Patterns of Gender Diversity and Funding Allocation in Post-2022 Publications on LLMs and ChatGPT

Moses Boudourides[a] and Evan Piepho[b]

[a] Former Professor of Applied Mathematics, University of Patras, Greece

[a] Adjunct Lecturer, School of Professional Studies, Northwestern University

[b] Graduate Student, School of Public Affairs, Arizona State University

October 31, 2024

Abstract

This study leverages data extracted from the Dimensions.ai database to investigate diversity and allocation patterns in post-2020 academic publications focused on Large Language Models (LLMs) and ChatGPT. It addresses key questions on gender representation, interdisciplinary collaboration, and funding dynamics within this rapidly evolving field. Using the Namsor app to infer author gender, the study analyzes gender disparities and inclusivity in LLM and ChatGPT research. It also examines research areas as interdisciplinary intersections, identifying emerging trends and their implications for innovation. The study explores grant support, funding sources, and potential disparities in allocation, providing insights into the financial ecosystems driving advancements.

Since this study is part of an ongoing project, the following is a brief report on a portion of our findings that we had completed before the end of 2024.

Introduction

This paper presents a bibliometric analysis of scholarly contributions focusing on the research productivity and impact of publications related to Large Language Models (LLMs) and ChatGPT post-2022. The analysis is conducted using data extracted from the Dimensions.ai database, which provides comprehensive metadata and programmatic access via its Python client, `dimcli`.

For an overview of the methodology behind the construction of the Dimensions database and user interface, see [8].[1] For key literature on bibliometrics as a tool for research evaluation, see [3], [7] and [2].

[1] One might wonder, “Why choose Dimensions over Scopus, Web of Science, or another bibliometric database?” The primary reasons for me in selecting Dimensions are twofold: (i) it provides a high degree of completeness and quality in publication metadata ([4], [11]), and (ii) it offers a Python client, `dimcli`, which facilitates programmatic access to the Dimensions API for efficient querying [14].

Data Collection and Methodology

The dataset was compiled using a Dimensions query for publications containing the terms "ChatGPT," "Large Language Model," or "LLM" in their titles or abstracts, published between 2022 and 2025. After removing duplicate publications, the query returned 47,855 publications with 20 fields of publication and grant metadata, of which the following 14 were analyzed in the present study: `id`, `authors`, `title`, `date`, `doi`, `type`, `journal`, `category_for`, `concept`, `concept_scores`, `journal`, `times_cited`, `supporting_grant_ids`, and `funding_usd`. Most of these fields are self-explanatory (documentation for all the Dimensions.ai fields can be found in [6]). Specifically, the following are descriptions of some lesser-recognizable fields:

- `type`: This is the type of a publication, such asn `article`, `proceeding`, `chapter`, `preprint` etc.

- `category_for`: In Dimensions, it pertains to the Field of Research (FoR) classification system. This field categorizes various academic outputs—including research publications, grants and patents—into standardized fields of study, enabling clearer organization and analysis of research data ([16]). This classification aligns with the Australian and New Zealand Standard Research Classification (ANZSRC), which arranges research outputs into a hierarchical structure, where major fields are subdivided into more specific minor fields ([15], [1]).

- `concept` and `concept_scores`: In Dimensions, the `concept` field represents normalized noun phrases that encapsulate the primary topics of a publication. These phrases are automatically extracted from publication abstracts using machine learning techniques. The relevance of each `concept` is ranked through the `concept_scores` field, which ranges from 0 to 1 ([17]). Scores closer to 1 indicate that the `concepts` are highly relevant to the subject matter of the publication, while scores approaching 0 suggest a lack of relevance to the publication topics.

- `supporting_grant_ids` and `funding_usd`: These fields correspond to grants supporting a publication, returned as a list of Dimensions grants IDs, and to the amount of funding, in U.S. dollars (USD), awarded to a publication through a grant or research project. Dimensions normalizes funding data from different funders to USD to facilitate comparisons across countries and funding bodies ([5]).

Top 10 Most Cited Publications

As one may observe in Figure 1, the key trends and characteristics observed in these publications are the following:

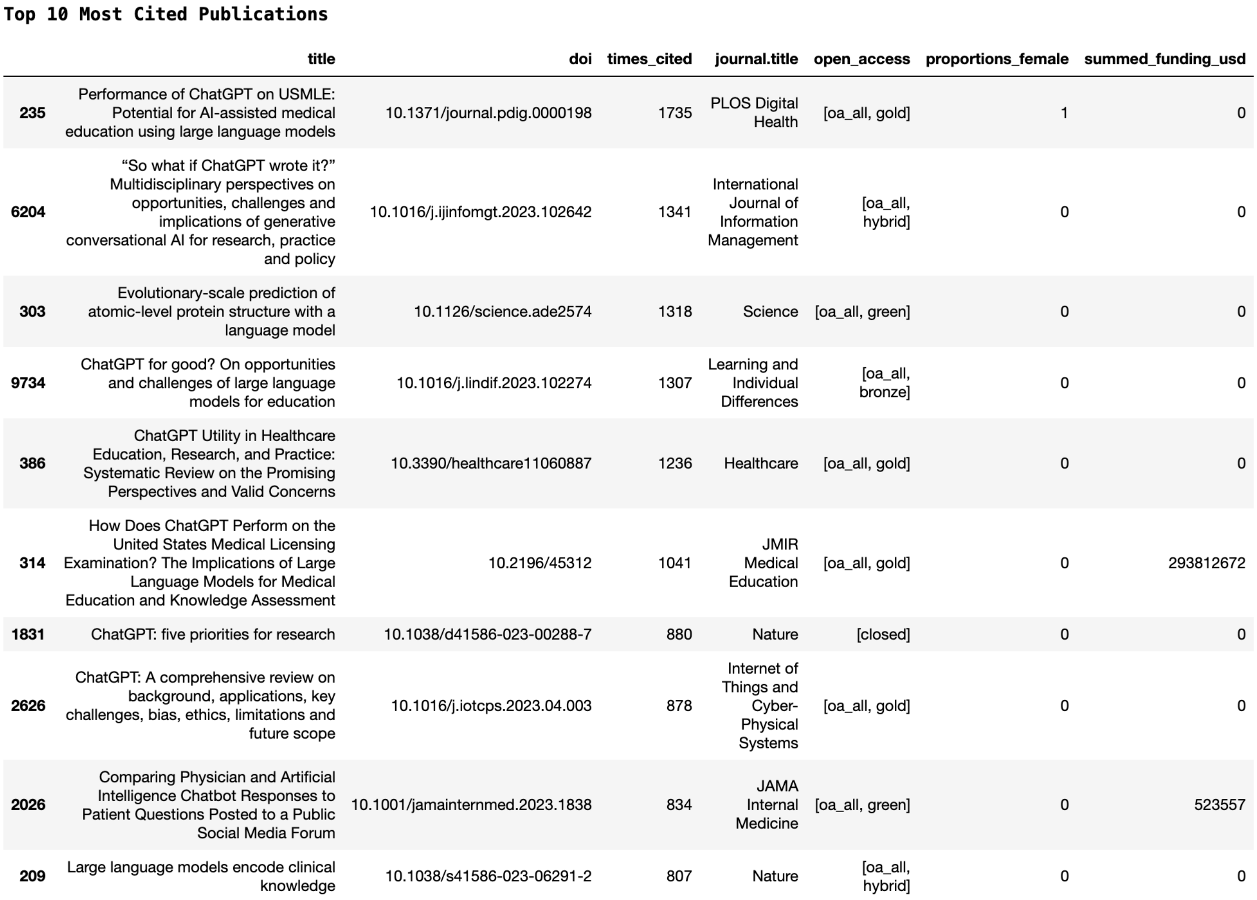
- **High Impact on Medical Education and Healthcare**: Several of the top-cited publications focus on the application of ChatGPT and LLMs in medical education and healthcare. For example, the most cited paper, *"Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models"* (cited 1,735 times), explores ChatGPT's potential in medical training. Similarly, *"How Does ChatGPT Perform on the United States Medical Licensing Examination?"* (cited 1,041 times) and *"ChatGPT Utility in Healthcare Education, Research, and Practice"* (cited 1,236 times) underscore the significant interest in AI's role in healthcare and education.

- **Broad Multidisciplinary Interest**: The second most cited paper, *"So what if ChatGPT wrote it?"* (cited 1,341 times), takes a multidisciplinary perspective, discussing the opportunities and challenges of generative AI across research, practice, and policy. This reflects the widespread relevance of ChatGPT beyond specific fields.
- **Scientific and Technical Applications**: Publications like *"Evolutionary-scale prediction of atomic-level protein structure with a language model"* (cited 1,318 times) demonstrate the use of LLMs in advanced scientific research, particularly in protein structure prediction, highlighting their utility in cutting-edge scientific domains.
- **Educational Focus**: Several papers, such as *"ChatGPT for good? On opportunities and challenges of large language models for education"* (cited 1,307 times), focus on the implications of ChatGPT for education, reflecting the growing interest in AI's role in transforming learning and knowledge assessment.
- **Open Access Dominance**: Most of the top publications are open access, indicating a strong preference for freely accessible research in this field. For instance, *"Performance of ChatGPT on USMLE"* and *"ChatGPT Utility in Healthcare Education, Research, and Practice"* are both gold open access, ensuring wide dissemination.
- **Limited Gender Diversity**: The *proportions_female* column shows that none of the top-cited publications have a majority of female authors, with most having no female representation. This highlights a gender disparity in authorship within this research domain.
- **Varied Funding Levels**: While some publications, like *"How Does ChatGPT Perform on the United States Medical Licensing Examination?"*, received significant funding (over $293 million), others have no reported funding. This suggests variability in financial support for research in this area.
- **High-Impact Journals**: Many of these publications appear in prestigious journals such as *Nature*, *Science*, *JAMA Internal Medicine*, and *PLOS Digital Health*, indicating the high academic and scientific impact of this research.

In summary, the top 10 publications in our dataset reflect a strong focus on the applications of ChatGPT and LLMs in healthcare, education, and scientific research, with a notable emphasis on open access and high-impact journals. However, across these top 10 publications, gender diversity in authorship remains limited, and funding levels vary widely among studies.

Publication Trends Over Time

The time evolution of the volume of publications in our dataset, as depicted in Figures 2 and 3, shows a clear trend of increasing research output over time. In particular, the



Top 10 Most Cited Publications

	title	doi	times_cited	journal.title	open_access	proportions_female	summed_funding_usd
235	Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models	10.1371/journal.pdig.0000198	1735	PLOS Digital Health	[oa_all, gold]	1	0
6204	“So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy	10.1016/j.ijinfomgt.2023.102642	1341	International Journal of Information Management	[oa_all, hybrid]	0	0
303	Evolutionary-scale prediction of atomic-level protein structure with a language model	10.1126/science.ade2574	1318	Science	[oa_all, green]	0	0
9734	ChatGPT for good? On opportunities and challenges of large language models for education	10.1016/j.lindif.2023.102274	1307	Learning and Individual Differences	[oa_all, bronze]	0	0
386	ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns	10.3390/healthcare11060887	1236	Healthcare	[oa_all, gold]	0	0
314	How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment	10.2196/45312	1041	JMIR Medical Education	[oa_all, gold]	0	293812672
1831	ChatGPT: five priorities for research	10.1038/d41586-023-00288-7	880	Nature	[closed]	0	0
2626	ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope	10.1016/j.iotcps.2023.04.003	878	Internet of Things and Cyber-Physical Systems	[oa_all, gold]	0	0
2026	Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum	10.1001/jamainternmed.2023.1838	834	JAMA Internal Medicine	[oa_all, green]	0	523557
209	Large language models encode clinical knowledge	10.1038/s41586-023-06291-2	807	Nature	[oa_all, hybrid]	0	0

Figure 1: Top 10 most cited publications.

monthly evolution of publications indicates a steady increase in the number of publications from July 2022 to January 2025. The growth appears to accelerate significantly starting around January 2023, coinciding with the widespread adoption and public release of ChatGPT in late 2022. This suggests a surge in research interest following the tool’s availability. By July 2023, the number of publications continues to rise, reflecting sustained academic and industrial interest in LLMs and their applications. The upward trend continues through the first eight months of 2024 in our surveyed data (examined in the present report) and extends further, as indicated by our complementary dataset covering the entire year. Overall, our ongoing bibliometric analysis strongly demonstrates that research in this area remains highly active and is likely to keep growing.

