AI DISCLOSURES PROJECT

WORKING PAPER SERIES

Real-World Gaps in AI Governance Research

AI safety and reliability in everyday deployments

Ilan Strauss
Program Director
Social Science Research Council

Isobel Moure
Program Associate
Social Science Research Council

Tim O'Reilly
Program Director
Social Science Research Council

Sruly Rosenblat
Program Associate
Social Science Research Council

100 YEARS | SOCIAL SCIENCE RESEARCH COUNCIL

WORKING PAPER NO. 2
2025/04

About The AI Disclosures Project

Led by technologist Tim O'Reilly and economist Ilan Strauss, *the AI Disclosures Project* addresses the potentially harmful societal impacts of AI's unrestrained commercialization. By improving corporate and technological transparency and disclosure mechanisms, it aims to ensure that economic incentives don't compromise safety or equity, or foster excessive risks. Disclosures are vital for well-functioning markets yet remain lacking in AI. Just as financial disclosure standards fostered robust securities markets, standardized AI disclosures can build trust, expedite adoption, and spur innovation. Through research, collaboration, and policy engagement, *the AI Disclosures Project* aims to develop a systematic framework for meaningful "Generally Accepted AI Management Principles." The project is generously funded by the Omidyar Network, The Alfred P. Sloan Foundation, and Patrick J. McGovern Foundation.

ISSN: 3067-1361
DOI: 10.35650/AIDP.4112.d.2025

This working paper can be referenced as follows:

Ilan Strauss, Isobel Moure, Tim O'Reilly and Sruly Rosenblat. "Real-World Gaps in AI Governance Research: AI Safety and Reliability in Everyday Deployments." SSRC AI Disclosures Project Working Paper Series (SSRC AI WP 2025-04), Social Science Research Council, April 2025. https://www.ssrc.org/publications/real-world-gaps-in-ai-governance-research/

Real-World Gaps in AI Governance Research

AI safety and reliability in everyday deployments

Ilan Strauss[1,2], Isobel Moure[1], Tim O'Reilly[1,3], and Sruly Rosenblat [*1]

[1]AI Disclosures Project, Social Science Research Council
[2]Institute for Innovation and Public Purpose, University College London
[3]O'Reilly Media

Abstract

Drawing on 1,178 safety and reliability papers from 9,439 generative AI papers (January 2020 through March 2025), we compare research outputs of leading *AI companies* (Anthropic, Google DeepMind, Meta, Microsoft, and OpenAI) and *AI universities* (CMU, MIT, NYU, Stanford, UC Berkeley, and University of Washington). We find that Corporate AI research increasingly concentrates on pre-deployment areas — model alignment and testing & evaluation — while attention to deployment-stage issues, such as model bias, has waned as commercial imperatives and existential risk concerns have taken precedence. We identify significant research gaps in high-risk deployment domains, including healthcare applications, commercial and financial contexts, misinformation, persuasive and addictive features, hallucinations, and copyright usage in training and inference. Without concerted efforts to enhance external observability into AI's deployment, the growing concentration of AI research with corporations could deepen knowledge deficits in these critical deployment areas. We recommend measures to expand external researcher access to deployment data and improve systematic observability of AI systems' in-market behaviors.

Keywords: *AI Research*, *AI Alignment*, *AI Interpretability*, *commercialization risks*, *cloud providers*, *AI model developers*.

[*]We gratefully acknowledge funding support from The Alfred P. Sloan Foundation, the Omidyar Network, and the Patrick J. McGovern Foundation. We extend our appreciation to the people we have had conversations with that have shaped our thinking. *Contact:* istrauss@ssrc.org / ilan@aidisclosures.org. *Code and Data:* https://github.com/AI-Disclosures-Project/Real-World-Gaps-in-AI-Governance-Research/.

Contents

1 Introduction

As generative AI becomes integrated into every facet of our work and social lives, there is an urgent need to understand the performance and impact of AI products in such commercial "post-deployment" contexts (Chayes, Cuèllar, and Li 2025; Weidinger, Raji, et al. 2025). Yet corporate research, now increasingly dominant, focuses on AI risks in *pre-deployment* laboratory settings through model alignment and testing (Figure 4).[1] User, system, and society-level impacts remain neglected.[2]

Unless AI governance research follows AI systems into the real world, areas currently considered highest risk by AI companies themselves will remain underexplored. These include model persuasiveness, emergent behaviors from reinforcement learning exploitation ('reward hacking'), and misinformation (Phuong et al. 2024; Weidinger, Barnhart, et al. 2024; Jaech et al. 2024b). De-prioritization of *research* into such areas both impedes developing industry-wide best *practices* for deployed AI systems and confines essential AI safeguards to siloed corporate efforts, limiting knowledge diffusion and public accountability.

Growing corporate concentration in AI research risks exacerbating these deficiencies. The commercial 'AI race' prioritizes an engaging user experience over broader societal impacts (Horwitz and Wells 2025). Evidence of this shift includes corporate research teams becoming tightly integrated with product teams (Wiggers 2025), research findings increasingly kept internal (Heikkilä and Morris 2025) (Figure 3), and alignment research overlooking dangerous side-effects, such as sycophancy and degraded answer quality (Amodei and Clark 2016; Sharma et al. 2023; Denison et al. 2024; Zeff 2025).

Method

We analyze AI governance research using a dataset of 1,178 safety and reliability papers from 9,439 generative AI papers written by five dominant *AI companies* (Anthropic, Google DeepMind, Meta, Microsoft, and OpenAI), and six prominent *AI research universities* (Carnegie

[1] AI alignment covers 'post-training' interventions, fine-tuning & reinforcement learning from human and AI feedback.
[2] AI companies do revise their models based based on red-teaming and user experience feedback (Jaech et al. 2024a).

Mellon University (CMU), MIT, New York University (NYU), Stanford, UC Berkeley, and University of Washington) between January 2020 and March 2025. We call these two groups 'Corporate AI' and 'Academic AI', respectively. Our dataset combines generative AI research papers from Anthropic and OpenAI's websites (Delaney, Guest, and Williams 2024) with OpenAlex's database. We define AI governance research as technical and applied safety and reliability research pre- and post-deployment. In conjunction with OpenAI's o3-mini, we determine if papers are "safety & reliability" research, and then classify them into one of eight sub-categories. We also conduct separate 'regex' key word searches in paper abstracts and titles for high-risk deployment domains (medical, finance, commercial, & copyright) and capabilities (misinformation, disclosures, behavioral, & accuracy).

Core Findings

(1) ***AI governance research is highly concentrated within a handful of uniquely resourced and integrated AI tech companies, with a disproportionately influential research impact***. Anthropic, OpenAI, and Google DeepMind each have far more citations for their AI safety & reliability work than any of the major U.S. academic institutions we track. Google DeepMind has more citations for its general generative AI research than the top four AI academic institutions combined.

(2) ***As leading AI companies race to commercialize powerful AI systems, their research priorities are increasingly shaped by business incentives rather than by comprehensive risk assessments and mitigations***. Most of the corporate governance research we review focuses on model performance divorced from its applications. Ethics & bias research – needed to understand systematic, unjustified differences in LLM behavior or outputs – now only receives attention from academic researchers.

(3) ***Corporate AI labs severely neglect deployment-stage behavioral and business risks***. Only 4% of Corporate AI papers (6% Academic AI) tackle high-stakes areas like persuasion, misinformation, medical & financial contexts, disclosures, or core business liabilities (IP violations, coding errors, hallucinations) – despite emerging lawsuits showing these risks to already be material.

Policy Considerations

To guard against commercialization-driven risks, third-party researchers (and auditors) need data on AI systems operating in real-world environments. Commercial incentives drive innovation but also foster corporate risk-taking, potentially lowering safeguards when they conflict with profit-maximizing business models (Horwitz and Wells 2025; Edwards 2025). *Post-deployment monitoring research is therefore publicly vital but currently limited to piecemeal AI incident databases* (Marchal et al. 2024; Willison 2024; Mylius 2024), *old or overly aggregated user-LLM chat data* (Tamkin and al. 2024; Zhao et al. 2024), and public testing of models. Real-world visibility into the effects of AI systems is negligible.

