of ChatGPT in November 2022, references fell into only two categories: those matched to paper records in existing databases and those left unmatched. We assume that unmatched references in this period were not LLM-generated hallucinations. After ChatGPT's release, a new category—LLM-hallucinated references—began to emerge alongside the pre-existing unmatched references, and the two became indistinguishable in the raw data, since both appeared simply as unmatched.

We define p_t and q_t as the probabilities that a given unmatched reference in t is an LLM hallucination and is not an LLM hallucination, respectively. Before the introduction of ChatGPT in November 2022, p_t remained 0 by definition. By definition, $p + q = 1$ for unmatched references, and $p = q = 0$ for matched references. Then, we estimate the following regression model:

$$y_i = \alpha + \gamma_{\text{category}} + \delta_t + \beta p_t + \theta q_t + \varepsilon_i$$

Here, β captures the difference between hallucinated references and matched references, net of category and time fixed effects.

In Section S2.2, we introduced the method for estimating ϵ_t, the hallucination rate of citations in a given period t. Here we estimate ϵ_t for each quarter using Q3 2022 as the reference period. We then calculate p_t:

$$p_t = \frac{\max\{0, \epsilon_t\}}{\max\{0, \epsilon_t\} + u_0}$$

where u_0 is the baseline unmatched-reference rate in Q3 2022. Values of p_i are bounded to the theoretical range [0, 1].

S2.6 Characterizing hallucination citers

For each paper with at least one unmatched reference, we randomly sampled one paper without unmatched references published in the same year, month and research area as the control group. Consistent with S2.3, for arXiv and bioRxiv papers, we rely on their own categorical systems; for SSRN and PMC papers, we used the paper's "primary topic" classified by OpenAlex. While these practices are consistent with common practice in the science of science literature[7], readers should keep in mind that these classifications may not be perfectly accurate at the individual paper level.[8]

To compare authors who cited hallucinated references against those who did not, we counted the number of papers each author had published and the citations they had received in two windows: (1) before 2023 and (2) in 2025 alone.

For arXiv authors, we constructed profiles by aggregating all papers associated with a given author name, after removing middle names to harmonize name variants. We then filtered to authors whose names appeared as full first and last names rather than initials only, since initial-only entries cannot be reliably disambiguated. Citations to these papers were retrieved from the Semantic Scholar database (version 2025-09-04). For bioRxiv, SSRN, and PMC authors, we constructed profiles using the author identifiers provided by OpenAlex and counted associated papers and citations from the OpenAlex database (legacy data accessed November 24, 2025).

We used the framework described in S2.5 to estimate the difference between the authors regarding the publication and citation counts in different time periods, separated by authorship position. We then exponentiated the estimated coefficients to obtain the ratio of paper or citation counts between authors who cited hallucinated references and those who did not.

S2.7 Estimating the rate of hallucinated references from bioRxiv to PMC

Among the 261,928 bioRxiv preprints we analyzed, 9,245 had at least one unmatched reference. We used the bioRxiv API to track whether these preprints were later published and to obtain their DOIs after publication in journals or conference proceedings. By December 2025, 3,992 of them had been published in other venues. We then used the PMC ID Converter API[9] to connect the DOIs of these published papers to PMC records. Of these, 2,241 were included in the PMC Open Access Subset under either the "commercial use allowed" or "non-commercial use only" license.

We next retrieved the corresponding PMC records and applied the same pipeline to examine their references. To determine whether an unmatched reference in the bioRxiv version was retained in the published version, we matched references across the two versions based on their normalized titles.

Using all references in the 2,241 paired bioRxiv–PMC papers, we applied the framework described in S2.5 to estimate the retention rate of hallucinated references. The outcome variable is binary, indicating whether an unmatched reference in bioRxiv persisted into the PMC version.

S2.8 Finding cited authors in hallucinated references

To assess the distribution of author names listed in references, we built a systematic linkage between raw author names and existing author profiles in Semantic Scholar. For each reference, we first extracted the names of the cited authors. To improve accuracy, we focused on names with full first and last names and excluded names containing only first initials. We also removed middle parts of each name for standardization, since such information is not always available in Semantic Scholar.

For each extracted name, we first retrieved all papers associated with the name in Semantic Scholar. Note that many of these candidate papers may have been contributed by different same-name authors – a common challenge in author disambiguation. To address this, we further filtered the list of papers based on their semantic distance to the focal paper. Specifically, we used paper-level embeddings calculated by SPECTER2, a scientific document embedding model based on title and abstract text. In the main text, we only consider papers whose Euclidean distance to the focal citing paper is smaller than 11.5. A cited author is considered as "identified" if they have at least one paper within this threshold. Robustness checks using different thresholds can be found in S3.5.

To construct a comparison group of likely human-generated references, for each unmatched reference, we randomly selected a matched reference cited by a paper published in the same month and subfield, with an identical total reference count and no unmatched references. We then applied the same procedure to their references to extract the cited authors and build their profiles. Fig. S4 plots the distribution of the number of papers linked to each author before and after embedding-based filtering, showing that the filter has effectively reduced the numbers to a more reasonable range. We then apply the regression method in S2.5 to compare the cited authors in matched and hallucinated references by estimating the following model:

$$y_i = \alpha + \gamma_{\text{category}} + \delta_t + \beta p_t + \theta q_t + \varepsilon_i$$

We consider seven outcomes in our analysis.

At the author level, we measure 1) whether a cited author could be identified. Then, among identifiable authors, we examined three author characteristics: 2) total publication count, 3) total citation count, and 4) gender.

To infer the gender of the authors from their name, we use an open-source name-based gender detector nomquamgender[10]. The tool builds on publicly available empirical name–gender associations from 36 distinct sources, covering more than 150 countries and over a century of observations. For each name, the tool provides a probability that a name is associated with a female-coded gender category instead of assigning a deterministic binary label. We binarize the probability score, labeling names with a female-gender probability below 0.5 as male-coded names.