Furthermore, the weekly evolution of the volume of publication provides a more granular view of the research output, showing fluctuations in the number of publications on a weekly basis. Similar to the monthly trend, the graph reveals a noticeable increase in publications starting around January 2023, with occasional spikes that may correspond to major conferences, breakthroughs, or events related to LLMs and ChatGPT. The overall trend is again one of growth, with the number of publications per week increasing over time, particularly from July 2023 onward. By January 2025, the weekly publication rate appears to stabilize at a higher level compared to earlier years, suggesting that research in this field appears to be reaching a mature but highly active phase.

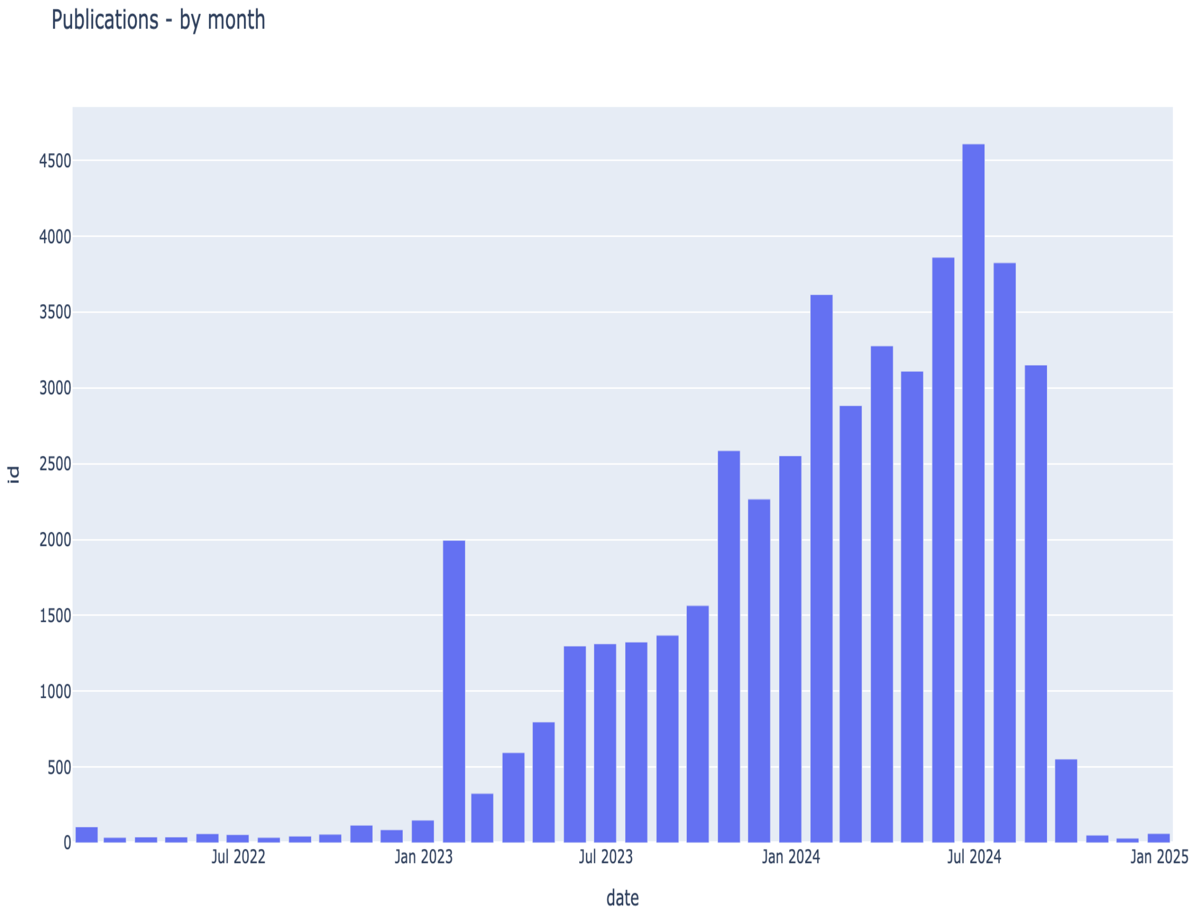


Figure 2: Number of publications per month.

Gender of Authors

The gender of authors was identified using the Namsor algorithm ([10]), which is based on a repository of 7.5 billion names, including those from 142 ethnicities, 249 countries, and 22 alphabets ([9]). Namsor's model recognizes morphemes—the smallest units of construction within languages that help comprise words—to incorporate patterns in naming conventions when assigning a name's gender, ethnic origin, and other elements offered through their service. The accuracy of Namsor's model has been verified by multiple studies and audits, including a 2018 Science-Metrix publication that found it correctly classified the gender of Olympic medalists' names from 25 countries to within 98-99% accuracy ([12]). As such, Namsor is used frequently within academic and international institutions—particularly in the context of examining gender disparities.

Paul Sebo ([13]) has suggested practical strategies and methodological insights to enhance the accuracy of gender inference in data science, particularly in research contexts where gender is inferred from first names. As Sebo's article highlights, we also encountered significant challenges—such as removing initials, diacritics, and accents from names, as well

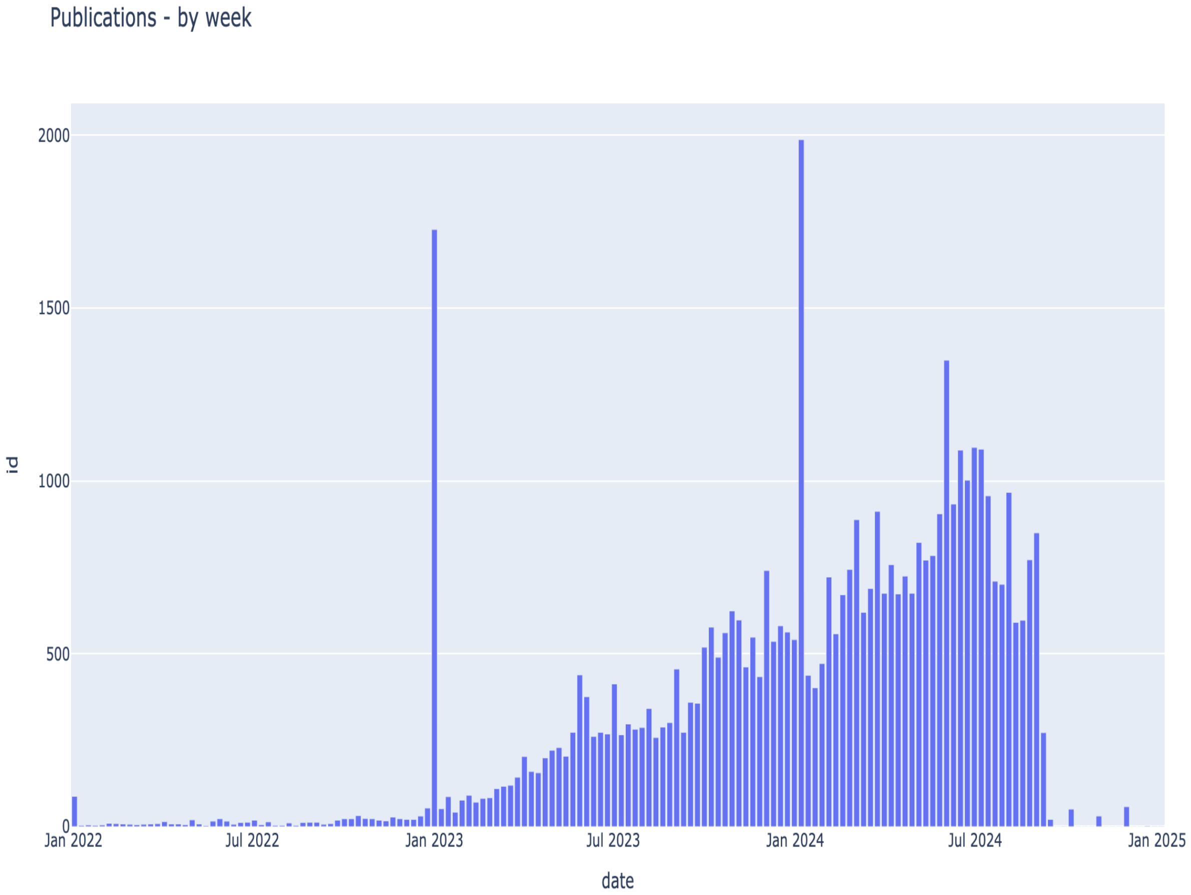


Figure 3: Number of publications per week.

as reversing first-last name order in certain languages—to address biases and limitations in automated gender inference and improve the reliability of Namsor's analyses.

To measure the gender composition of authors in our dataset, we are using two indices: the proportion of female and male authors within the group of (co-)authors for each publication. Both indices range from 0 to 1 and their histograms and distributions (density functions) are presented in 4.

The histogram for female authors peaks near 0, indicating that many publications have few or no female authors. In contrast, the histogram for male authors peaks near 1, reflecting a predominance of male authors. Both histograms show a gradual decline as the proportions move away from their peaks, with fewer publications exhibiting balanced gender representation. The density function for female authors is skewed toward 0, confirming their underrepresentation, while the male authors' density function is skewed toward 1, highlighting their dominance. The minimal overlap between the two density functions underscores the significant gender disparity in authorship. Overall, the data reveals a clear gender imbalance, with male authors overrepresented and female authors underrepresented in most publications. Publications with balanced gender proportions (around 0.5 for both indices) are relatively rare, as seen in the low counts and density in the middle range of the histograms.

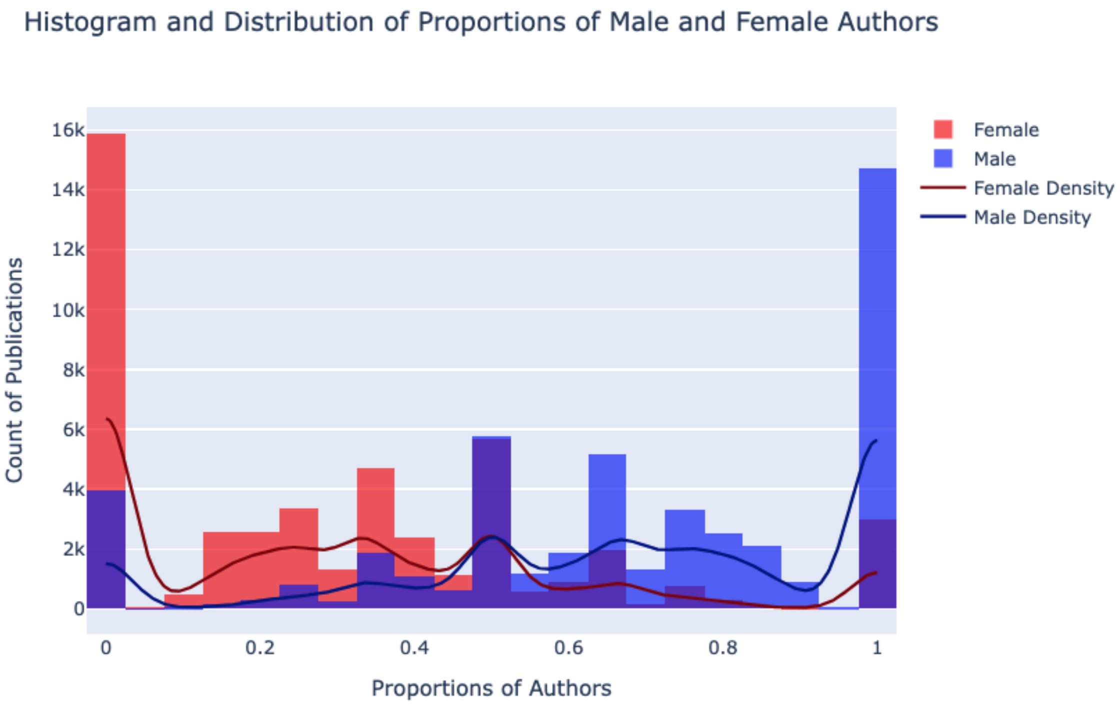


Figure 4: Gender distribution of authors.