Structured access is needed into deployed AI systems' telemetry data and artifacts to systematically analyze real-world risks and harms. Monitoring and evaluation of LLMs in real-world environments is now essential to quality assurance (QA), as in 'LLMOps' (Aryan 2024). But the data used for this is the preserve of corporate *practice*, resulting in society losing essential insight into AI's ongoing risks and harms. Disclosure of AI system *telemetry data* (logs, traces, & business metrics) and LLM model *data artifacts* (e.g., training/fine-tuning datasets) may expose corporations to liability. But emerging LLM monitoring frameworks – such as those from LangSmith, Langfuse, OpenTelemetry, & Weights and Biases – make structured & standardized external API access for researchers increasingly feasible. Liability safe harbors (Longpre et al. 2024; Arcila 2025) are likely required to support purpose-built external access; otherwise, deployment research will have to rely on public-private partnerships.

Literature and Roadmap. Important papers in AI research classification are Toner and Acharya (2022), Farber and Tampakis (2023), Cottier, Besiroglu, and Owen (2023), Klyman et al. (2024) – and most recently Delaney, Guest, and Williams (2024), which addresses pre-deployment technical AI safety research only. Next, Section 2 motivates our study's focus on AI's deployment, and describes our data and method (Appendix 6); Section 3 presents our key findings; Section 4 makes some policy suggestions; and Section 5 concludes.

2 Motivation, Data, and Methods

The research presented in this paper is motivated by three observations:

(1) There is a growing disconnect between the theoretical research being prioritized at the major corporate AI labs, which examines AI models in isolation, and the growing need for research on how AI systems function in real-world deployment contexts where their outputs vary greatly by prompt, context, and implementation (Strauss and O'Reilly 2024a; Anthropic 2025a; J. Cheng et al. 2025).

(2) Commercial activity is a major source of risk in post-deployment AI systems, yet those in the best position to monitor and understand those risks have economic and reputational incentives to underplay them, rather than conduct transparent research on emerging problems.

(3) The vast preponderance of AI research today is carried out by corporations, and public researchers have limited access to the data needed to assess risks during real-world deployment.

This paper therefore examines the critical gap between Corporate AI's research priorities and the real-world governance challenges emerging from commercial AI deployment, arguing for increased independent research access and transparency requirements to address these mounting concerns. We motivate this further below in Sections 2.1, 2.2, and 2.3.

2.1 Pre- versus post-deployment research

Without research on AI safety as practiced in the wild, we are flying blind. Research into model safety, reliability, and other AI governance that only examines the behavior under the controlled conditions of the AI lab and model developer is fundamentally insufficient. An AI model's risks and safeguards in practice often differ significantly from those in theory (Horwitz and Wells 2025; Edwards 2025), and these differences emerge through multiple deployment factors:

Deployment environments dramatically alter model behavior. LLMs' outputs vary greatly by prompt and context, requiring assessment of impacts over time arising from repeated use, the differentiated impact of fine-tuned applications, and the risks that arise from how LLMs are accessed and deployed (Strauss and O'Reilly 2024a,b). AI-driven search, coding assistants, chatbots, recommendation engines, and advertising all rely on extensive scaffolding that is part of AI's deployment stage, but may be absent from model evaluations and reliability research conducted pre-deployment.

API access introduces new risk vectors. There is a critical distinction between models as deployed directly by their developers (e.g., the user-facing ChatGPT or Claude applications) and models accessed via API by third-party developers. Much of the fine-tuning, scaffolding, and guardrails present in user-facing apps may not be in place when a model is accessed by API. Safety becomes explicitly the responsibility of the developer (OpenAI 2024a). While model developers provide guidance on implementing guardrails (OpenAI 2023), and third-party tools exist to help developers (Weights & Biases 2025), there is little to no published research into how widely or how well these guardrails are being implemented. This gap becomes increasingly dangerous in the emerging ecosystem of AI agents and other forms of distributed and cooperating AI systems. For example, Anthropic's privacy-preserving audit system called 'Clio', monitors end-user interactions within the consumer app but provides no coverage for enterprise traffic flowing through the API (Tamkin and al. 2024).

Infrastructure differences create varied risk profiles. Significant differences exist between AI applications deployed on the public cloud infrastructure of companies such as Amazon, Google, and Microsoft, and custom models (potentially based on open weight models such as Llama or DeepSeek) that are hosted in private data centers. Each deployment architecture introduces unique security, reliability, and governance challenges (Wilson 2024) that remain largely unresearched outside corporate environments.

Other critical post-deployment components affecting safety and reliability include: (i) Orchestration primitives that route information among users, models, and external systems; (ii) Data-retrieval layers such as RAG to supply knowledge to the model beyond its training

corpus; (iii) Safety and guardrail services that enforce company policies through moderation models and toxicity filters; and (iv) Observability and evaluation stacks ("LLMOps") that track quality, surface user feedback, and guide iterative improvement (Aryan 2024).

2.2 Why commercial incentives may drive research gaps

A structural misalignment exists between corporate profit incentives and rigorous safety research on deployed AI systems. Economic incentives may preclude corporate AI labs from thoroughly researching or publicizing findings that could negatively impact their products' market adoption or regulatory treatment.

Pattern of harm emergence and inadequate response. Emerging legal cases highlight this misalignment, serving as early warning signals about the inadequacy of leaving deployment safety research primarily to commercial AI labs, where: (1) Real-world harms emerge from deployed systems, (2) Companies respond with minimal changes or even counterproductive measures, and (3) Research focus remains predominantly on theoretical rather than applied risks. These torts provide a useful guide to what AI safety research to prioritize, showing what requires urgent analysis and monitoring (Spicer et al. 2024; Hughes 2025).

Character.ai faces lawsuits over 'addictive-by-design' bots allegedly encouraging self-harm among teenagers who formed romantic relationships with the AI (Spicer et al. 2024). Despite this evidence, Meta subsequently expanded permissions to allow explicit content for romantic role-play with its AI bots (Wells, Horwitz, and Seetharaman 2025). OpenAI removed impersonation restrictions for real-life figures with its Sora image generator, effectively enabling deepfakes (Mantzarlis 2025). Meanwhile, nearly 30 lawsuits target AI model developers over copyright infringement (Knibbs 2025), and AI hallucinations in legal content have created significant liability risks (Surani and Ho 2024; Merken 2025).

Misaligned research priorities. Corporate AI labs demonstrate a concerning disconnect between their research focus and documented real-world harms. The risk focus in sporadic AI company disclosures centers almost exclusively on *malicious use* (harmful intent), while ignoring commercial (profit-driven) uses that may cause equivalent harm (OpenAI 2024b;

Microsoft Corporation 2024; Anthropic 2025b; OpenAI 2025).

Anthropic's recent initiatives exemplify this misalignment. While announcing model interpretability work to find risks based on a "model's inner workings" (Amodei 2025) and testing Claude's values (S. Huang et al. 2025), Anthropic simultaneously documented actual malicious uses of Claude, including personalized recruitment fraud, malware development, credential scraping, and management of social media bot networks for political influence operations (Anthropic 2025b). The report noted: "As agentic AI systems improve we expect this trend [semi-autonomously orchestrated complex abuse systems] to continue." Yet these documented risks have not triggered proportionate research investment into post-deployment safeguards.

2.3 Data access challenges for independent research

The third critical factor driving the current research gap is the profound data access asymmetry between corporate and independent researchers. While corporations have complete visibility into their deployed models' behaviors, usage patterns, and failure modes, independent researchers face significant barriers to accessing equivalent data.

Asymmetric information access. Corporate AI labs have exclusive access to critical data including: (1) User interaction logs indicating how models respond to varied prompts across populations, (2) Safety incident reports documenting specific failure modes, (3) Fine-tuning datasets and algorithms used to shape model behavior, and (4) Internal evaluation metrics tracking performance across safety and reliability dimensions. This information asymmetry makes independent verification of safety claims and research nearly impossible.