We also measure outcomes at the team level. To avoid truncated author lists that, we excluded reference strings containing "et al." to avoid truncated author lists. We examine 5) the team size,

defined by the number of authors listed. For cases where both cited first and last authors can be linked to our dataset, we further estimate within-team hierarchy between first and last authors, by examining whether the last author has 6) a higher publication count or 7) a higher citation count than the first author.

S2.9 Screening and moderation of arXiv manuscripts

To probe the effectiveness of the moderation process on preprint servers, we analyzed a unique dataset of 30,981 rejected arXiv submissions. Unlike journals, arXiv does not conduct full peer review, and moderation decisions are not typically based on a comprehensive evaluation of the full text or reference list, if the submission is viewed at all. Instead, submissions may be rejected through multiple screening pathways. First, moderators may identify suspicious features from the title or abstract, then inspect the PDF to confirm whether the submission is appropriate for arXiv. Second, submissions may be flagged by automated screening tools using full-text and non-textual features, such as unusual bibliography patterns, bullet-point-like formatting, or authorship characteristics. These flagged submissions are then passed back to human moderators, who scrutinize the submission more carefully to make a final decision. For a broader discussion of arXiv's moderation model and recent challenges posed by AI-generated submissions, see the Data & Society article[11].

S2.10 Estimating journal impact

To quantify journal-level scientific impact in PMC, we use the journal hit rate: the probability that a paper published in a given journal reaches the 90th percentile of citations within its field and publication year[12]. For 41.7 million papers in PubMed and PMC, we map them to the SciSciNet-v2 database[13] which provided pre-calculated hit paper information. For each PubMed-indexed journal, we then calculate the proportion of its papers classified as hits.

S2.11 Measuring citation-only entries on Google Scholar

In Fig. 3c and Fig. S10, we estimate the prevalence of references that appear in Google Scholar only as citation records, i.e., entries that cannot be linked to a standalone indexed publication but are nevertheless indexed because they appear in the reference lists of other papers. To identify these cases, we revisit the validation pipeline described in S2.1 and focus on references that: (a) cannot be matched in Steps 1 and 2; (b) cannot be matched to a paper record under criterion (1) in Step 3; and (c) can still be found either as a Google Scholar citation record, or as an exact title string appearing in another Google Scholar-indexed paper, regardless of the number of mentions. We classify these references as citation-only entries and calculate their prevalence over time.

S3 Robustness checks

S3.1 Different thresholds for unmatched references

We varied the threshold for indel distance (used for title matching) and minimum citation count (used for paper filtering in Google Scholar search) and re-calculated the hallucination rate to examine how sensitive our results are to these different thresholds. Fig. S5 suggests the emergence of citation hallucinations is robust across various analytically plausible thresholds.

S3.2 Robustness check on SSRN data

In recent years, we observe a surge in non-social-science papers on SSRN, likely driven by the platform's partnerships with journals outside the social sciences (e.g., *The Lancet*). To confirm that our findings hold within the social sciences specifically, we restricted the SSRN sample to papers whose primary research topic falls within the "Social Science" domain in OpenAlex and re-estimated the hallucinated reference rate. Fig. S6 replicates Fig. 1c on this restricted sample, confirming the robust emergence of citation hallucination.

S3.3 Alternative specifications for characterizing hallucination citers

To test the robustness of the gap in academic achievements between hallucination citers and non-hallucination citers (Fig. 2a), we consider two alternative specifications: (1) in addition to examining the last author, we also examine the author with the most publications and all authors in the author list; (2) in addition to using the publication count as the outcome variable, we also use the total citation counts received before 2023 and in 2025. The results (Fig. S7) appear highly consistent, showing that the authors citing hallucinated references had relatively lower scientific output before 2023 compared with their counterparts, a gap that has largely narrowed in 2025.

Authors who published more papers have a higher chance to be selected into the control group for hallucination citers. To address this potential bias, we constructed an additional control group for each dataset by imposing a further restriction: for each unmatched reference, we counted the number of papers the last author published in the citing year and restricted the search for control group authors to those with the same publication count in that year, with all other criteria held constant. Figure S8 presents the results, which appear to be consistent.

S3.4 Comparing different reference extraction methods

We randomly selected 2,201 arXiv papers for which both the .bbl files and PDF files were available and evaluated three reference extraction methods. The first method used GROBID to extract references from the .bbl files. The second converted the PDF files into plain text using the open-source tool pdftotext-plus-plus (https://github.com/ad-freiburg/pdftotext-plus-plus), which was also used for arXiv rejected submissions, and then extracted references from the converted text.

The third method used GROBID to extract references directly from the PDF files. The three methods yielded 111,911, 107,241, and 112,353 references, respectively.

Within each paper, we compared the reference sets returned by each pair of methods by computing the sizes of their intersection and union and then calculating the Jaccard similarity. To account for minor discrepancies in reference strings arising from differences in the input files or extraction procedures, we used fuzzy matching and treated two reference strings as the same if their token set ratio, computed using the Python library RapidFuzz, is at least 90. Table S2 shows high pairwise agreement across all three methods.

S3.5 Different thresholds for finding cited authors

To test whether our results in Fig. 2c are sensitive to the semantic-distance thresholds for identifying relevant authors, we varied the threshold from 11.5 to 11 and 12, respectively, and replicated our analysis. The results displayed in Fig. S9 suggest high consistency across the thresholds.

S4 Manual validation of unmatched references

To test the validity of our pipeline for detecting hallucinated references, we examined the unmatched references identified by our algorithm and manually assessed whether they are hallucinated. We do not expect all unmatched references to be hallucinated; rather, the fraction of hallucinated references among the unmatched should approximate the p_t derived in Section S2.5.

For each dataset, we randomly selected 100 unmatched references in Q3 2025. For PMC, we excluded references whose cited venue appeared fewer than 10 times before 2023 before sampling (see S2.2). Manual verification was performed by a team of five graduate students and three undergraduate students, who were instructed to search the internet—without using LLM-based tools—to determine whether each reference pointed to a paper that does not exist.