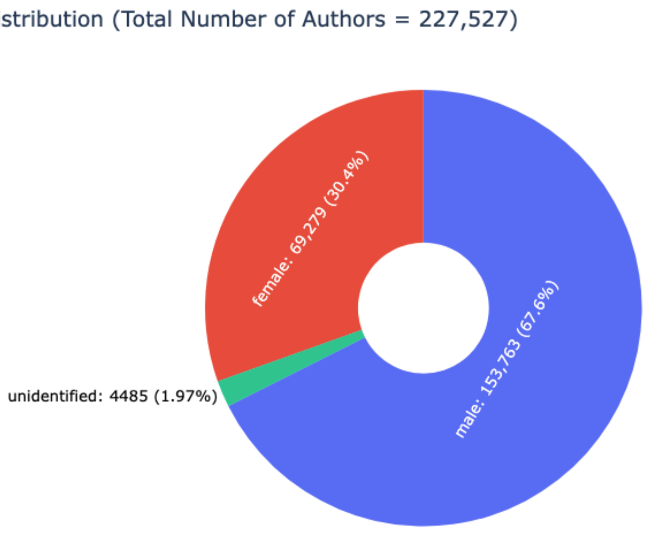


Figure 5: Gender distribution of authors.

The distribution of author gender in the publications of our dataset, shown in Figure 5, reveals the following composition:

- **Male Authors**: The majority of authors are male, representing 67.58% (153,763 out of 227,527) of the total distribution.

- **Female Authors**: Female authors account for 30.45% (69,279 out of 227,527) of the total, indicating a significant gender disparity in authorship.
- **Authors with Unidentified Gender**: A small proportion of authors (1.97%, or 4,485 out of 227,527) could not be identified in terms of gender.

In summary, the data highlights a significant overrepresentation of male authors in the publications, with female authors comprising about one-third of the total. The presence of authors with unidentified gender also suggests challenges in accurately inferring gender in some cases.

Next, we are examining the distribution of publications by gender composition over time, shown in Figure 6. The categories in the gender composition are male, female, mixed-gender teams (female, male), and combinations with authors with unidentified gender. The time evolution of weekly composition of author genders in publications, depicted in Figure 6, reveals the following trends: Publications with male authors dominate throughout the period, but there is a growing presence of mixed-gender teams (female, male), indicating some progress in gender diversity in collaborative research. The proportion of publications with authors with unidentified gender remains relatively small but consistent over time. Over time, there is a noticeable increase in the number of publications across all gender composition categories, particularly from January 2023 onward, coinciding with the rise in interest in LLMs and ChatGPT.

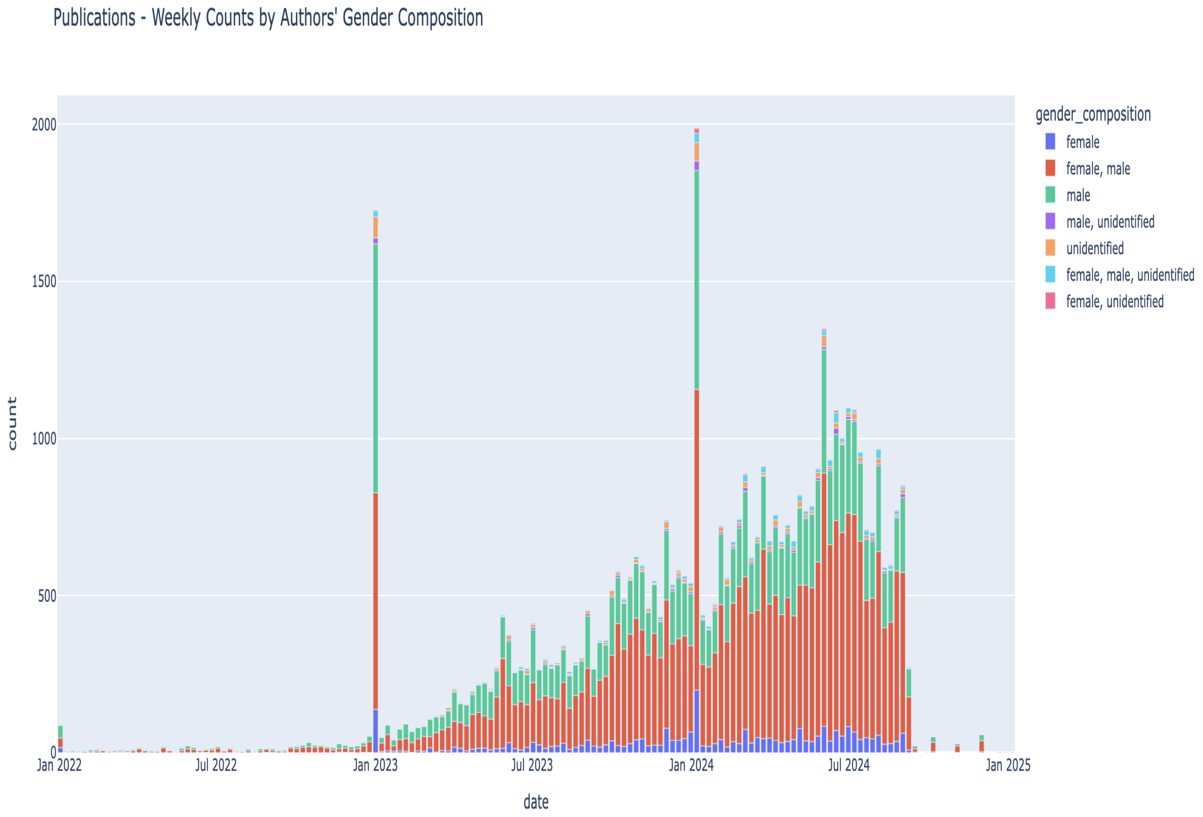


Figure 6: Weekly gender composition of authors.

As for the time distribution of female participation in the publications of the dataset, Figure 7 tracks the weekly count of publications with at least one female author (denoted by female participation taking the value 1) compared to those with no female authors (female participation = 0). The data shows a steady increase in the number of publications with female participation starting from January 2023, reflecting a growing inclusion of female authors in research. Despite this growth, publications with no female authors continue to outnumber those with female participation, highlighting ongoing gender disparities in authorship of publications on LLMs and ChatGPT. By January 2025, the gap between publications with and without female authors narrows slightly, suggesting gradual progress toward greater gender inclusivity.

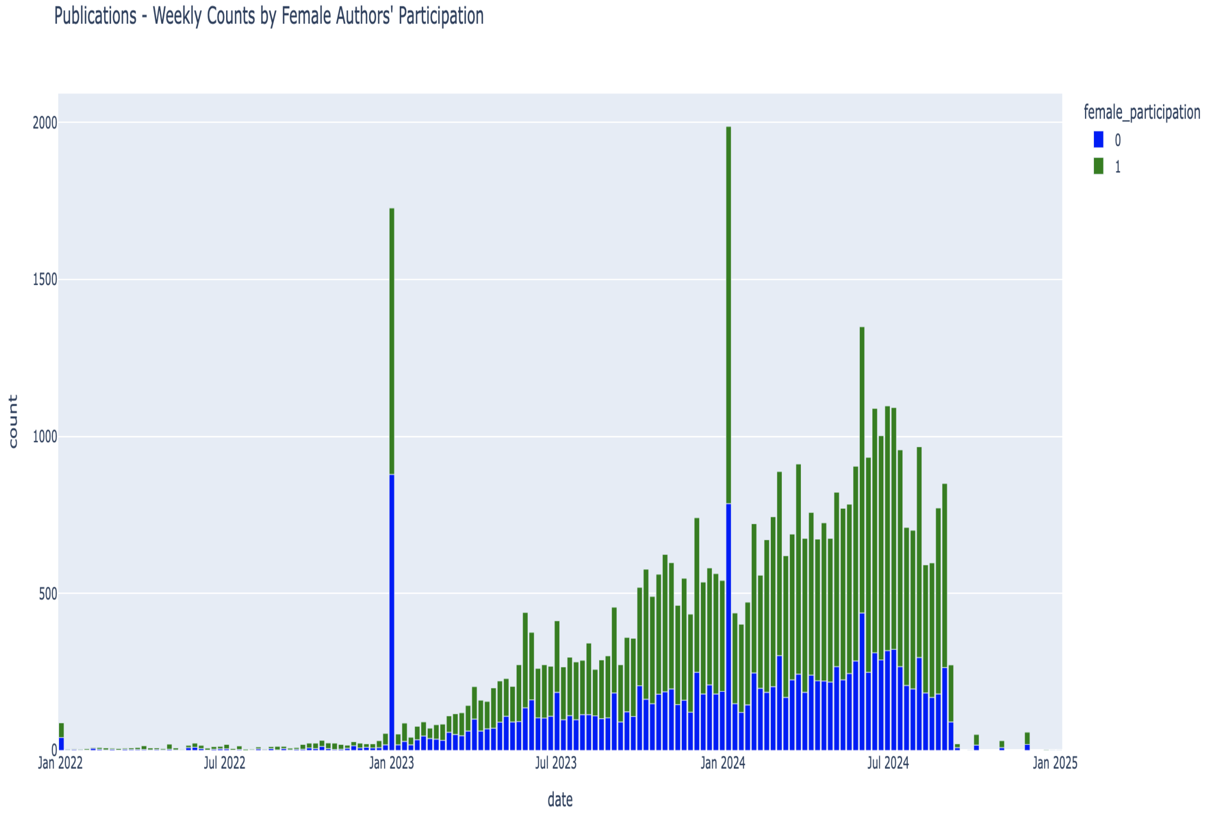


Figure 7: Weekly female participation in publications.

Publication Types

The distribution of publication types is shown in Figure 8. Clearly, the majority of publications are preprints, accounting for 46.6% (22,299 out of 47,855) of the total. This highlights the prevalence of early-stage research shared before formal peer review. Peer-reviewed articles represent 34.1% (16,331 out of 47,855) of the publications, indicating a significant portion of formally published research. This is likely driven by the rapid emergence of LLMs and ChatGPT, which has led researchers in the field to expedite the publication of their findings, while journals are eager to publish them swiftly to demonstrate their engagement with this dynamically growing area. Conference proceedings make up13.7% (6,566 out of 47,855), reflecting the importance of academic conferences in disseminating research. Book chapters constitute 5.3% (2,544 out of 47,855), suggesting a smaller but notable contribution to the literature. Monographs are rare, representing only 0.2% (108 out of 47,855) of the publications. Furthermore, seminar publications are the least common, accounting for just 0.01% (7 out of 47,855) of the total. In summary, preprints and articles dominate the publication landscape, together comprising over 80% of the total. Proceedings and chapters contribute moderately, while monographs and seminars are relatively rare. This distribution reflects the diverse ways in which research is shared and disseminated.

