Topics, Authors, and Institutions in Large Language Model Research: Trends from 17K arXiv Papers

Rajiv Movva[1] **Sidhika Balachandar**[1] **Kenny Peng**[1]
Gabriel Agostini[1] **Nikhil Garg**[2] **Emma Pierson**[2]
Cornell Tech
rmovva@cs.cornell.edu

Abstract

Large language models (LLMs) are dramatically influencing AI research, spurring discussions on what has changed so far and how to shape the field's future. To clarify such questions, we analyze a new dataset of 16,979 LLM-related arXiv papers, focusing on recent trends in 2023 vs. 2018-2022. First, we study disciplinary shifts: LLM research increasingly considers societal impacts, evidenced by 20× growth in LLM submissions to the Computers and Society sub-arXiv. An influx of new authors – half of all first authors in 2023 – are entering from non-NLP fields of CS, driving disciplinary expansion. Second, we study industry and academic publishing trends. Surprisingly, industry accounts for a *smaller* publication share in 2023, largely due to reduced output from Google and other Big Tech companies; universities in Asia are publishing more. Third, we study institutional collaboration: while industry-academic collaborations are common, they tend to focus on the same topics that industry focuses on rather than bridging differences. The most prolific institutions are all US- or China-based, but there is very little cross-country collaboration. We discuss implications around (1) how to support the influx of new authors, (2) how industry trends may affect academics, and (3) possible effects of (the lack of) collaboration.

1 Introduction

Recent advances in language modeling have caused disruptive shifts throughout AI research, spurring conversation about how the field is changing and how it should change. Discussions so far have drawn on surveys and interviews of NLP community members (Michael et al., 2022; Gururaja et al., 2023; Lee et al., 2023), and recurring themes under consideration include which topics are shifting in importance, which directions remain fruitful for academics and other compute-limited researchers, and which institutions hold power to shape LLM research.

In periods of flux like this one, *bibliometrics* – quantitative study of publication patterns – offers a useful lens. Prior bibliometric analyses have been clarifying in NLP, identifying topic shifts (Hall et al., 2008), flows of authors in and out of the field (Anderson et al., 2012), and the growing role of industry (Abdalla et al., 2023). Due to the rapid recent growth of the LLM literature, fundamental questions about the topics, authors, and institutions driving its growth remain understudied.

Addressing this gap, we conduct a bibliometric analysis of recent trends in the LLM literature, focusing on changes in 2023 compared to 2018-22. We collect and analyze 16,979 LLM-related papers posted to arXiv from Jan. 1, 2018 through Sep. 7, 2023. In addition to arXiv metadata, we annotate these papers with topic labels, author affiliations, and citation counts, and make all data and code publicly available.[1] We analyze three questions:

1. **Which topics and authors are driving the growth of LLM research?** Following prior work, we study topics and author movement as markers of a field's evolution (Kuhn, 1962; Uban et al., 2021; Anderson et al., 2012). In 2023, LLM research increasingly focuses on societal impacts: the Computers and Society sub-arXiv has grown faster than any other sub-arXiv in its proportion of LLM papers, up by a factor of 20× in 2023. A more granular topic-level analysis echoes this result: the "Applications of ChatGPT" and "Societal Implications of LLMs" topics have grown 8× and 4× respectively. We also see rapid growth in other sub-arXivs outside of core NLP, including

[1]Co-first authors.
[2]Co-senior authors.

[1]https://github.com/rmovva/LLM-publication-patterns-public

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1223–1243

Human-Computer Interaction, Security, and Software Engineering. Simultaneously, BERT and task-specific architectures are shrinking due to centralization around newer models (e.g., GPT-4 and LLaMA).