We report 95% Wilson score confidence intervals for the proportion of hallucinated references among all manually reviewed cases. Our results are broadly consistent with the theoretical estimates from S2.5 (Fig. S12).

S5 Supplementary Tables

Table S1. Mapping of dataset-specific field labels to macro research areas. Dataset-specific field labels from arXiv, SSRN/OpenAlex, and PMC/OpenAlex were harmonized into seven macro research areas used for cross-field comparisons in the main analysis.

	arXiv	SSRN	PMC
Social, Economic & Decision Sciences		Business, Management and Accounting, Economics, Econometrics and Finance, Decision Sciences, Social Sciences	
Computer Science	cs, eess		
Medical Science			Medicine, Dentistry, Health Professions, Neuroscience, Psychology
Mathematics	math, stat		
Biological Science			Biochemistry, Genetics and Molecular Biology, Immunology and Microbiology, Chemistry
Environmental Science			Agricultural and Biological Sciences, Environmental Science, Earth and Planetary Sciences
Physics	astro-ph, cond-mat, gr-qc, hep-ex, hep-lat, hep-ph, hep-th, math-ph, nlin, nucl-ex, nucl-th, physics, quant-ph		

Table S2. Pairwise agreement between reference extraction and parsing methods.
Intersection size, union size, and Jaccard similarity for each pair of methods. The three methods show high agreement, with Jaccard similarities ranging from 0.854 to 0.935, indicating that the choice of extraction method does not materially affect the set of references identified.

	Intersection size	Union size	Jaccard similarity
Method 1 & 2	100,875	118,125	0.854
Method 1 & 3	108,139	115,643	0.935
Method 2 & 3	101,172	118,010	0.857

S6 Supplementary Figures

Reference classification

Decide whether the following reference string refers to a scholarly work, including a research paper from academic sources (journals, conferences, scholarly publishers, institutional repositories, and pre-print platforms like arXiv, SSRN, etc.), and a monograph or chapter from scholarly publishers. Non-academic materials include but are not limited to news articles, blogs, magazine articles, theses, dissertations, government/industry reports, software documents, white papers, datasets, tutorials, etc.
Output one of the following labels with no explanation:

- Academic
- Non-Academic

The reference string is: {reference}

Title extraction

Extract the title of the referenced publication (can be a journal article, conference paper, or a book or chapter) from a string in the reference list of an academic paper.
– If the string clearly corresponds to a bibliographic reference (e.g., includes author names, publication year, journal name, or other citation elements), extract the title of the referenced work.
– If the string appears to be a reference but the title is not available, return "TITLENOTAVAILABLE".
– If the string is not a reference at all (i.e., it is explanatory text, inline commentary, or any content not meant to cite a publication), return "NON-REF".
Examples:
Input: Smith J, 2015. An Introduction to Quantum Mechanics. Cambridge University Press.
Output: An Introduction to Quantum Mechanics

Input: Further discussions of this concept can be found in various statistical physics textbooks.
Output: NON-REF

Input: Doe J, 2007. Nature.
Output: TITLENOTAVAILABLE

The string is: {ref}
The title is:

Figure S1. Prompt templates for classifying the citations and extracting the cited title. Reference classification is performed first; references classified as "Academic" are then queried again to extract the cited title.

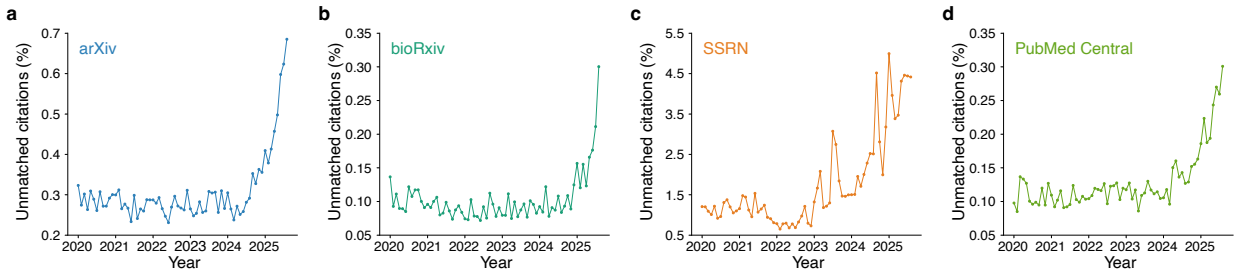

Figure S2. Monthly proportion of unmatched references. Trends across (**a**) arXiv, (**b**) bioRxiv, (**c**) SSRN, and (**d**) PubMed Central show a steep rise beginning in mid-2024, roughly 18 months after the initial release of ChatGPT in late 2022.

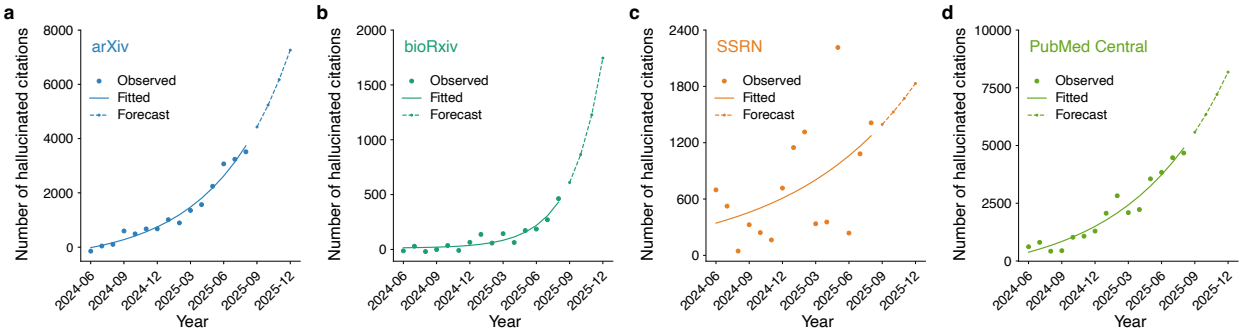

Figure S3. Fitting and projection of hallucinated references over time. Observed monthly counts, fitted exponential growth curves, and projected counts of hallucinated citations are shown for (a) arXiv, (b) bioRxiv, (c) SSRN, and (d) PubMed Central. Forecasts are used to project the final months of 2025, yielding an estimated annual total of 146,932 hallucinated citations across the four datasets.