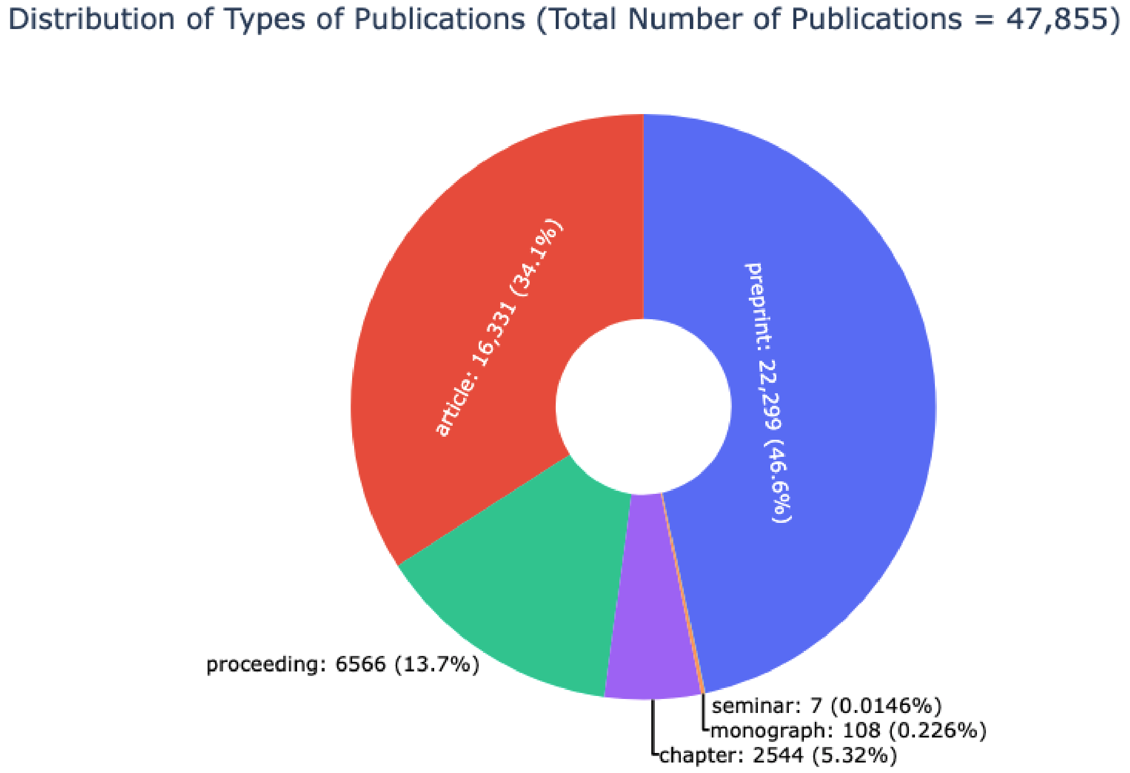


Figure 8: Distribution of publication types.

Similarly, the weekly distribution of publication types, shown in Figure 9, reveals trends that closely mirror those observed in the overall weekly distribution of publications: Preprints consistently dominate the weekly publication counts, showing a sharp increase starting around January 2023. This surge aligns with the growing adoption of preprint platforms for rapid dissemination of research, particularly in fast-moving fields of research on LLMs andChatGPT. Peer-reviewed articles also show a steady increase over time, with a noticeable rise from January 2023 onward. This reflects the continued importance of formal publication channels in academic research. Conference proceedings exhibit periodic spikes, likely corresponding to major academic conferences. These spikes are particularly prominent in 2023 and 2024, indicating the role of conferences in sharing cutting-edge research. Book chapters maintain a low but consistent presence throughout the period, with no significant spikes or declines. Monographs and seminars are rare, appearing only sporadically in the weekly counts. Their minimal presence highlights their niche role in the broader publication landscape.

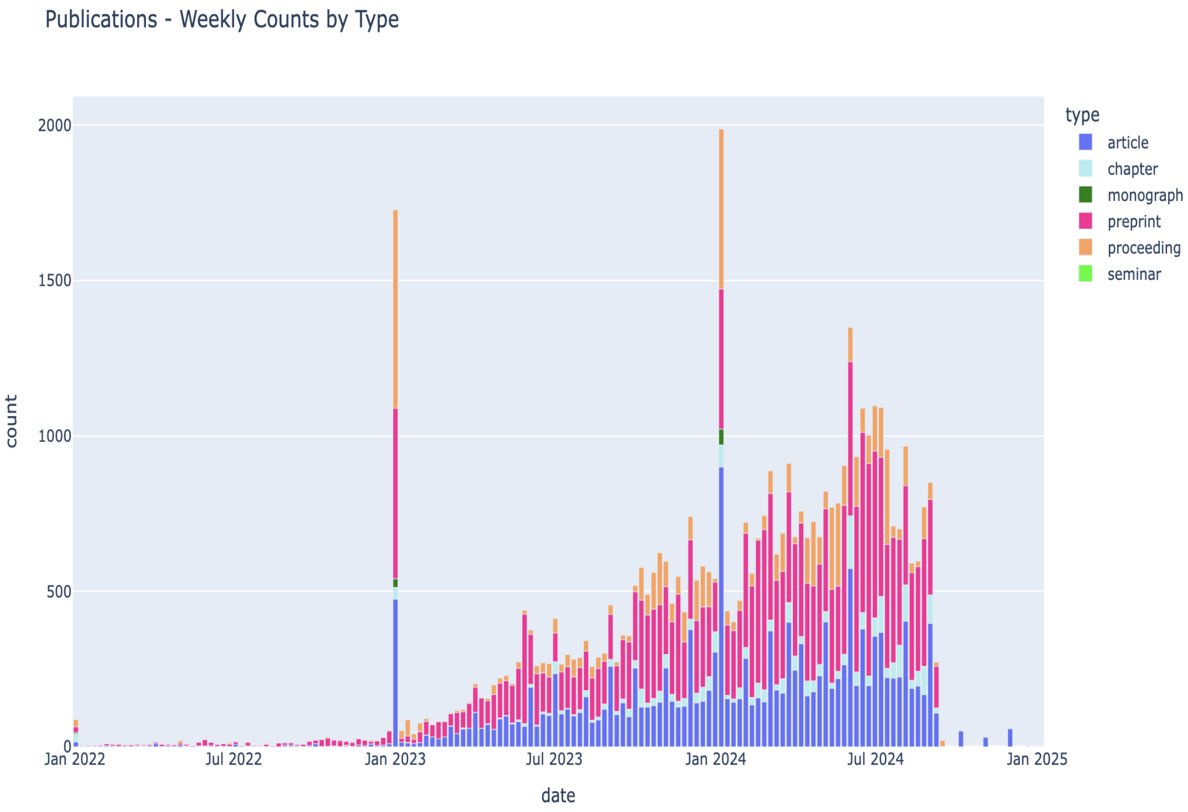


Figure 9: Weekly distribution of publication types.

It is worth noting that textbooks and monographs remain rare in the emerging fields of LLMs and ChatGPT studies for several reasons. The rapid pace of development in these fields makes textbooks less suitable, as they take longer to produce and require frequent updates to reflect the latest advancements. In contrast, peer-reviewed journal articles are the preferred method of communication for researchers in such dynamic fields, as they allow for quicker dissemination of new findings. Additionally, the highly specialized and niche nature of subfields within LLMs and ChatGPT means that researchers often focus on publishing short, focused papers rather than comprehensive books. Digital resources, such as preprints and online tutorials, are more common in these areas since they can be easily updated and distributed to a global audience. Furthermore, the community of practice in AI and machine learning often relies on online platforms, forums, and conference presentations as primary sources of knowledge, reducing the need for traditional textbooks and monographs.

Research Areas

The bar plot (Figure 10) and the word cloud (Figure 11) of the top research areas, each with a frequency exceeding 1,000, derived from the `category_for` field of publications on LLMs and ChatGPT, expose the following key insights:

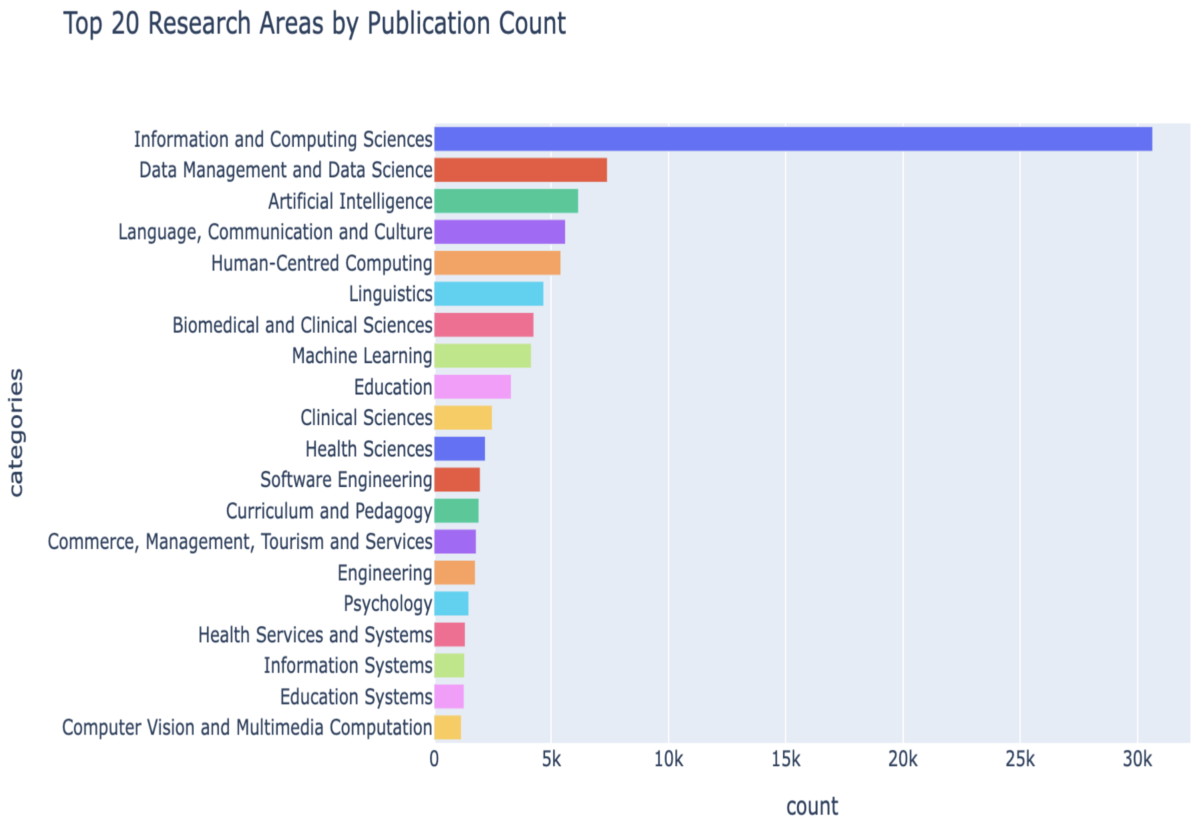


Figure 10: Top 20 research areas.

- **Dominant Research Areas**: The most prominent research area (FoR) is *Information and Computing Sciences* (30,651), reflecting the core technological foundation of LLMs and ChatGPT. Other significant areas of research include *Data Management and Data Science* (7,387), *Artificial Intelligence* (6,154), and *Machine Learning* (4,145), highlighting the central role of computational and AI-driven research.

- **Language and Communication**: Research areas such as *Language, Communication and Culture* (5,603) and *Linguistics* (4,674) are highly represented, emphasizing the focus on natural language processing and the linguistic capabilities of LLMs.

- **Healthcare and Biomedical Applications**: Areas of research like *Biomedical and Clinical Sciences* (4,250), *Clinical Sciences* (2,478), and *Health Sciences* (2,184) demonstrate the growing application of LLMs in healthcare and biomedical research.