Limited transparency mechanisms. Current transparency initiatives remain inadequate for enabling robust independent research. Model cards provide limited high-level information, API access is restricted and often fails to show safety-critical internals, and academic partnerships typically involve highly constrained access with corporate approval requirements for publication.

Regulatory implications. As AI systems become more deeply integrated into critical

infrastructures and decision systems, the absence of independent assessment mechanisms grows increasingly problematic from a regulatory perspective. Other regulated industries with substantial public safety implications, such as pharmaceuticals, automotive, and aviation, have established independent testing regimes and mandatory disclosure requirements that have no equivalent in AI development (O'Reilly 2024; Dillon et al. 2024).[3]

Growing corporate concentration in AI research risks exacerbating these oversight deficiencies, such that public research access has an essential role to play in addressing these gaps. Without targeted interventions to enhance independent research capabilities, our understanding of deployed AI risks will continue to lag behind the rapid pace of commercial development and deployment.

2.4 Data collection and sample construction

We construct a large dataset of 1,178 AI safety and reliability governance papers from a total of 9,439 generative AI papers published between January 2020 and March 2025. This sample includes research from both leading corporations (Anthropic, Google DeepMind, Meta, Microsoft, and OpenAI) and academic institutions (Carnegie Mellon University, Massachusetts Institute of Technology, New York University, Stanford University, University of California Berkeley, and University of Washington), chosen for their significant research contributions in the field.

Table 1. Research Dataset (by Type)

	Academic AI	**Corporate AI**
Safety & Reliability	795	383
All Generative AI	6,104	2,157

Note: Total unadjusted research papers and notes by research group, divided into 'safety & reliability' and all generative AI research, January 2020 through March 2025. OpenAlex and scraped data from Anthropic and OpenAI. When adjusted for relative authorship, the sample size declines by around two-fifths for papers and citations – Table 4.

Our research analyzes AI safety & reliability papers with an author from at least one of

[3] See also Lenhart and Myers West (2024).

the above academic and corporate institutions. This sample likely underestimates Corporate AI's research impact as we do not manually scrape research paper data from Meta's website.

In practice, paper numbers and citation counts used for much of the analysis conform more closely to Table 4 (Appendix), because we adjust our sample for each institution's relative authorship contribution to the paper. This **fractional authorship method** allocates to each institution its *prorated* share of the paper based on its relative authorship. For example, if a paper has four authors and only two are from OpenAI, then OpenAI receives only 0.5 of the citations and 0.5 of paper 'count'. This helps adjust for the fact that many computer science papers have dozens of authors spanning multiple institutions.[4]

Our data comes from two sources: (1) **OpenAlex database**: An open-access research repository with citation data,[5] which we filter for generative AI research with authors from the major AI companies and research universities; and (2) **Company Websites**. Because OpenAlex omits papers published on company websites – but includes most ArXiv papers – we scrape Anthropic's and OpenAI's research from their websites, including from the dataset assembled by Delaney, Guest, and Williams (2024).[6] We fill in missing citation numbers and abstracts using a range of APIs and web-scraping techniques (Appendix 6.2). Abstracts and titles are used to classify papers into the various categories below so filling in missing values for these two variables is vital. We have 92 missing abstracts in our final dataset.

Definitions & Categories. Our total sample is defined as all **generative AI research**, with an emphasis on text models.[7] We count all research and research blog posts published by Anthropic and OpenAI as generative AI research, but exclude their system cards, product promotions, and blogs that only duplicate papers.

We define AI *safety & reliability* research as technical and policy research covering the

[4] We allocate only a single institutional affiliation per author, choosing first from among the corporate and academic institutions we analyze in this paper as their primary one, and otherwise selecting the first one affiliation that appears.

[5] See: https://openalex.org/.

[6] OpenAlex does not contain any papers from Anthropic.

[7] We extract research papers containing the following regular expressions in their abstract or title in OpenAlex: "language model*" OR "large language model*" OR "LLM*" OR "GPT" OR "BERT" OR "transformer" OR "generative model*" OR "foundation model*".

entire model (product) life cycle: pre- and post-deployment. This includes research identifying and reducing harms from AI, and/or implementing measures to make models more reliable or safer. This contrasts with Delaney, Guest, and Williams (2024), which focuses on pre-deployment technical research only. But given that LLMs are widely deployed in a variety of commercial contexts we would expect AI research to extend into these contexts, and so we include these. The eight sub-category definitions used to further categorize 'safety & reliability' research can be found in Appendix 6.2.

3 Findings

3.1 Corporate vs. academic generative AI research

Corporate AI has an outsized impact on generative AI research, including in safety & reliability research. Table 2 compares the *general* generative AI research outputs from AI corporations – Anthropic, Google DeepMind (owned by Google), Meta, Microsoft, and OpenAI – with research from leading AI research universities – Carnegie Mellon University (CMU), Massachusetts Institute of Technology (MIT), New York University (NYU), Stanford University, University of California Berkeley (UC Berkeley), and University of Washington.

Table 2 highlights the outsized impact Corporate AI has on generative AI research, with far higher average – and for Google DeepMind and OpenAI total – citations per paper.[8] Although Corporate AI generally publishes fewer papers than Academic AI (1,527 vs. 3,578), its impact is far greater, with 119,845 citations compared with 78,858 for Academic AI. *Google DeepMind is uniquely impactful and well resourced in AI research, with more citations (69,453) than the top four academic institutions combined*. Despite very few papers, OpenAI (64 author adjusted papers) and Anthropic's (62) general AI research is also widely impactful, judged by total citations.[9]

[8] There will also be strong interplays between Academic AI and Corporate AI research that we do not explore here. We find surprisingly little co-authorship of papers between the two groups. But one can see from hiring decisions that academic experts constantly move to corporate AI research labs and back to academia.

[9] We run a regression to test if corporate AI research has a citation (impact) advantage after accounting for the eight possible sub-categories of 'safety and reliability' research that we use later on. Accounting for paper topic and whether it

Table 2. Academic vs. Corporate Generative AI Research (2020 - March 2025)

	Papers	Total Citations	Mean Cite
CMU	878	17,030	19
Stanford	828	19,701	24
MIT	607	12,276	20
University of Washington	433	13,010	30
UC Berkeley	421	9,705	23
New York University	411	7,136	17
Google DeepMind	969	69,453	72
Microsoft	369	11,973	32
Meta	64	12,584	196
OpenAI	64	17,709	278
Anthropic	62	8,127	131
Total: Academic AI	3,578	78,858	22
Total: Corporate AI	1,527	119,845	78

Note: January 2020 through March 2025. All generative AI research adjusted for authorship. Google DeepMind combines 'Google' and 'DeepMind'. Each institution's papers and citation numbers are adjusted for their 'fractional' contribution, based on the number of authors they have in the paper relative to a paper's total authors and institutions.

Figure 3 (Appendix) shows publications per year. There is some evidence of a broad-based decline in publicly available AI research published between 2023 and 2024, but it is particularly steep for Google DeepMind. Heikkilä and Morris (2025) discuss that Google DeepMind might be publishing less public research on purpose, for competitive reasons. This likely also reflects DeepMind's shift away from a pure research lab to housing the Gemini product (Wiggers 2025; Woo 2025).