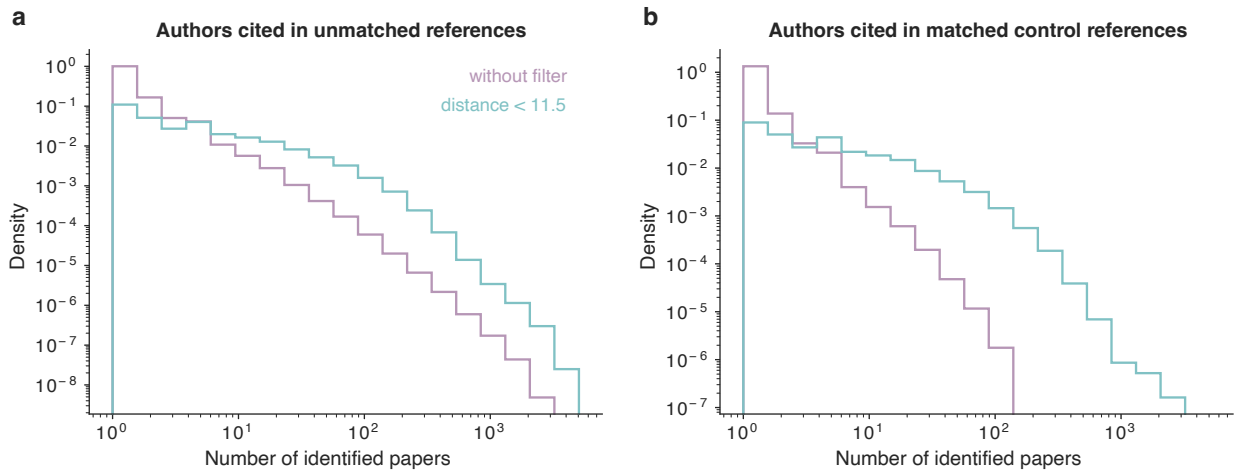

Figure S4. Effect of semantic-distance filtering on author paper counts. Distributions of the number of Semantic Scholar papers retrieved for each cited author name are shown before and after semantic-distance filtering, separately for (a) authors listed in unmatched references and (b) authors listed in matched control references. Filtering reduces implausibly large candidate sets, improving the precision of author-profile linkage.