- **Education**: The significant presence of *Education* (3,287), *Curriculum and Pedagogy* (1,908), and *Education Systems* (1,278) underscores the impact of LLMs on educational research and practice.

- **Interdisciplinary and Applied Research Areas**: Research areas such as *Human-Centred Computing* (5,404), *Software Engineering* (1,966), *Commerce, Management, Tourism and Services* (1,795), and *Engineering* (1,757) highlight the interdisciplinary applications of LLMs across various domains.

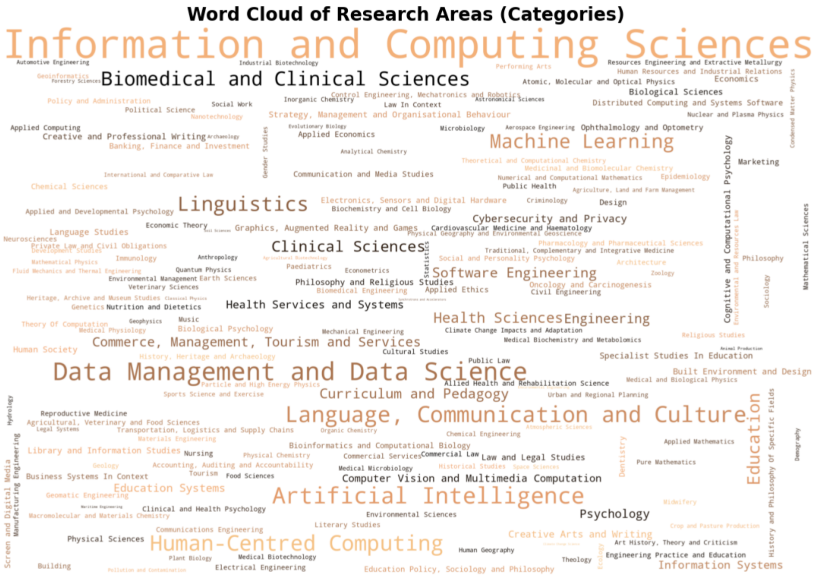

Word Cloud of Research Areas (Categories)
Information and Computing Sciences
Biomedical and Clinical Sciences
Machine Learning
Linguistics
Clinical Sciences
Software Engineering
Health Services and Systems
Health Sciences
Engineering
Commerce, Management, Tourism and Services
Data Management and Data Science
Curriculum and Pedagogy
Language, Communication and Culture
Education
Artificial Intelligence
Psychology
Human-Centred Computing
Information Systems
Computer Vision and Multimedia Computation
Cybersecurity and Privacy
Education Systems
Strategy, Management and Organisational Behaviour
Biological Sciences
Economics
Political Science
Applied Economics
Creative and Professional Writing
Communication and Media Studies
Philosophy and Religious Studies
Law and Legal Studies
Creative Arts and Writing
Bioinformatics and Computational Biology
Cognitive and Computational Psychology
Distributed Computing and Systems Software
Graphics, Augmented Reality and Games
Specialist Studies In Education
Built Environment and Design
Education Policy, Sociology and Philosophy
Engineering Practice and Education
Physical Sciences
Electrical Engineering
Medical Biotechnology
Environmental Sciences
Library and Information Studies
Marketing
Mathematical Sciences
Applied Computing
Public Health
Design

Figure 11: Word cloud of top research areas.

- **Emerging and Niche Research Areas**: Fields like *Cybersecurity and Privacy* (1,156), *Computer Vision and Multimedia Computation* (1,160), and *Creative Arts and Writing* (1,055) indicate the expanding role of LLMs in specialized and emerging areas.

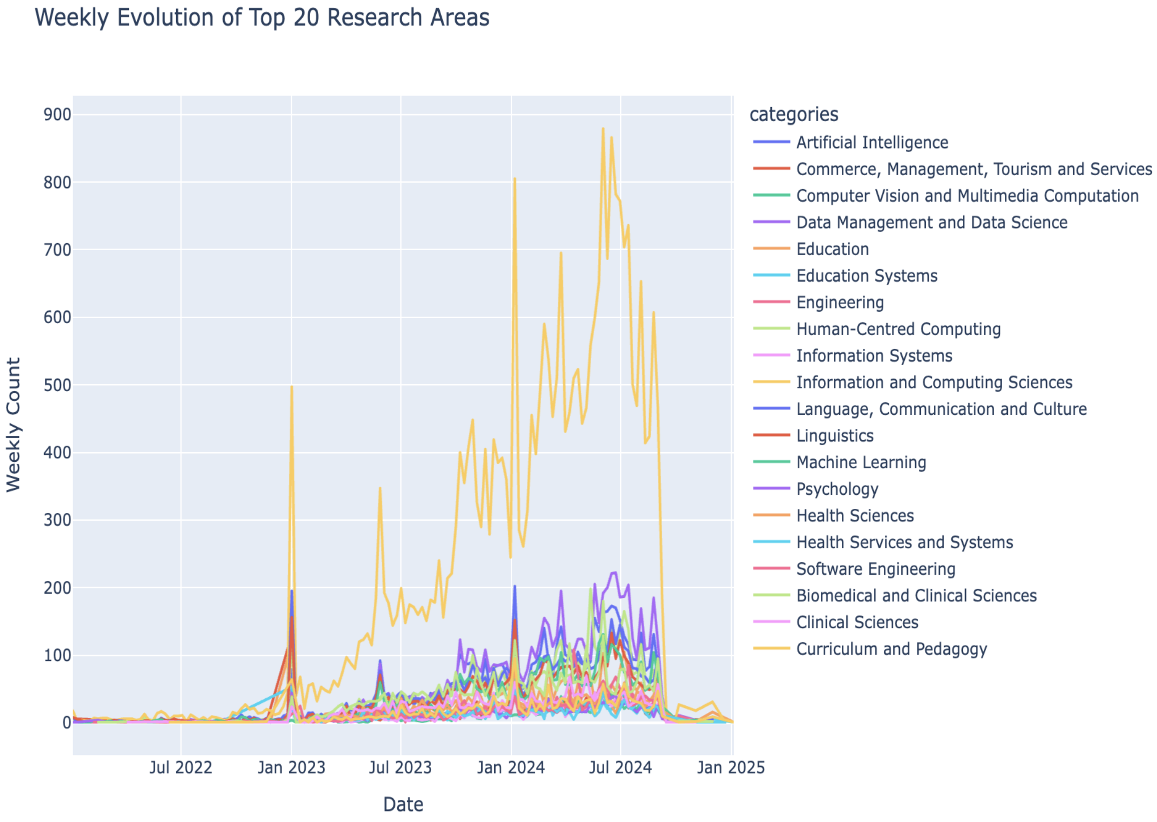


Figure 12: Weekly distribution of top 20 research areas.

In summary, the word cloud illustrates the extensive and diverse impact of LLMs and ChatGPT, with a strong emphasis on computational sciences, language and communication, healthcare, education, and interdisciplinary applications. The prominence of these research areas reflects both the technological core of LLMs and their broad societal and practical implications.

The weekly distribution of the top 20 research areas, presented in Figure 12, shows a marked increase in research activity from July 2022 onward, with a notable acceleration beginning in January 2023. *Information and Computing Sciences* consistently leads with the highest weekly counts, while *Artificial Intelligence*, *Data Management and Data Science*, and *Machine Learning* demonstrate steady, robust growth. Interdisciplinary fields like *Language, Communication and Culture* and *Education* show gradual increases, reflecting the expanding applications of LLMs. Meanwhile, healthcare-related fields, such as *Biomedical and Clinical Sciences* and *Health Sciences*, experience more moderate growth, indicating their niche but sustained presence in the research landscape.

Research Areas and Journals vs. Citations by Gender

The relationship between research areas and citation counts, differentiated by female author participation (ranging from 0 to 1), reveals certain interesting patterns. As Figure **??** shows, the highest female participation (score = 1) is observed in publications on LLMs and ChatGPT, particularly in fields such as Environmental Engineering, Climate Change Science, and Ecological Applications, where it coincides with high citation counts. In contrast, in fields like Visual Arts, Animal Production, and Oceanography, it is associated with low citation counts. At the other end of the spectrum, As Figure **??** shows, publications in Horticultural Production receive high citation counts despite the absence of female participation; additionally, fields such as Maritime Engineering, Mathematical Physics, and Pure Mathematics exhibit low female participation (below 0.25 but nonzero) and fall within the 40th percentile of the citation distribution. Overall, the dataset indicate that female authors tend to be more prominently represented in highly cited publications within certain fields, particularly those related to biology, engineering, and tourism.

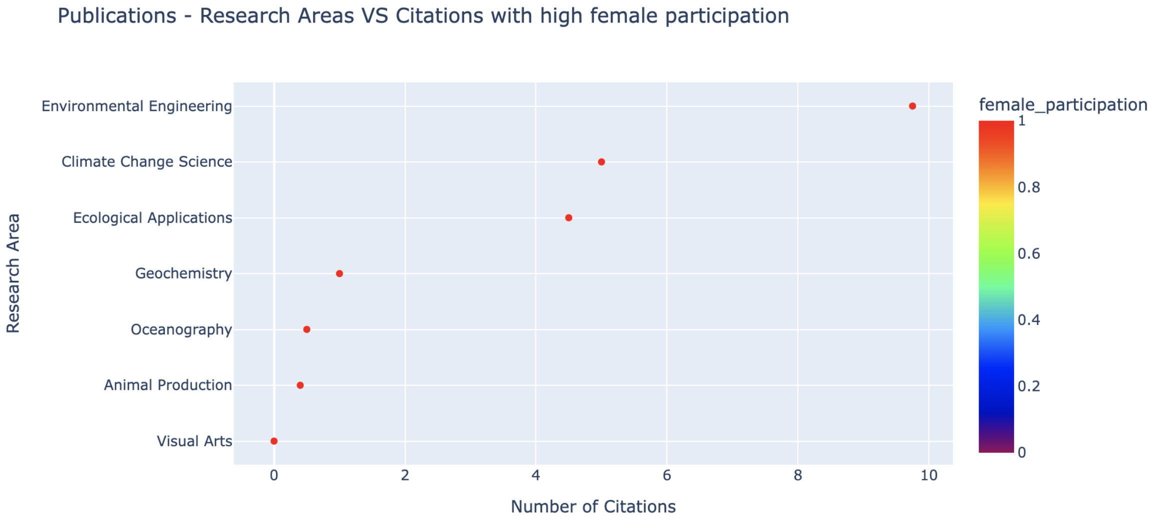


Figure 13: Research areas vs. citations by gender with high female participation.

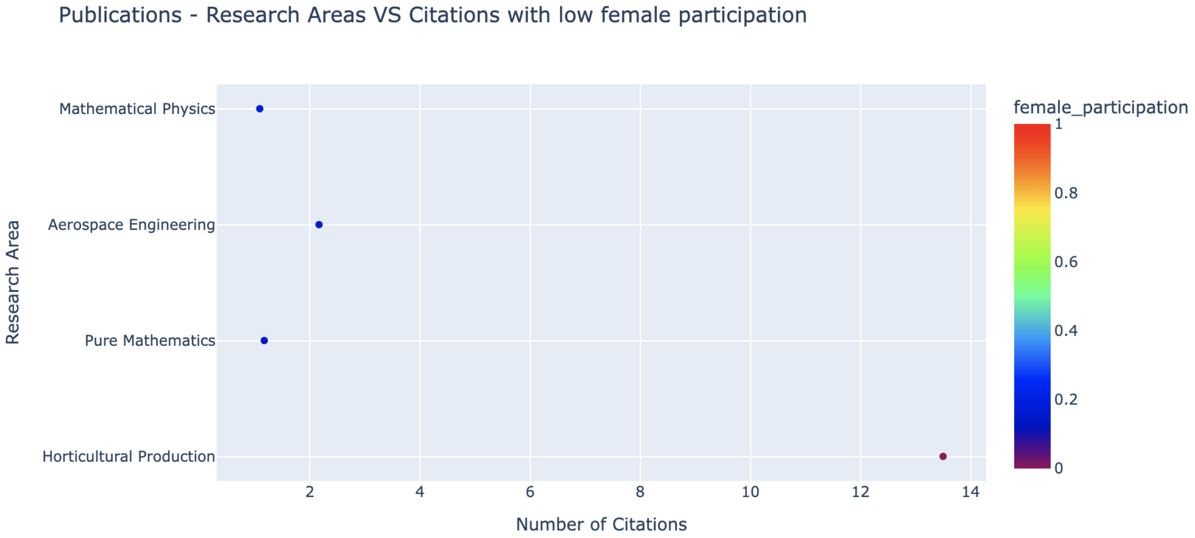


Figure 14: Research areas vs. citations by gender with low female participation.