The increased focus on societal impacts and non-NLP disciplines is driven by a strikingly large proportion of authors new to LLMs. Half (49.5%) of LLM first authors in 2023 have never previously co-authored an NLP paper (and nearly two-thirds haven't previously co-authored an LLM paper), and a substantial fraction (38.6%) of the last authors haven't either. These new authors are entering from other fields like Computer Vision, Software Engineering, and Security, and they are writing LLM papers on topics further out from NLP, e.g., vision-language models, applications in the natural sciences, and privacy/adversarial risks.

2. **What are the roles of industry and academia?** The role of industry – both what it is, and what it should be – has been a topic of prior empirical measurement and normative discussion (Abdalla et al., 2023; Michael et al., 2022); the latest wave of LLM research has raised concerns of centralization around industry models and increased industry secrecy (Bommasani et al., 2023). We identify two competing trends. On one hand, industry publishes an outsize fraction of top-cited research, including widely-used foundation models and methodological work on topics like efficiency. However, in 2023, industry is publishing *less*: Big Tech companies are accounting for 13.0% of papers in 2023, compared to 19.3% before then, with a particular drop from Google. This trend suggests that reduced openness is not only playing out in high-profile cases (e.g., the opaque GPT-4 technical report (Rogers, 2023)), but is emerging as a broader, industry-wide phenomenon of reduced publishing. Academics, meanwhile, account for a larger share of papers (particularly universities in Asia). Relative to industry, more academic work applies models to non-NLP tasks and studies social impacts.

3. **How are institutions collaborating?** Motivated by broader discussion around the risks of AI competition between nations and institutions (Cuéllar and Sheehan, 2023; Hao, 2023), we analyze the network of collaborations between the 20 most prolific institutions, all of which are either US- or China-based. We document a US/China schism: pairs of institutions which frequently collaborate are almost exclusively based in the same country. Microsoft, which collaborates with both American and Chinese universities, is the one exception. We also find that while industry-academic collaborations are common, rather than bridging differences, these papers skew significantly towards the topics usually pursued by industry. Collaborations between multiple companies are rare.

How might these insights inform the NLP community, policymakers, and other stakeholders in the future of LLM research? First, our analysis of topics and authors reveals that LLMs are increasingly being applied to diverse fields outside of core NLP. Researchers performing interdisciplinary work should involve domain experts in both NLP and the other areas of study (e.g. education, medicine, law); community leaders should reflect on how publication venues and peer review processes can best accommodate interdisciplinary work. We also show that LLM research is experiencing an influx of new authors, implying heightened value of educational resources, research checklists (Magnusson et al., 2023), and other frameworks to encourage good research practice (Dodge et al., 2019; Kapoor et al., 2023). Second, we find that while industry continues to lead much of the most impactful research, large tech companies are publishing less overall. Academics lead important research on society-facing applications and harms of AI, but closed-source models hinder detailed evaluations (Rogers, 2023). Open-source datasets and models are therefore increasingly valuable, and the community should consider how to better incentivize these contributions. Third, we provide evidence of a lack of collaboration between the US and China, substantiating concerns about AI-related competition (Cuéllar and Sheehan, 2023; Hao, 2023). Institutions may have incentives that make collaboration difficult, but efforts to create consensus may help avoid unethical or risky uses of AI. By characterizing these recent changes in the LLM research landscape, our work aims to ground discussions on the policies and practices that will shape the field's future.

2 Methods

We summarize our data and methods here and provide full details in Appendix A; Table S1 lists the fields we use in our analysis. Our primary dataset consists of all 418K papers posted to the CS and Stat arXivs between January 1, 2018 and September 7, 2023. Following past ML survey papers (Fan et al., 2023; Peng et al., 2021; Blodgett et al., 2020; Field et al., 2021), we identify an analysis subset by searching for a list of keywords in paper titles or abstracts. Keyword search has the benefits of transparency, simplicity, and consistency with past work, but also has caveats; see §A.2 for further details. Our keyword list surfaces 16,979 papers; the specific terms we include are {`language model`, `foundation model`, `BERT`, `XLNet`, `GPT-2`, `GPT-3`, `GPT-4`, `GPT-Neo`, `GPT-J`, `ChatGPT`, `PaLM`, `LLaMA`}. Details about this list are in §A.2.