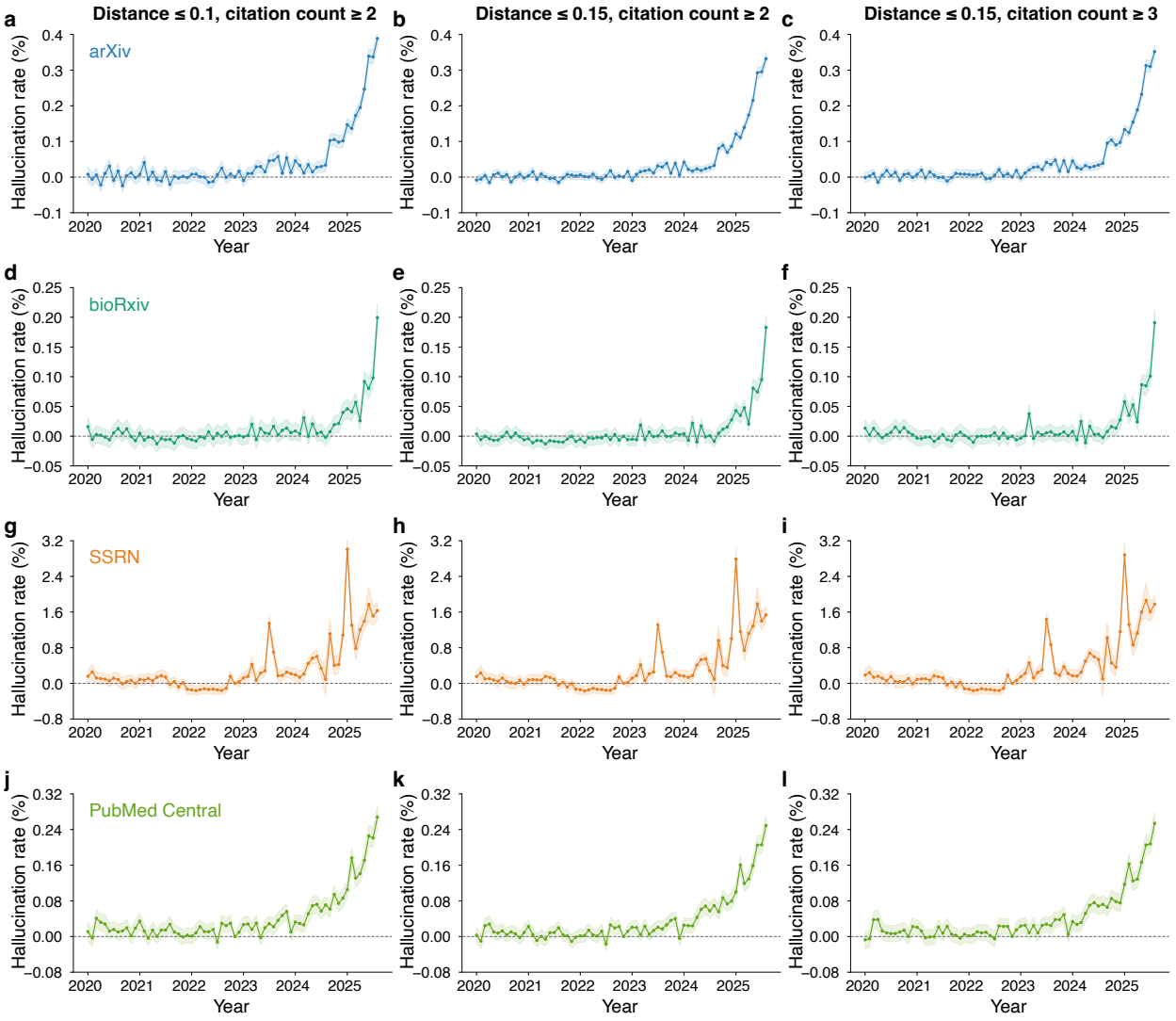

Figure S5. Robustness of the estimated hallucination rate to alternative indel-distance and citation-count thresholds. Hallucination rate over time across (**a**–**c**) arXiv, (**d**–**f**) bioRxiv, (**g**–**i**) SSRN, and (**j**–**l**) PubMed Central, under three threshold combinations: (**a**, **d**, **g**, **j**) relative indel distance ≤ 0.1 and citation count ≥ 2; (**b**, **e**, **h**, **k**) relative indel distance ≤ 0.15 and citation count ≥ 2; and (**c**, **f**, **i**, **l**) relative indel distance ≤ 0.15 and citation count ≥ 3. The rise in hallucination rate is consistent across all threshold choices, indicating that the emergence of citation hallucination is not an artifact of the specific cutoffs used in the main analysis.

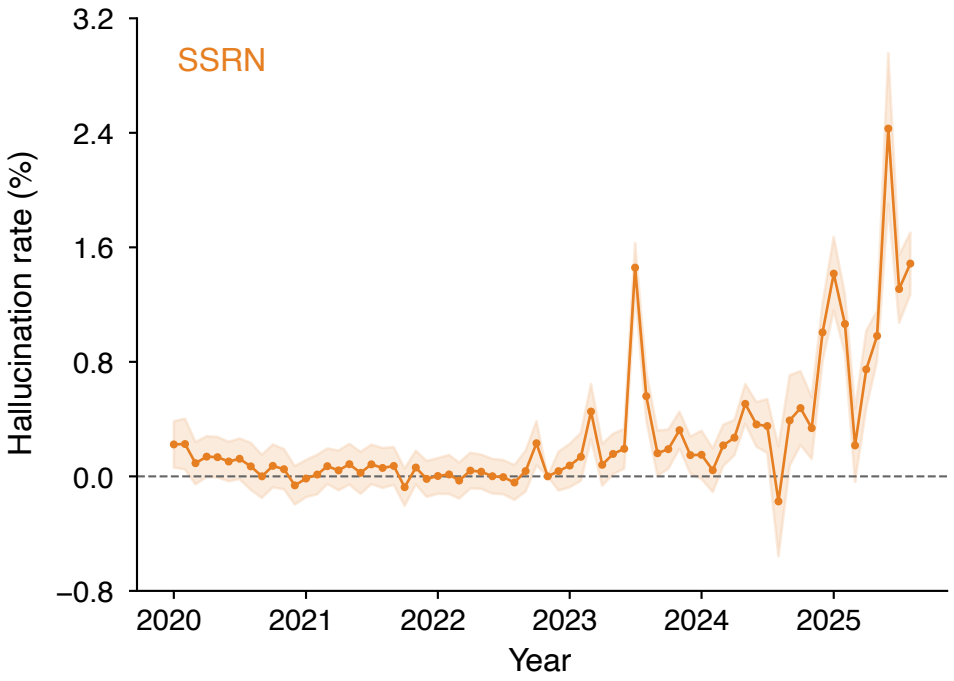

Figure S6. Hallucination rate in SSRN restricted to social science papers. Estimated hallucination rate over time after excluding SSRN papers whose primary research topic in OpenAlex falls outside the "Social Sciences" domain. The pattern replicates Fig. 1c, confirming that the emergence of citation hallucination is not driven by the recent influx of non-social-science papers on the platform.