Furthermore, in Figure 15, we examine how the influence of author gender on participation across research areas and publication journals relates to the number of citations (as indicated by the `times_cited` field). Notably, the journals with female participation and the highest citation counts (> 1000) include *PLOS Digital Health*, *International Journal of Information Management*, *Science*, *Learning and Individual Differences*, and *JMIR Medical Education*. Conversely, the journals with no female participation and the highest citation counts (> 500) are *Healthcare*, *Internet of Things and Cyber-Physical Systems*, *Science*, *Cureus*, *Smart Learning Environments*, *Education Sciences*, *Nature*, and *Frontiers in Artificial Intelligence*. This distribution suggests that female participation is more prominent in highly cited research within fields like digital health, information management, medical education, and learning sciences. The presence of *Science* in both categories indicates that elite journals publish highly cited research regardless of gender representation, but disparities in authorship persist. In contrast, fields related to engineering, AI, and cybersecurity, as reflected in journals like *Internet of Things and Cyber-Physical Systems* and *Smart Learning Environments*, continue to exhibit lower female representation despite high citation counts. The findings highlight ongoing gender imbalances in authorship, particularly in STEM fields, suggesting that structural barriers may still limit female researchers' participation in certain disciplines.

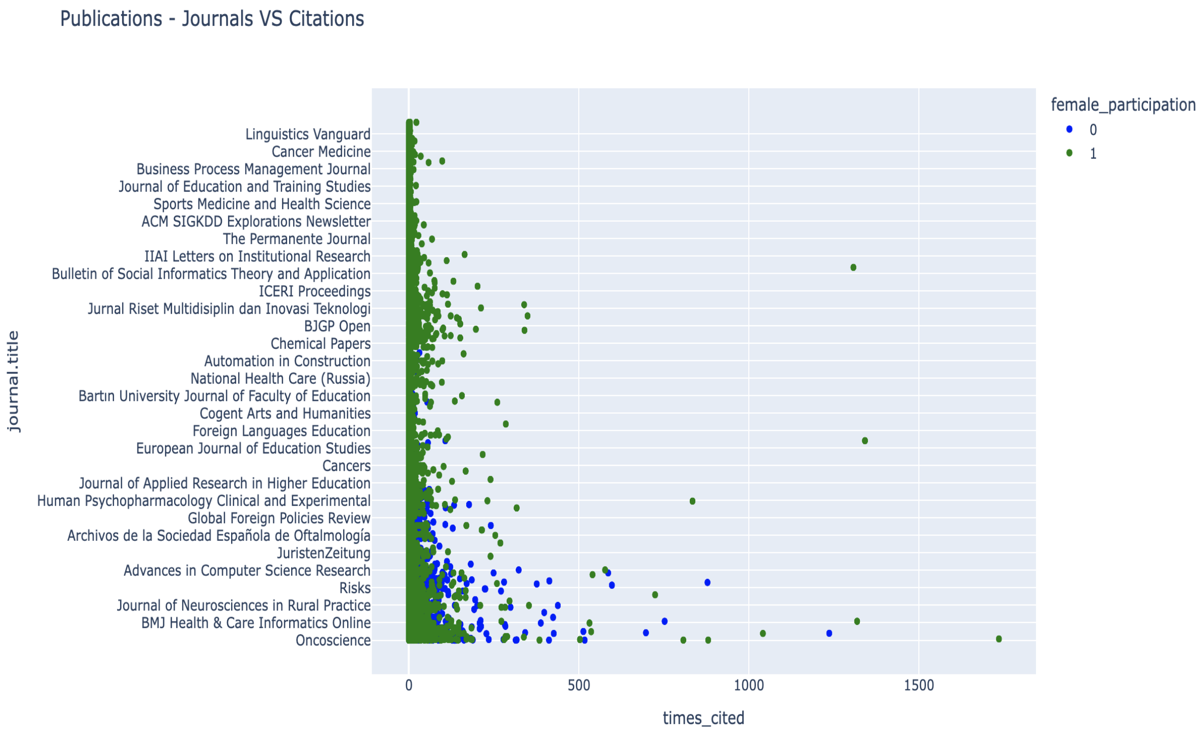


Figure 15: Journals vs. citations by gender.

Publications Grants Funding by Gender

Figure 16 illustrates the trajectory of active funding years for grants awarded to research on ChatGPT and large language models (LLMs). The data reveal a steady increase in both the total number of grants and those specifically awarded to female researchers, peaking around 2025. This upward trend highlights the growing recognition and financial support for advancements in this domain. However, the decline observed after 2025 can be attributed to the natural expiration of grants awarded in previous funding cycles. Since research grants typically have fixed durations, often spanning multiple years, the observed drop does not necessarily indicate a reduction in new funding but rather reflects the conclusion of earlier awarded grants. This pattern is a common feature of funding timelines, where periods of sustained growth are followed by declines as existing grants reach their completion dates.

The close alignment between the overall funding trend and the trajectory of grants awarded to female researchers suggests that while women in this field have benefited from the general increase in funding, their representation has not significantly diverged from broader patterns. This indicates that the rise in funding for female researchers is largely a reflection of overall funding growth rather than a targeted effort to address gender disparities. As a result, sustained and deliberate investment strategies are necessary to promote greater diversity and inclusion in AI research. Ensuring consistent financial support, particularly through initiatives designed to increase opportunities for underrepresented groups, is crucial for fostering a more equitable research environment. By prioritizing inclusive funding policies, the field can encourage a broader range of perspectives, drive innovation, and contribute to the long-term advancement of AI.

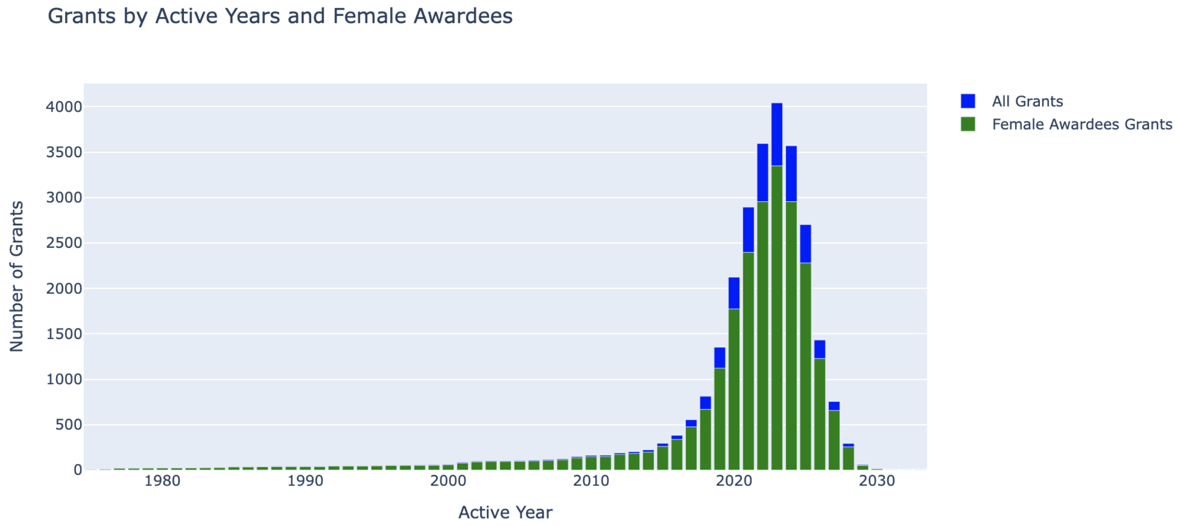
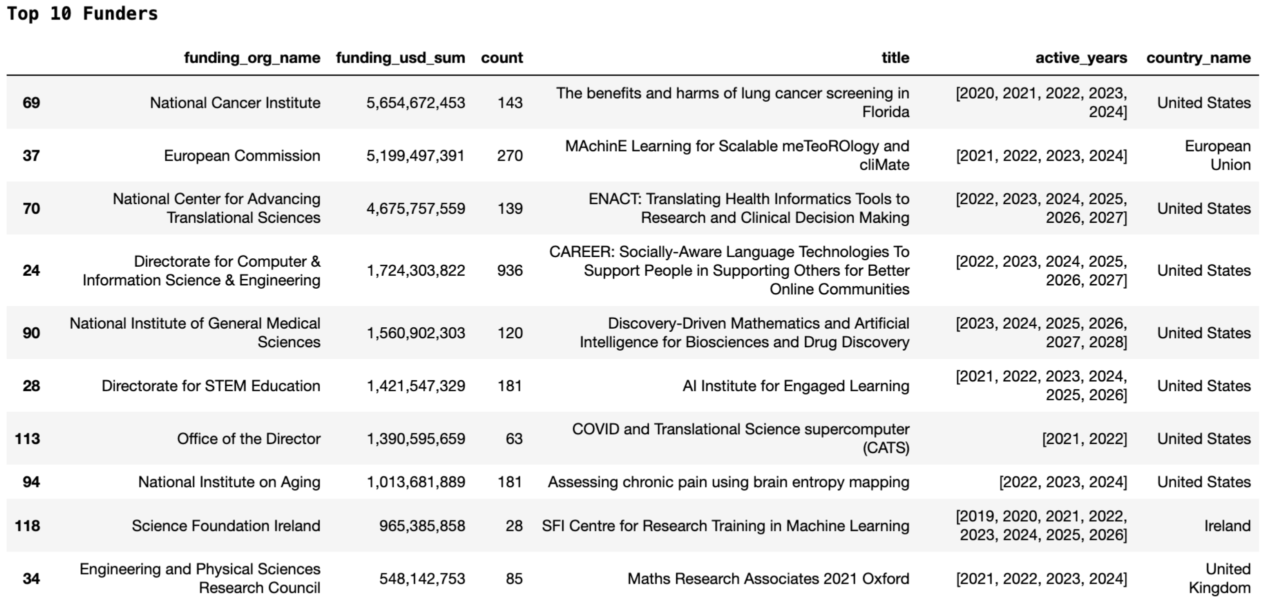


Figure 16: Funding received by publications, categorized by gender.

Top Funders

The table in Figure 17 lists the Top 10 Funders of grants awarded to publications related to ChatGPT and LLMs, providing insights into the funding landscape for this rapidly growing field. The National Cancer Institute (NCI) leads with the highest funding amount at 5.65 billion, followed by the European Commission at 5.20 billion. These organizations are investing heavily in research that intersects with AI, machine learning, and healthcare applications. Other significant funders include the National Center for Advancing Translational Sciences (NCATS) (1.72 billion), highlighting the importance of translational research and socially-aware AI technologies.