Corporate AI has an even more dominant impact on AI safety & reliability specific re-

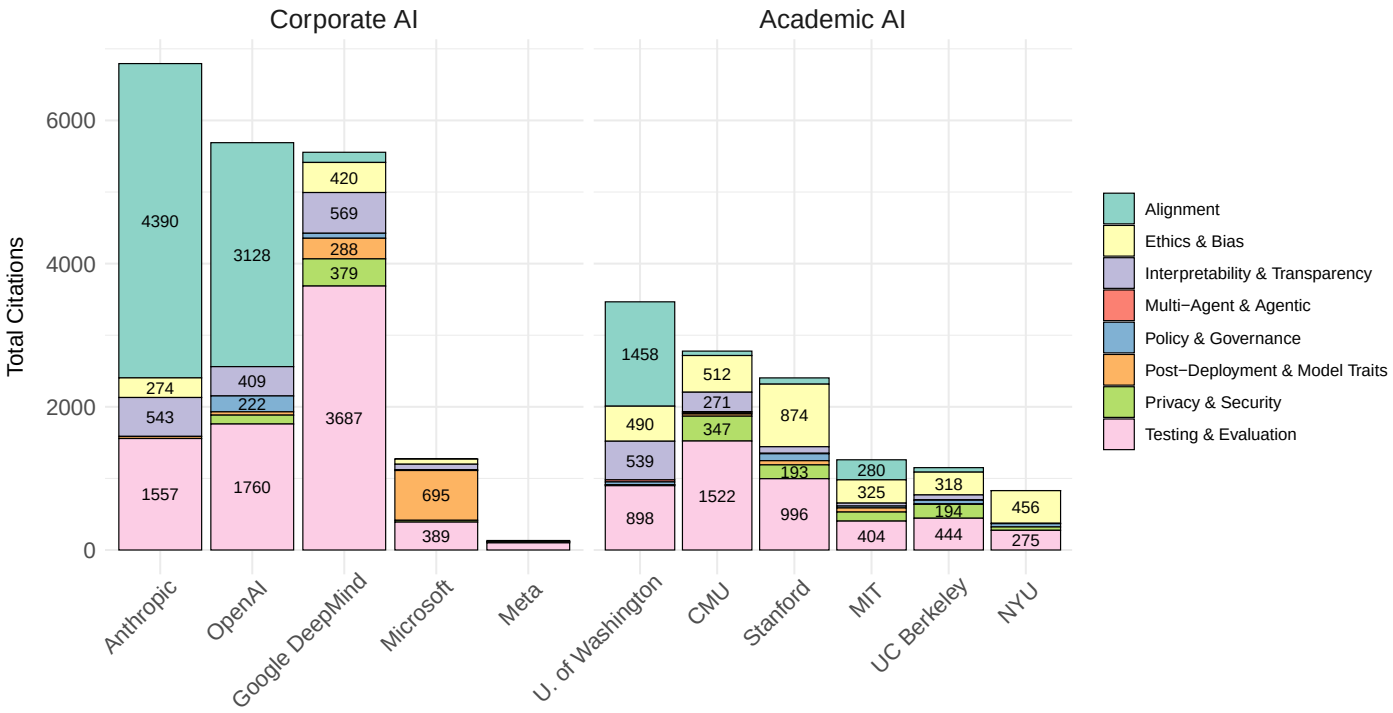
is a 'safety & reliability' paper or not, Corporate AI papers absolute probabilities of having a top 1% cited paper (versus Academic papers) increase from the sub-1–2% range up to around 9% – or 4.5x increase in the odds. NA values replaced with zeros:

$$\text{logit}\big(\Pr(\text{top01}_i = 1)\big) = \beta_0 + \beta_{s_i} + \gamma_{g_i} + \delta_{s_i,g_i},$$

where $\text{top01}_i = 1$ if paper i is in the top 1%, s_i is its safety_classification, g_i is its institution_group, β_0 is the intercept for the reference levels, β_{s_i} are safety-class effects, γ_{g_i} are institution effects (corporate or academic), and δ_{s_i,g_i} are the safety × institution interaction effects.

search, judged by total citations.[10] As shown in Figure 1, Anthropic, OpenAI, followed by Google DeepMind each have far more citations for their research in this field than established leading AI academic research institutions.

Figure 1. Total Citations for Safety & Reliability Research



Note: Fractionally adjusted for each institution's relative authorship contribution to each paper. Not showing numbers for a category with less than 150 citations. The eight categories are chosen and defined by authors and then categorized using GPT 4o-mini. See Appendix 6.3 for definitions.

Corporate AI's outsized impact on AI governance research stems from their differing research focus. Breaking this down by category, Figures 1 and 2 show that Corporate AI's research impact dominance is led by their model alignment and their testing & evaluation research, focused on model (pre-deployment) risks:

- Most *testing and evaluations* research involves pre-deployment contexts.[11] So-called 'in-the-wild' evaluations (Zhu, Yang, and Sun 2024; Bayat et al. 2025) aim to predict

[10] Though this does not account for originality of research. In many areas, academia will establish the fundamental research concepts within which corporate labs explore applications and refinements of, including for transformers, neural networks, and reinforcement learning.

[11] Our analysis of testing & evaluation papers using OpenAI's o3 Model and Claude 3.7, finds that

how a model will behave once deployed, yet they are inherently retrospective. They draw on benchmark datasets built around known failures and older model generations, leaving emergent risks invisible. Because every item must be labeled in advance, these tests are confined to what researchers already know how to measure – and to data the next model is almost certain to have seen during training. Consequently, they shed little light on unknown vulnerabilities or novel forms of misuse.

- *Applied alignment research* (mint green) helped bring Anthropic and OpenAI's research and products to prominence (Christiano et al. 2017; Stiennon et al. 2020; Ouyang et al. 2022; Glaese et al. 2022; Bai et al. 2022).[12]

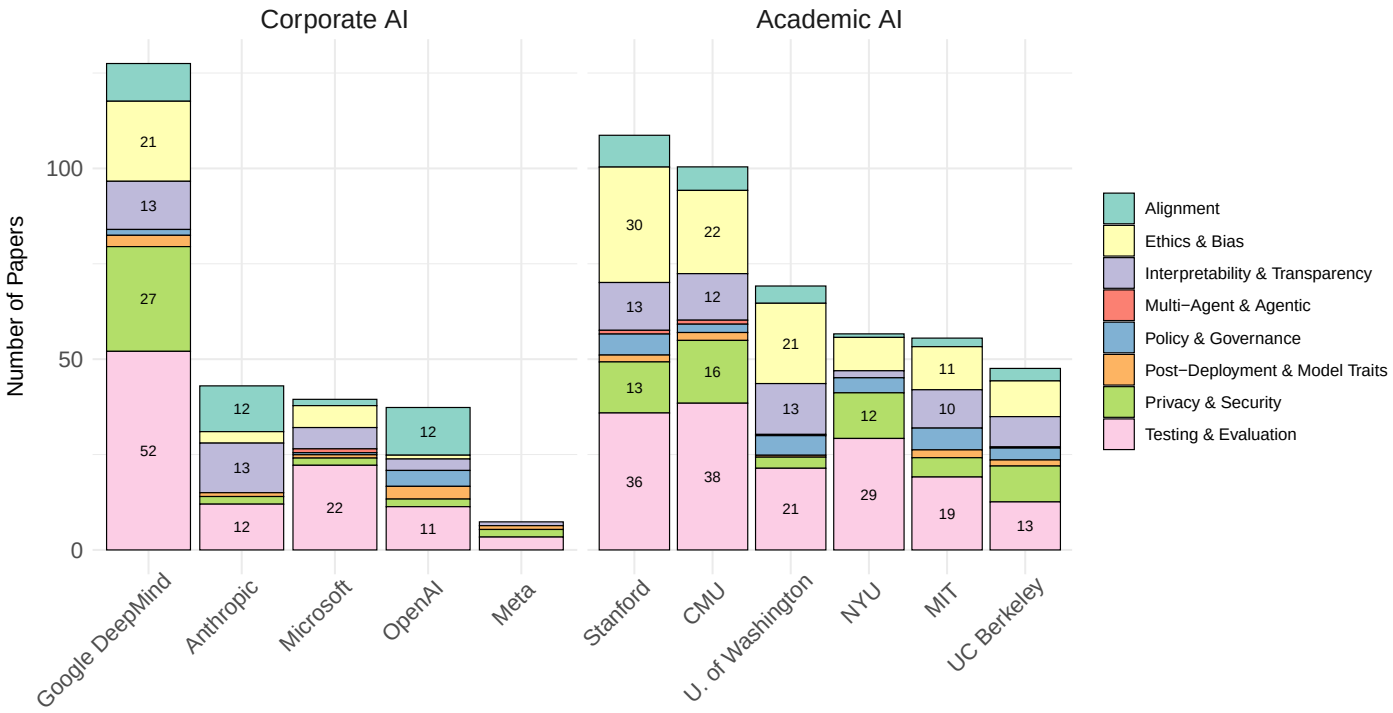
- *Research in the ethics & bias* in generative AI (yellow) is far more prominent within Academic AI's research impact (citations) than Corporate AI. Ethics & bias research includes some esoteric work in our sample, but also essential efforts to detect and explain systematic, unjustified errors (or disparities) in model behavior (predictions) that correlates with race, gender, income, education, age, language, geography, and other attributes. Reports on AI bias in medical triage, hiring, credit scoring, and in 'LLMs as a judge' motivates for why these errors are vital to study (Demchak et al. 2024; G. H. Chen et al. 2024).