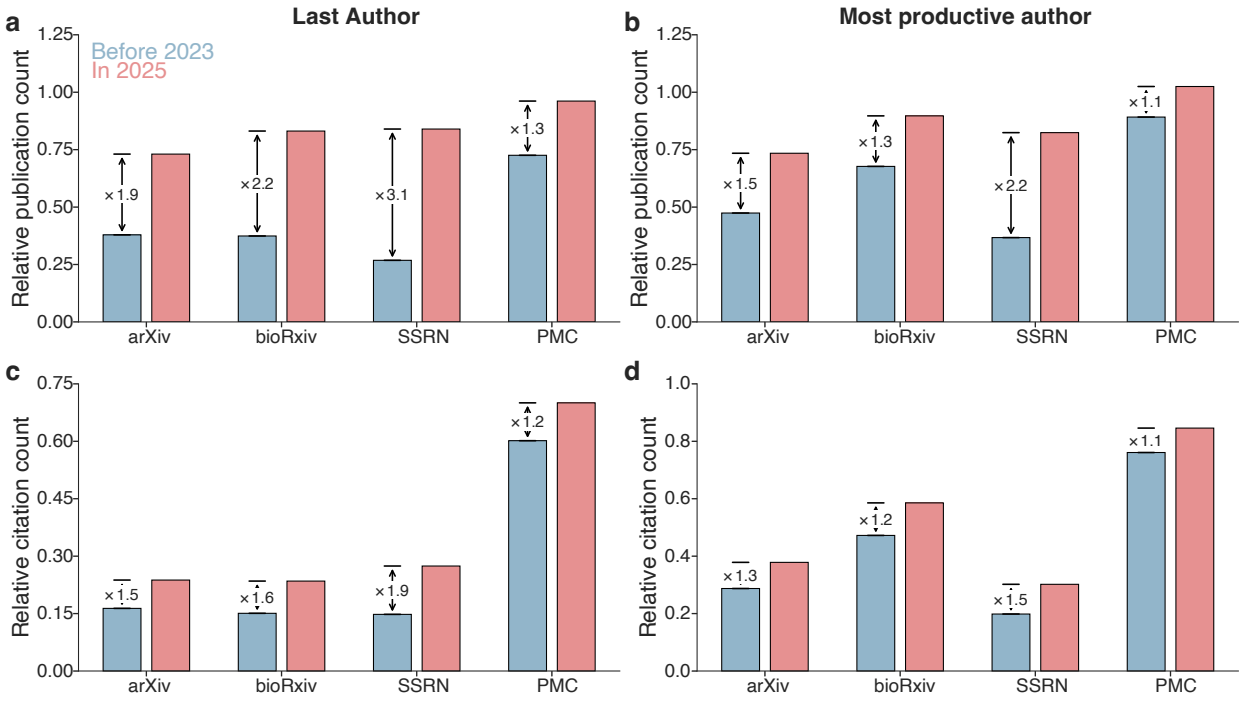

Figure S7. Robustness of the productivity gap to alternative authorship definitions and impact measures. Comparison of hallucination citers against control authors using: (**a**) the original specification reported in the main text; (**b**) publication count of the most prolific author in each paper's author list; (**c**) citation count of the last author; and (**d**) citation count of the most prolific author. Across all specifications, hallucination citers are on average less established than control authors, though this gap narrows in 2025.

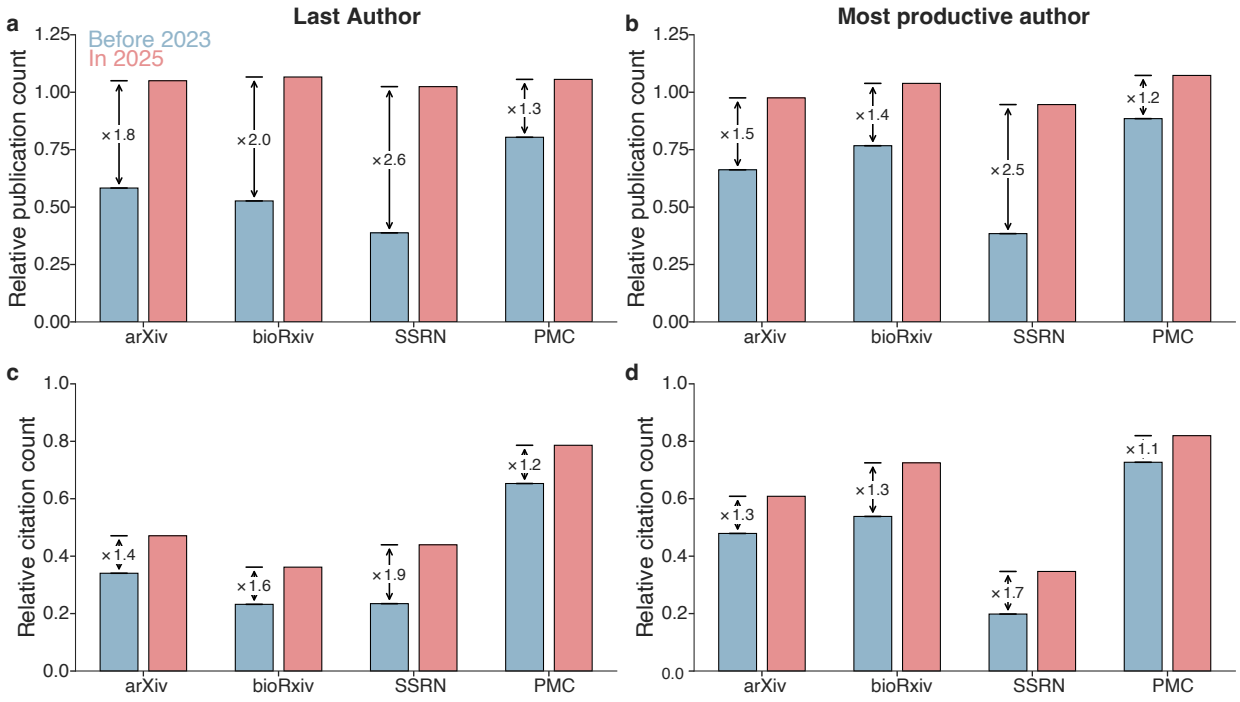

Figure S8. Robustness of the productivity gap using a publication-count-matched control group. Replication of Fig. S7 using an alternative control group, in which control authors are matched to hallucination citers on publication count. Panels (**a**–**d**) follow the same specifications as in Fig. S7. Results are consistent with the main analysis: hallucination citers are on average less established than control authors, with the gap narrowing in 2025.