The European Commission has awarded the most grants (270), indicating a broad and diverse portfolio of projects. In contrast, the National Cancer Institute has fewer grants (143) but with much larger individual funding amounts, suggesting a focus on high-impact, large-scale projects. The funded projects span a wide range of applications, including healthcare (e.g., lung cancer screening, chronic pain assessment), climate and meteorology (e.g., MAchinE Learning for Scalable meTeoROlogy), and education (e.g., AI Institute for Engaged Learning). This reflects the interdisciplinary nature of ChatGPT and LLM research. Several projects emphasize translational science and health



Top 10 Funders

	funding_org_name	funding_usd_sum	count	title	active_years	country_name
69	National Cancer Institute	5,654,672,453	143	The benefits and harms of lung cancer screening in Florida	[2020, 2021, 2022, 2023, 2024]	United States
37	European Commission	5,199,497,391	270	MAchinE Learning for Scalable meTeORology and cliMate	[2021, 2022, 2023, 2024]	European Union
70	National Center for Advancing Translational Sciences	4,675,757,559	139	ENACT: Translating Health Informatics Tools to Research and Clinical Decision Making	[2022, 2023, 2024, 2025, 2026, 2027]	United States
24	Directorate for Computer & Information Science & Engineering	1,724,303,822	936	CAREER: Socially-Aware Language Technologies To Support People in Supporting Others for Better Online Communities	[2022, 2023, 2024, 2025, 2026, 2027]	United States
90	National Institute of General Medical Sciences	1,560,902,303	120	Discovery-Driven Mathematics and Artificial Intelligence for Biosciences and Drug Discovery	[2023, 2024, 2025, 2026, 2027, 2028]	United States
28	Directorate for STEM Education	1,421,547,329	181	AI Institute for Engaged Learning	[2021, 2022, 2023, 2024, 2025, 2026]	United States
113	Office of the Director	1,390,595,659	63	COVID and Translational Science supercomputer (CATS)	[2021, 2022]	United States
94	National Institute on Aging	1,013,681,889	181	Assessing chronic pain using brain entropy mapping	[2022, 2023, 2024]	United States
118	Science Foundation Ireland	965,385,858	28	SFI Centre for Research Training in Machine Learning	[2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026]	Ireland
34	Engineering and Physical Sciences Research Council	548,142,753	85	Maths Research Associates 2021 Oxford	[2021, 2022, 2023, 2024]	United Kingdom

Figure 17: Top 10 funders.

informatics, indicating a strong focus on applying AI to real-world problems, particularly in medicine and public health.

Most grants have active funding periods spanning 4 to 6 years, with some extending up to 8 years (e.g., Science Foundation Ireland). This suggests long-term commitments to advancing research in ChatGPT and LLMs. The Office of the Director has a shorter active period (2 years) for its COVID and Translational Science supercomputer project, likely reflecting the urgency of pandemic-related research. The majority of top funders are based in the United States, with significant contributions from the European Union, Ireland, and the United Kingdom. This highlights the global nature of AI research funding, with Western countries leading the way.

The table underscores the strategic importance of ChatGPT and LLMs across various domains, particularly in healthcare, climate science, and education. The substantial funding amounts and long active years indicate that these technologies are seen as critical to addressing complex societal challenges. The dominance of U.S.-based funders reflects the country's leadership in AI research, while the involvement of European organizations demonstrates a collaborative, international effort. The focus on translational research and real-world applications suggests that funders are prioritizing projects with tangible societal impacts, rather than purely theoretical advancements. Overall, this funding landscape highlights the growing recognition of ChatGPT and LLMs as transformative technologies with wide-ranging implications.

Figure 18 illustrates that female researchers have secured a substantial proportion of funding from two of the largest funding bodies, the National Cancer Institute and the European Commission. However, discrepancies are observed with other major funders, such as the United States Department of the Air Force and the Novo Nordisk Foundation, where the total funding awarded contrasts notably with the amount allocated to female researchers. This suggests that female researchers may receive a smaller share of grants from these organizations. Overall, our analysis underscores persistent gender disparities in the allocation of research funding, highlighting the ongoing challenges faced by women in securing grants. These findings emphasize the urgent need for policies that promote equitable access to funding, particularly in high-investment fields like engineering, physical sciences, and biomedical research, where disparities are particularly pronounced.

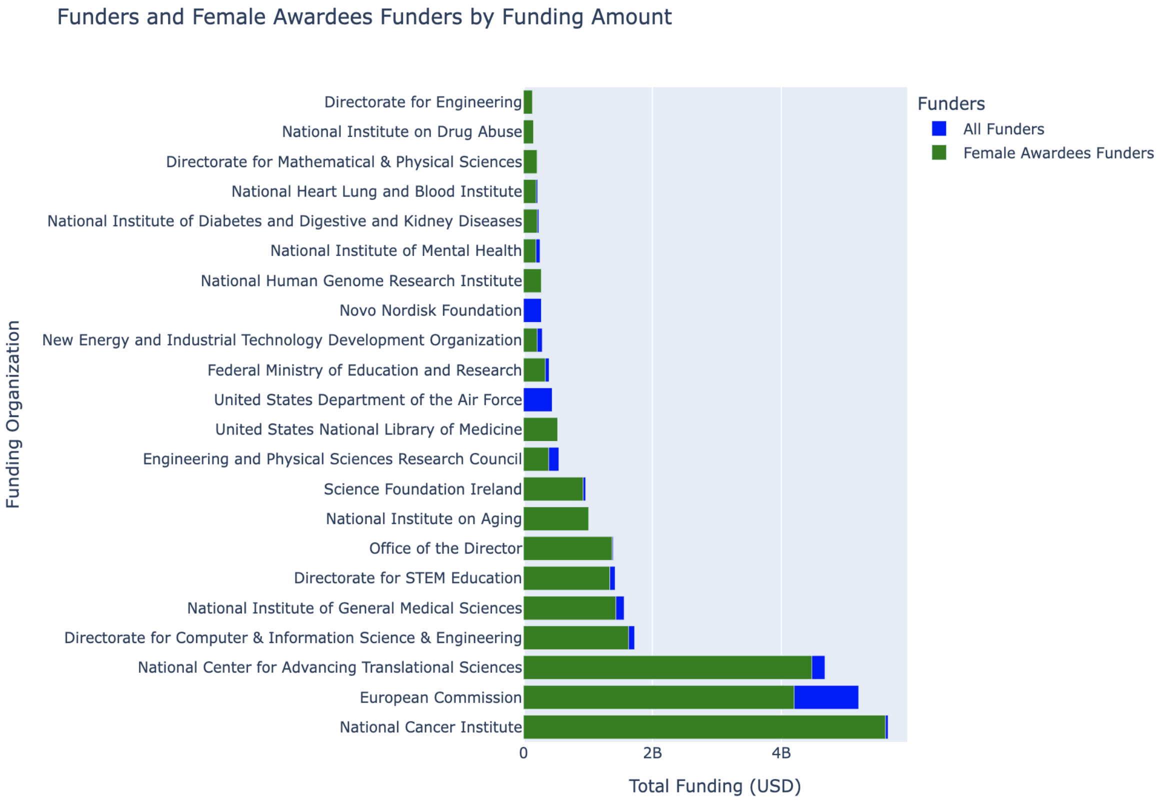


Figure 18: Funding distribution by gender for top funders.

Countries of the Top 50 Funders

Figure 19 presents data on funding organizations and the corresponding amounts of funding in U.S. dollars allocated by various countries for grants awarded to publications on ChatGPT and LLMs. The United States consistently emerges as the top contributor, both in terms of funding volume and the number of publications, with funding amounts ranging from billions to hundreds of millions of dollars. The National Cancer Institute (NCI) stands out as the highest-funded organization, with a substantial sum of 5.6 billion dollars, underscoring the U.S.'s prioritization of cancer research. Other notable U.S.-based organizations include the National Center for Advancing Translational Sciences (NCATS) and the National Institute of General Medical Sciences (NIGMS), which also receive billions in funding, highlighting the emphasis on translational research and basic medical sciences.

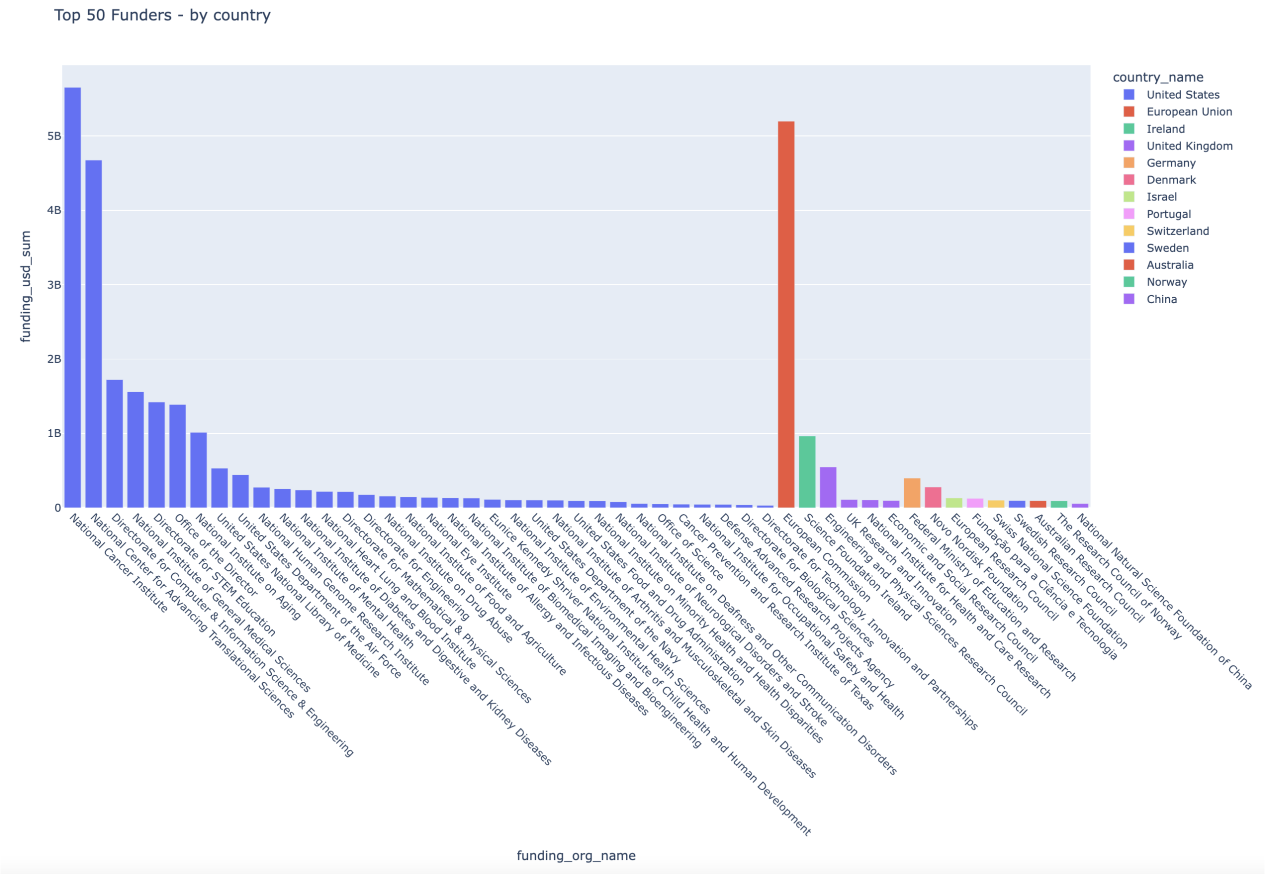


Figure 19: Countries of the top 50 funders.

The European Union, represented by the European Commission, ranks as the second-largest funder, with a notable 5.2 billion dollars allocated, indicating a strong commitment to research and innovation at the continental level. Other European countries, such as Ireland, the United Kingdom, Germany, and Portugal, also contribute significantly, though their funding amounts are notably lower than those of the U.S. and the EU. For example, the Science Foundation Ireland and the UK's Engineering and Physical Sciences Research Council (EPSRC) receive hundreds of millions of dollars, reflecting their focus on advancing scientific and technological fields.