This shift in research emphasis broadly confirms earlier findings by Delaney, Guest, and Williams (2024) and Toner and Acharya (2022).

only around 15-35% of testing & evaluation deal substantively with post-deployment issues. GPT defined post-deployment as involving real-world telemetry, user-study, or live-monitoring work: https://chatgpt.com/share/680279a5-f6f4-800f-85ec-2dd9f39f1ab6. Both had a large portion of papers as unclassified. Claude allocated most unclassified to pre-deployment when pushed https://claude.ai/public/artifacts/0440fef2-c030-45a8-ba50-427d3268b714 and https://claude.ai/chat/d9a32859-a725-4efb-9aac-3111ef75901f. Both used a combination of word and word combination searches within semantic search, using each paper's abstract and title. The split was roughly even between pre- and post-deployment for Academic and Corporate AI research in this area.

[12] Ouyang et al. (2022) seems to be omitted from our data since it has 13,000 citations with exclusively OpenAI authorship. We have an earlier version in our dataset, as 'Aligning language models to follow instructions' (05wx9n238 = ror id), but with no citation and other information.

Figure 2. Number of AI Safety & Reliability Papers



Note: Total number of papers fractionally adjusted for authorship. See note above. Numbers are rounded. Not showing numbers for categories with less than 10 (fractionally adjusted) papers.

Corporate AI's research influence extends to how post-deployment problems are framed. For example, Corporate AI increasingly approaches bias as a pre-deployment model personality issue, rather than a post-deployment (practical) statistical issue. This is reflected by them giving greater consideration to the existential risks from a model's autonomy – and even a model's consciousness & values (S. Huang et al. 2025; Anthropic 2025c; Witten 2025).[13] Yet a generative model's 'bias' is traditionally considered to be a function of its pre-training or post-training data, its weights, or exact fine-tuning algorithms.

Lastly, Figure 2 shows that when not accounting for research impact (citations) – looking just at total papers written (adjusted for authorship contributions) – Corporate AI's research dominance subsides, except for Google DeepMind, who still publishes more papers than any other academic research lab.[14] We show Corporate AI's research focus in greater detail in

[13] Thereby "anthropomorphizing inert weights" (Zoeller 2025; Khan, Casper, and Hadfield-Menell 2025)

[14] A similar topic emphasis is evident but now with Google DeepMind's research into privacy and security being evident,

Figure 4 (Appendix).

3.2 Post-deployment research gaps

Table 3 highlights minimal AI research in post-deployment contexts and high-risk areas, especially among Corporate AI. Only 217 Academic AI papers and 67 Corporate AI papers (adjusted for authorship) cover these high-risk areas, representing just 6% of academic and 4% of corporate papers and citations. The table breaks down papers by contexts (medical, commercial, finance) and risks (misinformation, behavioral issues, disclosure requirements, and business liabilities).

Many high-risk areas are especially underrepresented in Academic AI research, highlighting the importance of non-private research. While the usual ratio is 2.5 academic papers for every 1 corporate paper (3,578 vs. 1,527), the gap widens to 3 or 5 times for misinformation risks (53 vs. 8 papers) and medical contexts (57 vs. 9 papers).

Business and behavioral risks remain significantly under-researched. Business risks like intellectual property (IP) violations, liability for coding errors, and misinformation are rarely addressed despite early lawsuits indicating their significance (Gifford 2018). Similarly, behavioral risks of AI systems influencing human behavior receive minimal attention (Phuong et al. 2024). System cards acknowledge persuasion risks without corresponding safeguards (Jaech et al. 2024a; Phuong et al. 2024).[15] Of our sample, just 6 Corporate AI and 16 Academic AI papers address behavioral topics, with none covering addiction and relationship-forming risks despite known concerns (Ibrahim et al. 2024; Turkle 2024).

reflecting commercial incentives to operationalize its product through secure cloud and related deployments. Academic work in fact leans less towards safety & reliability (12%) compared to Corporate AI research (16%) of papers, both adjusted for authorship (not shown in Figure).

[15] For more see Weidinger et al. (2021) and Ngo, Chan, and Mindermann (2022).

Table 3. Generative AI Research Papers by Risk Areas and Context

	Papers		By Citations		% Safety
Risk Area	*Academic AI*	*Corporate AI*	*Academic AI*	*Corporate AI*	
Medical	53	9	880	1,239	26%
Misinfo	53	8	1,385	548	38%
Accuracy	28	24	282	971	55%
Finance	36	9	1,737	1,748	18%
Disclosure	15	7	226	122	85%
Behavioral	16	6	198	581	38%
Commercial	16	5	279	39	28%
Copyright	3	2	41	19	94%

Note: Author adjusted. Keyword matching in abstract or title using regex: **Disclosure** includes model cards, data cards, auditing/audits, model standards, evaluation standards, and testing standards; **Medical** includes hospital(s), health insurance, and clinician(s); **Commercial** includes adverts/advertisements, marketing, hiring, and recruiting; **Misinfo** includes spam, phishing, disinformation, and misinformation; **Finance** includes finance/financial; **Behavioral** includes sycophant(s)/sycophantic, sycophancy, addictive, persuasion(s)/persuasive, and reward-hacking; **Copyright** includes access violations, copyright violations, content attribution, dataset licensing, data attribution, copyrighted material, copyright law, C2PA, and the Content Authenticity Initiative; **Accuracy** includes hallucinations, coding errors, coding inaccuracy, factual inaccuracy, factual error.

Research on disclosures, auditing, and standards — preventing companies from "grading their own homework" — is also sparse. Ganguli et al. (2023) offers one of few examples detailing lessons from voluntary external auditing.

Actual AI safety practices are largely absent in post-deployment research. Alignment research (Guan et al. 2024) ties safety to the model itself rather than product architecture involving moderation, filtering, and security systems. Only four papers in our database address moderation and filtering practices (Hsieh et al. 2023; Y. Zhang et al. 2023; Qiao et al. 2024; Luo et al. 2025).

Post-deployment considerations do appear in Corporate AI research but remain peripheral. Notable examples include DeepMind's socio-technical approach (Weidinger et al. 2021; Weidinger, Rauh, et al. 2023), Microsoft's red-teaming & mitigations research (Abdali et al. 2024; Bullwinkel et al. 2025), Anthropic's work on reward hacking and sycophancy

(Sharma et al. 2023; Denison et al. 2024; Perez et al. 2022), regulatory markets research (Clark and Hadfield 2019; Hadfield and Clark 2023), and standard setting (Anderljung et al. 2023).

Discussion of causes. *Commercial incentives and "x-risk" ideology shape research priorities*. Early OpenAI work, for example, addressed post-deployment evaluations (Radford et al. 2019; Solaiman et al. 2019; Brundage et al. 2022), but this focus has shifted toward existential risks and profitable applications, exemplified by their image generator now allowing creation of brands and real people (Edwards 2025).