We define several fields for each paper in this subset. In doing so, we follow past work and conduct manual audits to assess the reliability of our annotations; however, there remain inherent limitations in how these fields are defined, as we discuss in Appendix A. We tracked a paper's primary sub-arXiv category, e.g., Computation and Language (cs.CL). For more fine-grained topic analysis, we assigned each paper one of 40 LLM-related topics (§A.3). We clustered embeddings of paper abstracts (Zhang et al., 2022; Grootendorst, 2022), then titled the clusters using a combination of LLM annotation and manual annotation. We annotated papers for whether their authors list academic or industry affiliations (§A.4). We pulled citation counts from Semantic Scholar (Kinney et al., 2023), and tracked the *citation percentile* for each paper: the percentile of its citation count relative to papers from the same 3-month window (§A.5).

3 Results

Past work has shown that the raw count of LLM papers has risen steeply (Fan et al., 2023; Zhao et al., 2023). These trends replicate on arXiv: 12% of all CS/Stat papers were LLM-related in mid-2023 (Figure S1). Compared to all other topics (Figure S2) and all words/bigrams in paper abstracts (Table S2), LLM- & generative AI-related topics and terms are growing fastest. We dissect these ongoing changes by studying the topics, authors, and institutions that are accounting for them.

3.1 Which topics and authors are driving the growth of LLM research?

3.1.1 How have topics shifted in 2023?

We begin by analyzing the changing topic distribution of language modeling research – taxonomized by sub-arXiv category and semantic clusters – to identify which threads within LLM research are expanding and shrinking fastest.

LLM papers increasingly involve societal impacts and fields beyond NLP. For a coarse analysis of where LLMs are growing fastest, we use a paper's designated primary sub-arXiv category. We rank sub-arXivs by how quickly their proportion of LLM papers is increasing, i.e., according to the ratio $\frac{p(\text{LLM paper}|\text{paper on sub-arXiv \& 2023})}{p(\text{LLM paper}|\text{paper on sub-arXiv \& pre-2023})}$. Figure 1 displays all sub-arXivs (with at least 50 LLM papers) sorted by their 2023 to pre-2023 ratios. Computers and Society (cs.CY) ranks first, with a ratio of 20×: in 2023, 16% of its papers are about LLMs, compared to just 0.8% pre-2023. This society-facing work ranges widely, including discussions of the impacts of LLMs on education (Kasneci et al. 2023; Chan and Hu 2023, inter alia), ethics and safety (Ferrara, 2023; Sison et al., 2023), and law (Henderson et al., 2023; Li, 2023). Other sub-arXivs with both rapid growth and at least 10% prevalence of LLM papers in 2023 include HCI (up to 10% of all papers in 2023), AI (16%), and Software Engineering (19%). Strikingly, 55% of Computation and Language (cs.CL) papers are LLM-related, but due to its already-large fraction before 2023 (29%), its rate of increase ranks last. LLMs are clearly impacting much of CS research beyond NLP, especially in society- and human-facing fields.

The fastest-growing LLM topics cover applications, capabilities, and methods. To study topic shifts at a more granular level, we observe the changing distribution of the 40 LLM-related topics with which we annotated the corpus. Since these topics are learned only on the LLM paper distribution, they are more specific than sub-arXivs, which span all of Computer Science. Figure S3 lists the five fastest growing and shrinking topics in 2023 according to $\frac{p(\text{topic} \mid \text{published in 2023})}{p(\text{topic} \mid \text{published pre-2023})}$, and the results corroborate the sub-arXiv analysis (full results in Table S3). The fastest-growing topic is "Applications of LLMs/ChatGPT", which has risen from 0.9% of LLM papers before 2023 to 7% in 2023, an 8× increase. This cluster of papers,