Additionally, non-U.S. and non-EU entities, such as the Novo Nordisk Foundation in Denmark and the National Natural Science Foundation of China, are also significant contributors, although their funding amounts are more modest in comparison. This suggests that while these countries are active in supporting research, their investments are not yet on the same scale as those of the U.S. or the EU.

Overall, the table illustrates the global landscape of research funding, with the U.S. and the EU leading the way. It emphasizes the prioritization of health-related research, especially in areas such as cancer, translational sciences, and general medical sciences, while also acknowledging the contributions of other countries and organizations to the global research ecosystem. The disparities in funding amounts reflect differing national priorities, economic capacities, and research strategies.

The dominance of the United States in both funding allocation and publication output on ChatGPT and LLMs is further confirmed by Figure 20. With an astounding 21.76 billion dollars in funding and 2,917 publications, the U.S. demonstrates its unparalleled commitment to research and development, as well as its capacity to produce a vast amount of scholarly work. This reflects the country's robust infrastructure, significant investment in science and technology, and the presence of numerous leading research institutions.

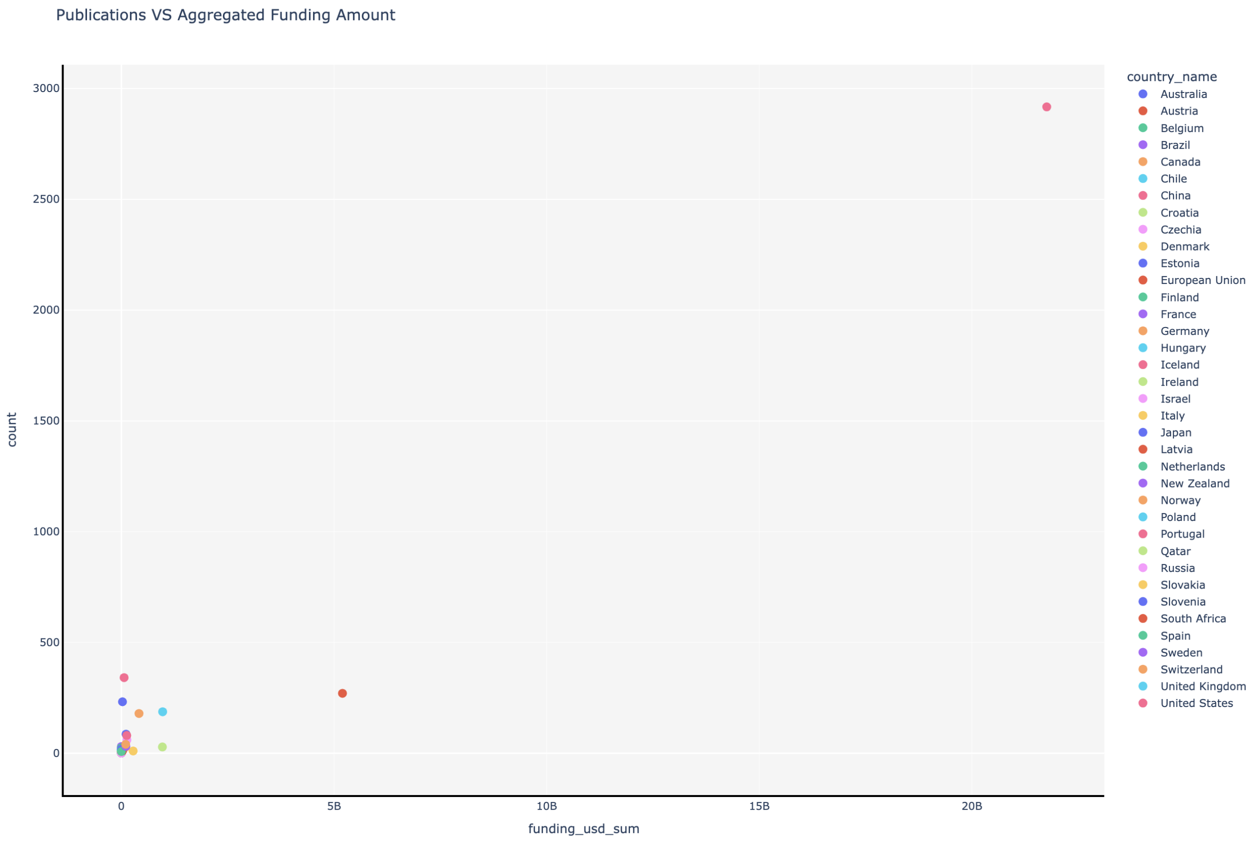


Figure 20: Number of publications vs. aggregated funding by country.

China, with 64.9 million dollars in funding and 341 publications, demonstrates a high level of research productivity relative to its funding. This suggests efficient resource use and a focused effort to generate scholarly output, aligning with China's growing emphasis on becoming a global leader in science and technology.

Other countries with notable funding and publication counts include Germany (415 million dollars and 179 publications), the United Kingdom (971 million dollars and 187 publications), and Ireland (965 million dollars and 28 publications). While Germany and the UK maintain a balanced relationship between funding and output, Ireland's high funding relative to its publication count may indicate that their research is focused on high-cost areas or longer-term projects that have not yet resulted in a proportional number of publications.

Denmark and Portugal also stand out, with Denmark's 278 million dollars in funding yielding only 10 publications, while Portugal's 127 million dollars resulted in 80 publications. Denmark's higher funding per publication suggests a focus on specialized or resource-intensive research, while Portugal's relatively higher publication count compared to its funding indicates more efficient research output.

Countries like Belgium, Brazil, Chile, Croatia, Hungary, Iceland, Latvia, New Zealand, Qatar, Russia, and Slovenia report zero funding but still contribute to the publication count. This suggests that research for these publications may have been supported through non-monetary means, possibly through international collaborations or funding sources not captured in this dataset.

In summary, this figure emphasizes the strong correlation between research funding and publication output, with the United States leading by a significant margin. It also highlights the varying efficiencies and research priorities across countries, with some nations achieving high publication counts with lower funding, while others focus heavily on specialized or long-term research projects. These findings offer valuable insights into global research capacities, productivity, and strategic priorities.

Conclusion

This study provides a comprehensive analysis of gender diversity and funding allocation in post-2022 publications on LLMs and ChatGPT. The findings reveal significant patterns in gender diversity and funding allocation within this rapidly evolving field. The study demonstrates that despite the explosive growth in research output, gender disparities persist in authorship and funding distribution. Female authors comprise approximately 30% of total contributors, with male authors representing nearly 68%, highlighting a substantial gender gap in representation within this technological domain.

The temporal analysis reveals interesting dynamics in publication trends and gender participation. Following ChatGPT's public release in late 2022, research output increased dramatically, with publications growing steadily through January 2025. While female participation in authorship showed gradual improvement during this period, publications without female authors consistently outnumbered those with female contributors, indicating persistent gender disparities despite overall growth in research activity.

In our findings, the analysis of citation patterns across journals reveals that female participation is more prevalent in highly cited research within fields like digital health, information management, and medical education, whereas journals in engineering, AI, and cybersecurity continue to show lower female representation despite high citation counts. This underscores persistent gender disparities in authorship, particularly in STEM-related disciplines.

Funding patterns reveal both opportunities and challenges for diversity in LLM research. The National Cancer Institute and European Commission emerge as leading funders, with total allocations exceeding 5 billion dollars each. While female researchers secured substantial funding from major organizations, significant discrepancies exist across different funding bodies, suggesting uneven access to resources. The United States dominates both funding allocation and publication output, with 21.76 billion dollars invested and 2,917 publications produced, followed by the European Union and China.

Several key implications emerge from these findings. First, while the field demonstrates remarkable growth and innovation, the persistence of gender disparities suggests a need for targeted initiatives to promote diversity. Second, the correlation between funding levels and publication output indicates that resource allocation plays a crucial role in research productivity. Finally, the dominance of certain regions and organizations highlights the importance of international collaboration and equitable distribution of resources.

Future research directions should focus on several areas:

1. Investigating the relationship between gender diversity and innovation outcomes in LLM research.
2. Analyzing the impact of funding mechanisms on promoting gender equality in AI research.
3. Examining regional differences in gender representation and funding patterns.
4. Studying the influence of institutional policies on gender diversity in technological research.
5. Developing frameworks for measuring and addressing systemic barriers to inclusion in AI research.

These findings contribute to our understanding of diversity patterns in emerging technologies while highlighting the need for sustained efforts to promote inclusivity in AI research. The study demonstrates that despite rapid progress in the field, achieving greater gender diversity remains an essential challenge for ensuring diverse perspectives and equitable opportunities in LLM and ChatGPT research.

References

[1] ANZSRC classifications (published by the Australian Bureau of Statistics and Statistics New Zealand), https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-research-classification-anzsrc

[2] Bornmann, L., (2017), Measuring impact in research evaluations: A thorough discussion of methods for, effects of and problems with impact measurements, https://link.springer.com/article/10.1007/s10734-016-9995-x.

[3] De Bellis, N. (2009), *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybernetics*, The Scarecrow Press

[4] Delgado-Quirós, L., & Ortega, J. L. (2024), Completeness degree of publication metadata in eight free-access scholarly databases, *Quantitative Science Studies*, https://doi.org/10.1162/qss_a_00286.

[5] Dimensions grants, https://docs.dimensions.ai/dsl/2.0.0/datasource-grants.html.

[6] Dimensions publications fields, https://docs.dimensions.ai/dsl/2.0.0/datasource-publications.html.

[7] Hicks, D., & Melkers, J. (2012), Bibliometrics as a tool for research evaluation. In Link, A. N., & Vonortas, N. S. (ed.), *Handbook on the Theory and Practice of Program Evaluation*, Edward Elgar Publishing https://doi.org/10.4337/9780857932402.00019

[8] Hook, D. W., Porter, S. J., & Herzog, C. (2018), Dimensions: Building context for search and evaluation, *Frontiers in Research Metrics and Analytics*, https://www.frontiersin.org/journals/research-metrics-and-analytics/articles/10.3389/frma.2018.00023/full.

[9] Namsor, a name checking technology. https://namsor.app/about-us/.

[10] Namsor Gender API

Software

. https://namesorts.com/api/.

[11] Nguyen, B. X., Luczak-Roesch, M., Dinneen, J. D., & Larivière, V. (2022), Assessing the quality of bibliographic data sources for measuring international research collaboration, *Quantitative Science Studies*, https://doi.org/10.1162/qss_a_00211.

[12] Science-Metrix. (2018). *Analytical support for bibliometrics indicators: Development of bibliometric indicators to measure women's contribution to scientific publications* (Final Report), https://namsor.app/files_to_download_p/science-metrix_bibliometric_indicators_womens_contribution_to_science_report.pdf.

[13] Sebo, P. (2021), Using genderize.io to infer the gender of first names: How to improve the accuracy of the inference, *Journal of the Medical Library Association*, https://dx.doi.org/10.5195/jmla.2021.1252.

[14] The Dimensions Search Language (DSL), https://api-lab.dimensions.ai/cookbooks/1-getting-started/5-Deep-dive-DSL-language.html.

[15] What is the background behind the Fields of Research (FoR) classification system? https://plus.dimensions.ai/support/solutions/articles/23000018826-what-is-the-background-behind-the-fields-of-research-for-\protect\penalty\z@classification-system-

[16] Which research categories and classification schemes are available in Dimensions? https://plus.dimensions.ai/support/solutions/articles/23000018820-which-research-categories-and-classification-schemes-are-\protect\penalty\z@available-in-dimensions-

[17] Working with concepts in the Dimensions API, https://api-lab.dimensions.ai/cookbooks/1-getting-started/7-Working-with-concepts.html.