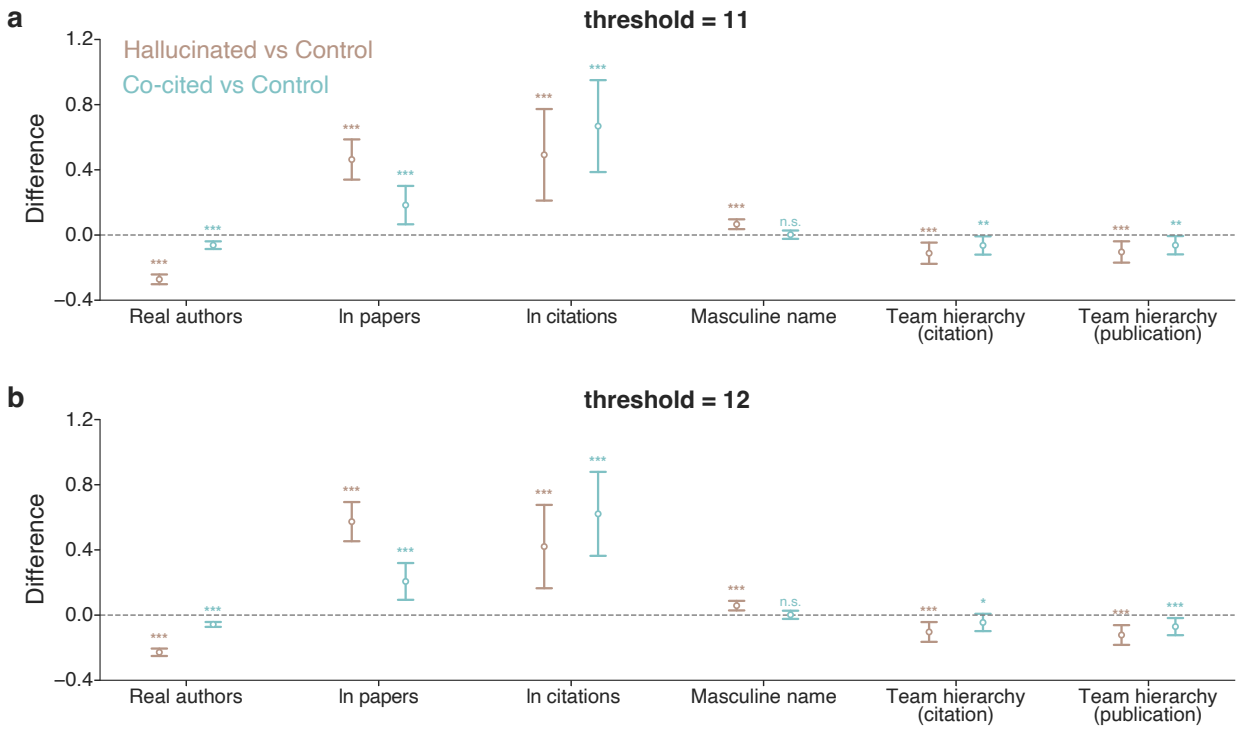

Figure S9. Robustness of cited-author comparisons to alternative matching thresholds. Replication of Fig. 2c using matching thresholds of (**a**) 11 and (**b**) 12. Results are qualitatively consistent with the main analysis, indicating that the observed differences are not sensitive to the chosen linkage threshold.

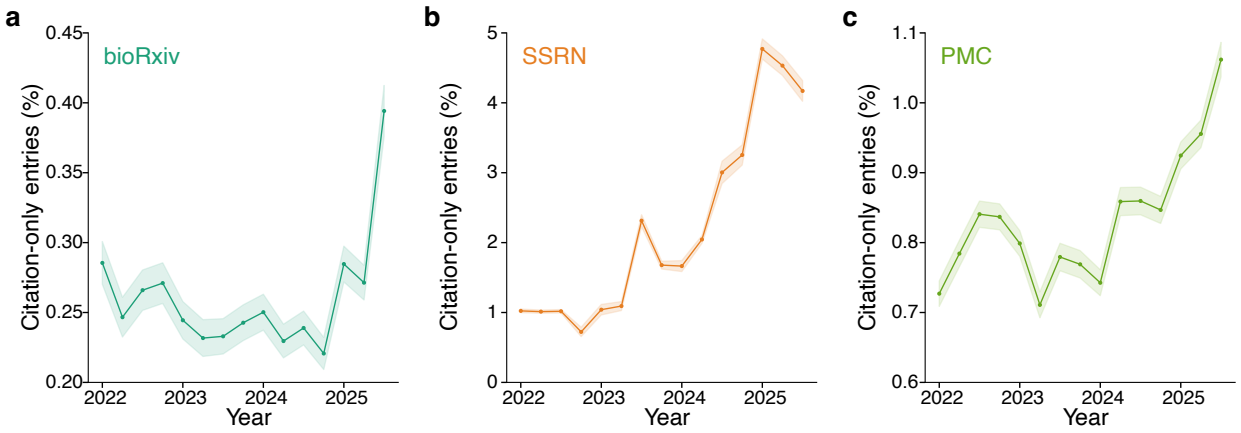

Figure S10. Emergence of unmatched citation-only entries across datasets. Prevalence of cited papers that cannot be matched to any real publication yet already appear as references in other papers, shown for (**a**) bioRxiv, (**b**) SSRN, and (**c**) PubMed Central. All three datasets show a steep increase beginning in mid-2024, replicating the pattern reported for arXiv in Fig. 3c.

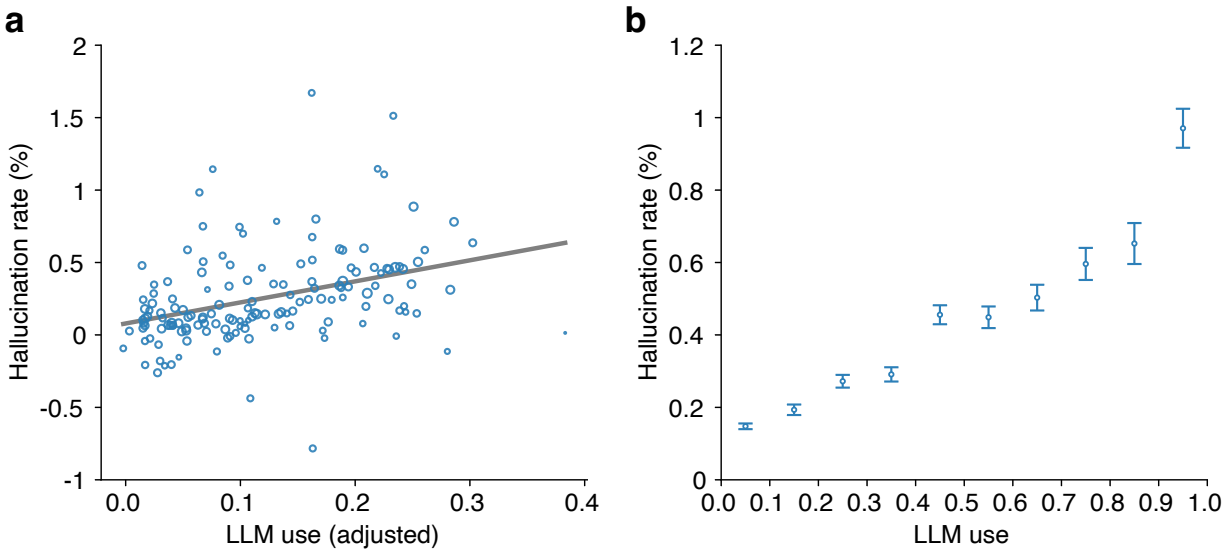

Figure S11. Association between citation hallucination and estimated LLM use, using an alternative detection method. Replication of Fig. 1k,l using the method developed by Liang et al.[5] (**a**) Subfield-level association: each point represents an arXiv subfield, and the positive correlation between subfield-level hallucination rate and estimated LLM use is consistent with the main analysis ($r = 0.363$, $P < 0.001$). (**b**) Paper-level association: higher estimated LLM use is associated with higher citation hallucination, consistent with the main analysis.