The shift in corporate labs stems from both commercial motivations and ideological influences. Alignment research and evaluation work share origins in existential risk philosophy (Yudkowsky 2002; Bostrom 2014; Yudkowsky 2020), which emphasizes low-probability but potentially catastrophic future scenarios. In this philosophy, the model itself is the source of risk due to its potentially autonomous capabilities, prioritizing speculative future dangers over immediate post-deployment concerns. This perspective has shaped corporate risk frameworks and appears now in emerging research on AI model 'values' and consciousness (S. Huang et al. 2025; Roose 2025). This philosophy permeated Corporate AI research (Olson 2024) and eventually academia too, through centers like Berkeley's Center for Human-Compatible AI (CHAI) and Stanford's Institute for Human-Centered AI (HAI).[16]

4 Policy Discussion

The commercial rollout of large-scale AI systems has created an information asymmetry that makes rigorous, public-interest oversight almost impossible. Firms now operate powerful models behind proprietary interfaces, collecting exhaustive telemetry — everything from prompts and error traces to user-level engagement metrics — but that data seldom leaves the corporate dashboard. Independent scholars must rely on studying "in-

[16] Stuart Russell at CHAI and Nick Bostrom's Future of Humanity Institute at Oxford connected technical alignment approaches with formal modeling of risks from advanced AI, drawing on concepts like Pascal's Wager - acting on low-probability but infinite-stakes events - and expected utility theory to address potential catastrophic outcomes.

cidents" after they spill into the press (Marchal et al. 2024; Willison 2024; Mylius 2024; Mylius and Bernadi 2024) or mining limited chat logs released by chance (Tamkin and al. 2024; Zhao et al. 2024; ShareGPT 2023). While companies have comprehensive instrumentation, external researchers work with fragmentary glimpses.

This opacity is not accidental; it is an economically rational response to litigation risk and competitive pressure. Detailed corporate logs can indicate bias, privacy leakage, or manipulative behavior — liabilities no firm wants to advertise. Yet these same *traces* – detailed records of system operations, inputs, outputs, and decision paths – are precisely what outside researchers require to measure real-world harms and propose effective safeguards.

One potential pathway is to treat AI telemetry like financial-market trade data, using a tiered disclosure regime (CFTC 0017; Martinen et al. 2018). For high-risk applications, firms would expose a secure API that streams three privacy-protected data feeds: differentially private event logs, system-operation traces, and model artifact manifests that record key metadata such as version numbers, training methods, and documented limitations. Together, this could allow external researchers to link behaviors observed in traces to the specific model characteristics that produced them.

Next, verified academics could access capped samples, while accredited auditors could obtain deeper access under NDAs, and regulators would retain subpoena-level rights. Liability safe harbors would be needed to incentivize participation from firms and from researchers (Longpre et al. 2024; Arcila 2025). This is comparable to suspicious activity reports (SARs) in banking: firms are compelled to share, researchers are protected when they probe, and misuse carries penalties.[17]

Technically, the pieces of this approach already exist. OpenTelemetry, LangSmith, Langfuse, and Weights & Biases have converged on JSON trace formats that can be versioned and rate-limited. Extending those with LLM-specific fields would allow companies

[17] Under the Bank Secrecy Act, financial institutions must file Suspicious Activity Reports (SARs) with the Financial Crimes Enforcement Network (FinCEN) when they detect transactions that may involve illicit activity. The regime provides for: (i) mandatory reporting, (ii) a statutory safe harbor shielding institutions and their personnel from civil liability for good-faith filings, (iii) strict confidentiality requirements that prohibit disclosing a SAR's existence, and (iv) civil and criminal penalties for failure to report or for misuse or disclosure of SAR information (Gadinis and Mangels 2016).

to create external access to their disclosures with minimal effort. A reference standard, similar to SOC-2 but with principles relevant to business metrics, could streamline this process and should ideally align with emerging regulatory frameworks like ISO/IEC standards and the EU AI Act.

With structured visibility into deployed systems, researchers could run studies of model bias, detect early signs of catastrophic jailbreaks, and quantify whether engagement-optimized assistants nudge users toward extreme content or addictive patterns. Policymakers would gain an empirical foundation for interventions rather than relying on headline-driven panic. Systematic telemetry access would allow AI governance research to escape speculative theory and directly shape evidence-based practices. Without addressing this systematic gap in observability, governance frameworks will remain constrained by ex-ante assessment limitations.

5 Conclusion

This paper analyzed 1,178 safety and reliability papers from 9,439 generative AI research publications (2020 through March 2025), detailing a worrying trend: as commercial deployment accelerates, research increasingly concentrates on pre-deployment areas while high-risk post-deployment research remains significantly underrepresented.

AI research has become highly concentrated within a small number of tech companies wielding disproportionate influence. Google DeepMind, Anthropic, and OpenAI significantly now drive AI's research agenda (reflected in outsized citation impacts), shaping priorities toward technical model alignment and evaluation approaches that improve performance, but with an emphasis on safety concerns that align with commercial interests.

Most concerning is the lack of attention to deployment-stage risks. Only 4% of Corporate AI papers and citations tackle high-stakes areas such as persuasion, misinformation, medical and financial contexts, or core business liabilities — even as lawsuits demonstrate these risks are already material. Widely deployed mitigations like content moderation and

telemetry-based monitoring remain virtually unresearched.

These findings suggest a governance paradox: corporations with comprehensive data on live AI systems are the least incentivized to study resulting harms publicly. Without structured access to deployment telemetry, external researchers cannot build the empirical base that regulators require.

The policy implication is clear: access to post-deployment evidence – logs, traces, and incident data – should become the norm for high-impact AI deployments. Existing observability stacks already capture these data internally; extending them to accredited researchers would impose minimal overhead while dramatically expanding the public risk-assessment toolkit. Safe-harbor provisions and tiered-access APIs can balance liability concerns with transparency.

In summary, as the field's center of gravity has migrated from university labs to corporate product groups, society's need for independent oversight has never been greater. Bridging that gap requires not just incident tracking, but continuous, structured observability of AI in the wild for governance through tiered public research, governance, and audit access.

References

Chayes, Jennifer Tour, Mariano-Florentino Cuèllar, and Fei-Fei Li (Mar. 2025). *Draft Report of the Joint California Policy Working Group on AI Frontier Models*. Tech. rep. Draft report. Joint California Policy Working Group on AI Frontier Models. URL: https://www.cafrontieraigov.org/wp-content/uploads/2025/03/Draft_Report_of_the_Joint_California_Policy_Working_Group_on_AI_Frontier_Models.pdf.

Weidinger, Laura, Deb Raji, et al. (2025). "Toward an evaluation science for generative AI systems". *arXiv preprint arXiv:2503.05336*.

Jaech, Aaron et al. (2024a). "Openai o1 system card". *arXiv preprint arXiv:2412.16720*.

Phuong, Mary et al. (2024). "Evaluating frontier models for dangerous capabilities". *arXiv preprint arXiv:2403.13793*.

Weidinger, Laura, Joslyn Barnhart, et al. (2024). "Holistic safety and responsibility evaluations of advanced AI models". *arXiv preprint arXiv:2404.14068*.

Jaech, Aaron et al. (2024b). "Openai o1 system card". *arXiv preprint arXiv:2412.16720*.

Horwitz, Jeff and Georgia Wells (Apr. 2025). "Meta's 'Digital Companions' Will Talk Sex With Users—Even Children". *The Wall Street Journal*. URL: https://www.wsj.com/tech/ai/meta-ai-chatbots-sex-a25311bf.

Wiggers, Kyle (Jan. 2025). *Google folds more AI teams into DeepMind to 'accelerate the research-to-developer pipeline'*. Accessed: 2025-01-16. URL: https://techcrunch.com/2025/01/09/google-folds-more-ai-teams-into-deepmind-to-accelerate-the-research-to-developer-pipeline/.

Heikkilä, Melissa and Stephen Morris (Apr. 2025). "DeepMind slows down research releases to keep competitive edge in AI race". *Financial Times*. Accessed: 2025-04-10. URL: https://www.ft.com/content/2ee1ffde-008e-4ea4-861b-24f15b25cf54.

Amodei, Dario and Jack Clark (Dec. 2016). *Faulty Reward Functions in the Wild*. OpenAI Blog. URL: https://openai.com/blog/faulty-reward-functions/.

Sharma, Mrinank et al. (2023). "Towards understanding sycophancy in language models". *arXiv preprint arXiv:2310.13548*.

Denison, Carson et al. (2024). "Sycophancy to subterfuge: Investigating reward-tampering in large language models". *arXiv preprint arXiv:2406.10162*.

Zeff, Maxwell (Apr. 2025). *OpenAI's New Reasoning AI Models Hallucinate More*. TechCrunch, accessed April 23, 2025. URL: https://techcrunch.com/2025/04/18/openais-new-reasoning-ai-models-hallucinate-more/.

Delaney, Oscar, Oliver Guest, and Zoe Williams (2024). "Mapping Technical Safety Research at AI Companies: A literature review and incentives analysis". *arXiv preprint arXiv:2409.07878*.

Edwards, Benj (Mar. 2025). "OpenAI's New AI Image Generator Is Potent and Bound to Provoke". *Ars Technica*. URL: https://arstechnica.com/ai/2025/03/openais-new-ai-image-generator-is-potent-and-bound-to-provoke/.

Marchal, Nahema et al. (2024). "Generative AI misuse: A taxonomy of tactics and insights from real-world data". *arXiv preprint arXiv:2406.13843*.

Willison, Steve (2024). *OWASP Top Ten*. https://owasp.org/www-project-top-ten/. Accessed: 2025-04-21.
Mylius, Simon (2024). *MIT AI Incident Tracker*. Accessed: February 6, 2025. URL: https://airisk.mit.edu/ai-incident-tracker.
Tamkin, Alex and et al. (2024). "Clio: Privacy-Preserving Insights into Real-World AI Use". *arXiv preprint arXiv:2412.13678*.
Zhao, Wenting et al. (2024). "WildChat: 1M ChatGPT Interaction Logs in the Wild". *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Bl8u7ZRlbM.
Aryan, Abi (2024). *What is LLMOps?: large language models in production*. O'Reilly Media, Inc.
Longpre, Shayne et al. (2024). "A safe harbor for AI evaluation and red teaming". *arXiv preprint arXiv:2403.04893*.
Arcila, Beatriz Botero (2025). *AI Liability Along the Value Chain*. Mozilla. URL: https://blog.mozilla.org/netpolicy/files/2025/03/AI-Liability-Along-the-Value-Chain_Beatriz-Arcila.pdf.
Toner, Helen and Ashwin Acharya (2022). "Exploring clusters of research in three areas of AI safety". *Center for Security and Emerging Technology*. URL: https://cset.georgetown.edu/wp-content/uploads/Exploring-Clusters-of-Research-in-Three-Areas-of-AI-Safety.pdf.
Farber, Michael and Lazaros Tampakis (Oct. 2023). *Analyzing the Impact of Companies on AI Research Based on Publications*. arXiv preprint. Accessed: 2025-01-10. URL: https://arxiv.org/pdf/2310.20444.
Cottier, Ben, Tamay Besiroglu, and David Owen (2023). "Who is leading in AI? An analysis of industry AI research". *arXiv preprint arXiv:2312.00043*.
Klyman, Kevin et al. (Dec. 2024). *Expanding Academia's Role in Public Sector AI*. Issue Brief. Accessed: 2025-04-21. Stanford, CA: Stanford Institute for Human-Centered Artificial Intelligence, Stanford University. URL: https://hai.stanford.edu/policy/expanding-academias-role-in-public-sector-ai.
Strauss, Ilan and Tim O'Reilly (Oct. 2024a). *AI is Entirely New, AI is Exactly the Same: Thoughts on the New White House AI Memorandum*. Asimov's Addendum.
Anthropic (Mar. 2025a). *Anthropic Economic Index: Insights from Claude 3.7 Sonnet*. Accessed: 2025-04-28. URL: https://www.anthropic.com/news/anthropic-economic-index-insights-from-claude-sonnet-3-7.
Cheng, Jingwen et al. (Mar. 2025). *REALM Dataset Dashboard*. Accessed: 2025-04-28. URL: https://realm-e7682.web.app/.
Strauss, Ilan and Tim O'Reilly (2024b). "Risk without Uncertainty? OpenAI Would Like Us to Think So..." *Asimov's Addendum*. AI model evaluations, such as those conducted by OpenAI in its GPT system cards, aim to quantify model risks but often fail to account for uncertainty. URL: https://asimovaddendum.substack.com/p/can-we-have-ai-model-risk-evaluation.
OpenAI (2024a). *Safety Best Practices*. https://platform.openai.com/docs/guides/safety-best-practices. Accessed: 2025-04-29.

OpenAI (2023). *How to Implement LLM Guardrails*. Accessed: 2025-04-29. URL: https://cookbook.openai.com/examples/how_to_use_guardrails.
Weights & Biases (2025). *Responsible AI: A Guide to Guardrails and Scorers*. https://wandb.ai/site/articles/ai-guardrails/. Accessed: 2025-04-29.
Wilson, Steve (2024). *The Developer's Playbook for Large Language Model Security*. O'Reilly Media, Incorporated.
Spicer, Katy et al. (2024). *Artificial Intelligence and the Rise of Product Liability Tort Litigation: Novel Action Alleges AI Chatbot Caused Minor's Suicide*. Accessed: 2025-01-05. URL: https://www.privacyworld.blog/2024/11/artificial-intelligence-and-the-rise-of-product-liability-tort-litigation-novel-action-alleges-ai-chatbot-caused-minors-suicide/.
Hughes, Chris (Feb. 2025). *Can we govern AI without breaking it?* Accessed: 2025-04-28. URL: https://chrishughes.substack.com/p/can-we-govern-ai-without-breaking.
Wells, Georgia, Jeff Horwitz, and Deepa Seetharaman (Apr. 2025). "Meta's 'Digital Companions' Will Talk Sex With Users—Even Children". *The Wall Street Journal*. URL: https://www.wsj.com/tech/ai/meta-ai-chatbots-sex-a25311bf.
Mantzarlis, Alexios (Apr. 2025). "OpenAI says "f**k it, we're doing impersonation now"". *Faked Up*. URL: https://fakedup.org/openai-says-fk-it-were-doing-impersonation-now/.
Knibbs, Kate (Mar. 2025). "Every AI Copyright Lawsuit in the US, Visualized". *WIRED*. URL: https://www.wired.com/story/ai-copyright-case-tracker/.
Surani, Faiz and Daniel E. Ho (May 2024). "AI on Trial: Legal Models Hallucinate in 1 out of 6 (or More) Benchmarking Queries". *Stanford HAI*. Accessed: 2025-04-28. URL: https://hai.stanford.edu/news/ai-trial-legal-models-hallucinate-1-out-6-or-more-benchmarking-queries.
Merken, Sara (Feb. 2025). "AI 'hallucinations' in court papers spell trouble for lawyers". *Reuters*. Accessed: 2025-04-28. URL: https://www.reuters.com/technology/artificial-intelligence/ai-hallucinations-court-papers-spell-trouble-lawyers-2025-02-18/.
OpenAI (Oct. 2024b). *Influence and Cyber Operations: An Update*. URL: https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations.
Microsoft Corporation (Oct. 2024). *Microsoft Digital Defense Report 2024*. Tech. rep. Accessed: 2025-04-28. Microsoft Corporation. URL: https://www.microsoft.com/en-us/security/security-insider/intelligence-reports/microsoft-digital-defense-report-2024.
Anthropic (Apr. 2025b). "Detecting and Countering Malicious Uses of Claude: March 2025". *Anthropic News*. Accessed: 2025-04-28. URL: https://www.anthropic.com/news/detecting-and-countering-malicious-uses-of-claude-march-2025.
OpenAI (Feb. 2025). *Disrupting Malicious Uses of Our Models: February 2025 Update*. Tech. rep. OpenAI. URL: https://cdn.openai.com/threat-intelligence-reports/disrupting-malicious-uses-of-our-models-february-2025-update.pdf.
Amodei, Dario (Apr. 2025). *The Urgency of Interpretability*. Accessed: 2025-04-28. URL: https://www.darioamodei.com/post/the-urgency-of-interpretability.

Huang, Saffron et al. (2025). "Values in the Wild: Discovering and Analyzing Values in Real-World Language Model Interactions". *arXiv preprint arXiv:2504.15236*.

O'Reilly, Tim (Nov. 2024). *What Auto Safety Teaches Us About AI Safety*. URL: https://asimovaddendum.substack.com/p/what-auto-safety-teaches-us-about (visited on 04/29/2025).