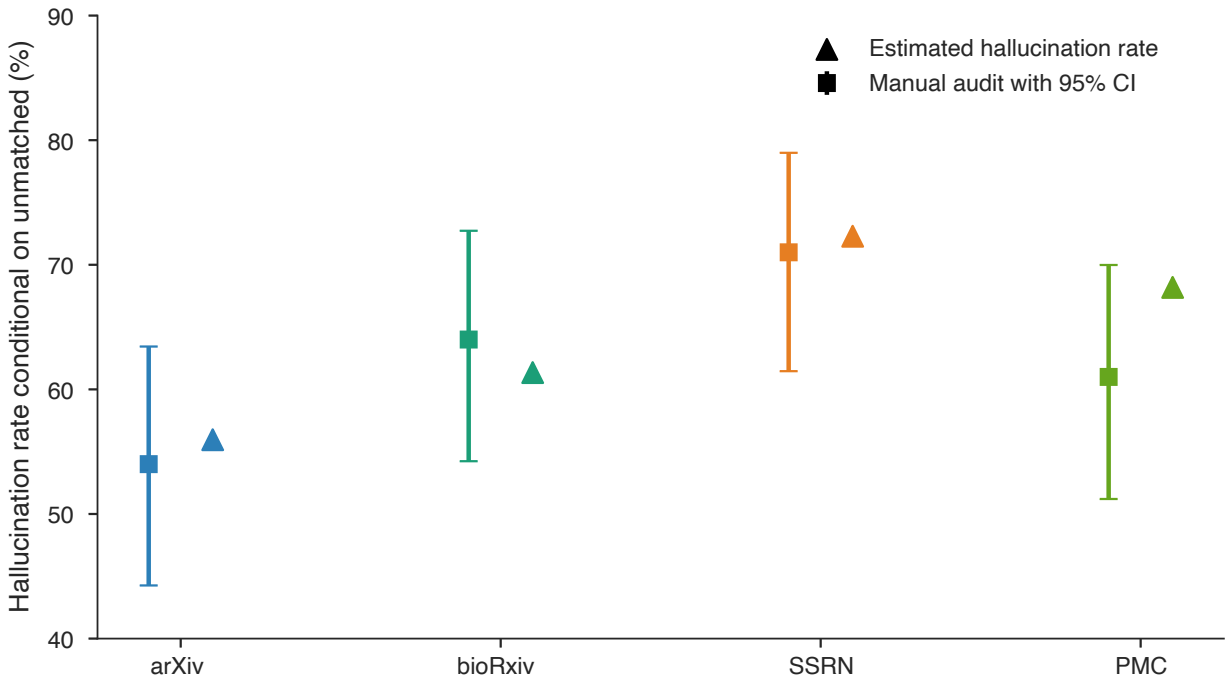

Figure S12. Validation of estimated hallucination rates against manual audit. Hallucination rate conditional on a reference being unmatched, for arXiv, bioRxiv, SSRN, and PubMed Central. Triangles show rates estimated from the data; squares show rates from manual audit, with error bars indicating 95% confidence intervals. Estimated rates fall within the 95% confidence interval of the manual audit in all four datasets.

References

1. GROBID. Introduction - GROBID Documentation. https://grobid.readthedocs.io/en/latest/Introduction/.
2. National Library of Medicine. PubMed Central: NIH preprint pilot. *PubMed Central (PMC)* https://pmc.ncbi.nlm.nih.gov/about/nihpreprints/ (2026).
3. Jabarian, B. & Imas, A. Artificial writing and automated detection. SSRN at https://doi.org/10.2139/ssrn.5407424 (2025).
4. Gartenberg, C., Hasan, S., Murray, A. & Pierce, L. More versus better: artificial intelligence, incentives, and the emerging crisis in peer review. *Organ. Sci.* https://doi.org/10.1287/orsc.2026.ed.v37.n3 (2026).
5. Liang, W. *et al.* Mapping the increasing use of LLMs in scientific papers. Preprint at https://doi.org/10.48550/arXiv.2404.01268 (2024).
6. Kusumegi, K. *et al.* Scientific production in the era of large language models. *Science* **390**, 1240–1243 (2025).
7. Wang, D. & Barabási, A.-L. *The Science of Science*. (Cambridge University Press, 2021).
8. Haunschild, R. & Bornmann, L. The use of OpenAlex to produce meaningful bibliometric global overlay maps of science on the individual, institutional, and national levels. *PLOS ONE* **19**, e0308041 (2024).
9. National Library of Medicine. PubMed Central: PMC ID converter API. *PubMed Central (PMC)* https://pmc.ncbi.nlm.nih.gov/tools/id-converter-api/ (2025).
10. Van Buskirk, I., Clauset, A. & Larremore, D. B. An open-source cultural consensus approach to name-based gender classification. Preprint at https://doi.org/10.48550/arXiv.2208.01714 (2022).

11. Singh, R. On arXiv, an Influx of AI slop pits surface against substance. *Data & Society* https://datasociety.net/points/on-arxiv-an-influx-of-ai-slop-pits-surface-against-substance/ (2025).

12. Hill, R. *et al.* The pivot penalty in research. *Nature* **642**, 999–1006 (2025).

13. Lin, Z., Yin, Y., Liu, L. & Wang, D. SciSciNet: A large-scale open data lake for the science of science research. *Sci. Data* **10**, 315 (2023).