Dillon, Robin et al. (2024). "How AI Can Help Learn Lessons from Incident Reporting Systems". *2024 IEEE Aerospace Conference*. IEEE, pp. 1–15.

Lenhart, Anna and Sarah Myers West (Aug. 2024). *Lessons from the FDA for AI*. Research Report. AI Now Institute. URL: https://ainowinstitute.org/publications/research/lessons-from-the-fda-for-ai (visited on 04/29/2025).

Woo, Erin (Mar. 2025). "Google's AI Unit Reorganizes Product Work, Announces Changes to Gemini App Team". *The Information*. Accessed: 2025-04-18. URL: https://www.theinformation.com/briefings/googles-ai-unit-reorganizes-product-work-announces-changes-to-gemini-app-team?rc=7em78a.

Zhu, Zhiying, Yiming Yang, and Zhiqing Sun (2024). "Halueval-wild: Evaluating hallucinations of language models in the wild". *arXiv preprint arXiv:2403.04307*.

Bayat, Farima Fatahi et al. (2025). "FactBench: A Dynamic Benchmark for In-the-Wild Language Model Factuality Evaluation". *arXiv preprint arXiv:2410.22257*.

Christiano, Paul F et al. (2017). "Deep reinforcement learning from human preferences". *Advances in neural information processing systems* 30.

Stiennon, Nisan et al. (2020). "Learning to summarize with human feedback". *Advances in neural information processing systems* 33, pp. 3008–3021.

Ouyang, Long et al. (2022). "Training language models to follow instructions with human feedback". *Advances in neural information processing systems* 35, pp. 27730–27744.

Glaese, Andreas et al. (2022). "Improving alignment of dialogue agents via targeted human feedback". *arXiv preprint arXiv:2209.14375*. URL: https://arxiv.org/abs/2209.14375.

Bai, Yuntao et al. (2022). "Constitutional AI: Harmlessness from ai feedback". *arXiv preprint arXiv:2212.08073*.

Demchak, Nathaniel et al. (2024). "Assessing Bias in Metric Models for LLM Open-Ended Generation Bias Benchmarks". *arXiv preprint arXiv:2410.11059*.

Chen, Guiming Hardy et al. (2024). "Humans or LLMs as the judge? a study on judgement biases". *arXiv preprint arXiv:2402.10669*.

Anthropic (Apr. 2025c). *Exploring Model Welfare*. Accessed: 2025-04-25. URL: https://www.anthropic.com/research/exploring-model-welfare.

Witten, Zack (2025). *Measuring Models' Special Interests*. https://zswitten.github.io/2025/04/14/model-special-interests.html. Accessed: 2025-04-18.

Zoeller, Georg (Apr. 2025). *Comment on Ethan Mollick's post about model preferences and Claude's behavior*. https://www.linkedin.com. LinkedIn post, April 15, 2025. Accessed via Ethan Mollick's public post.

Khan, Ariba, Stephen Casper, and Dylan Hadfield-Menell (2025). "Randomness, Not Representation: The Unreliability of Evaluating Cultural Alignment in LLMs". *arXiv preprint arXiv:2503.08688*.

Gifford, Donald G (2018). "Technological triggers to tort revolutions: steam locomotives, autonomous vehicles, and accident compensation". *Journal of tort law* 11.1, pp. 71–143.
Weidinger, Laura et al. (2021). "Ethical and social risks of harm from language models". *arXiv preprint arXiv:2112.04359*.
Ngo, Richard, Lawrence Chan, and Sören Mindermann (2022). "The alignment problem from a deep learning perspective". *arXiv preprint arXiv:2209.00626*.
Ibrahim, Lujain et al. (2024). "Beyond static AI evaluations: advancing human interaction evaluations for LLM harms and risks". *arXiv preprint arXiv:2405.10632*.
Turkle, Sherry (2024). *Who Do We Become When We Talk to Machines?* URL: https://www.youtube.com/watch?v=yY1fGc0YR3Y.
Ganguli, Deep et al. (2023). *Challenges in Evaluating AI Systems*. Accessed: 2025-01-23. URL: https://www.anthropic.com/news/evaluating-ai-systems.
Guan, Melody Y et al. (2024). "Deliberative alignment: Reasoning enables safer language models". *arXiv preprint arXiv:2412.16339*.
Hsieh, Jane et al. (2023). ""Nip it in the Bud": Moderation Strategies in Open Source Software Projects and the Role of Bots". *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW2, pp. 1–29.
Zhang, Yiming et al. (2023). "Biasx:" thinking slow" in toxic content moderation with explanations of implied social biases". *arXiv preprint arXiv:2305.13589*.
Qiao, Wei et al. (2024). "Scaling Up LLM Reviews for Google Ads Content Moderation". *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 1174–1175.
Luo, Enming et al. (2025). "Zero-Shot Image Moderation in Google Ads with LLM-Assisted Textual Descriptions and Cross-modal Co-embeddings". *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 1092–1093.
Weidinger, Laura, Maribeth Rauh, et al. (2023). "Sociotechnical safety evaluation of generative AI systems". *arXiv preprint arXiv:2310.11986*.
Abdali, Sara et al. (2024). "Securing large language models: Threats, vulnerabilities and responsible practices". *arXiv preprint arXiv:2403.12503*.
Bullwinkel, Blake et al. (2025). "Lessons From Red Teaming 100 Generative AI Products". *arXiv preprint arXiv:2501.07238*.
Perez, Ethan et al. (2022). "Discovering language model behaviors with model-written evaluations". *arXiv preprint arXiv:2212.09251*.
Clark, Jack and Gillian K Hadfield (2019). "Regulatory markets for AI safety". *arXiv preprint arXiv:2001.00078*.
Hadfield, Gillian K and Jack Clark (2023). "Regulatory markets: The future of AI governance". *arXiv preprint arXiv:2304.04914*.
Anderljung, Markus et al. (2023). "Frontier AI regulation: Managing emerging risks to public safety". *arXiv preprint arXiv:2307.03718*.
Radford, Alec et al. (2019). *Better language models and their implications*. Accessed: 2025-01-23. URL: https://openai.com/index/better-language-models/.
Solaiman, Irene et al. (2019). "Release Strategies and the Social Impacts of Language Models". *arXiv preprint* arXiv:1908.09203. URL: https://arxiv.org/abs/1908.09203.

Brundage, Miles et al. (Mar. 2022). *Lessons learned on language model safety and misuse*. Accessed: 2025-01-23. URL: https://openai.com/index/language-model-safety-and-misuse/.

Yudkowsky, Eliezer (2002). *The AI-Box Experiment*. Accessed: 2025-02-03. URL: http://yudkowsky.net/singularity/aibox.

Bostrom, Nick (2014). *Superintelligence: Paths, dangers, strategies*.

Yudkowsky, Eliezer (2020). *The Sequences (LessWrong)*. https://www.lesswrong.com/tag/sequences. Accessed: 2025-02-03.

Roose, Kevin (Apr. 24, 2025). "If A.I. Systems Become Conscious, Should They Have Rights?" *The New York Times*. Accessed: 2025-04-28. URL: https://www.nytimes.com/2025/04/24/technology/ai-welfare-anthropic-claude.html.

Olson, Parmy (2024). *Supremacy: AI, ChatGPT, and the Race that Will Change the World*. St. Martin's Press.

Mylius, Simon and Jamie Bernadi (2024). *Scalable AI Incident Classification*. Blog post. URL: https://simonmylius.com/blog/incident-classification (visited on 12/23/2024).

ShareGPT (2023). *ShareGPT Vicuna Unfiltered*. https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered. Apache 2.0 License.

CFTC, US (17). "CFR Part 43; RIN 3038-AD08: Real-Time Public Reporting of Swap Transaction Data". *Federal Register* 77.5, pp. 1182–266.

Martinen, Michael et al. (2018). "Consolidated Audit Trail: Strategic planning and best practices". *Journal of Securities Operations & Custody* 10.1, pp. 77–83.

Gadinis, Stavros and Colby Mangels (2016). "Collaborative Gatekeepers". *Wash. & Lee L. Rev.* 73, p. 797.

6 Appendix

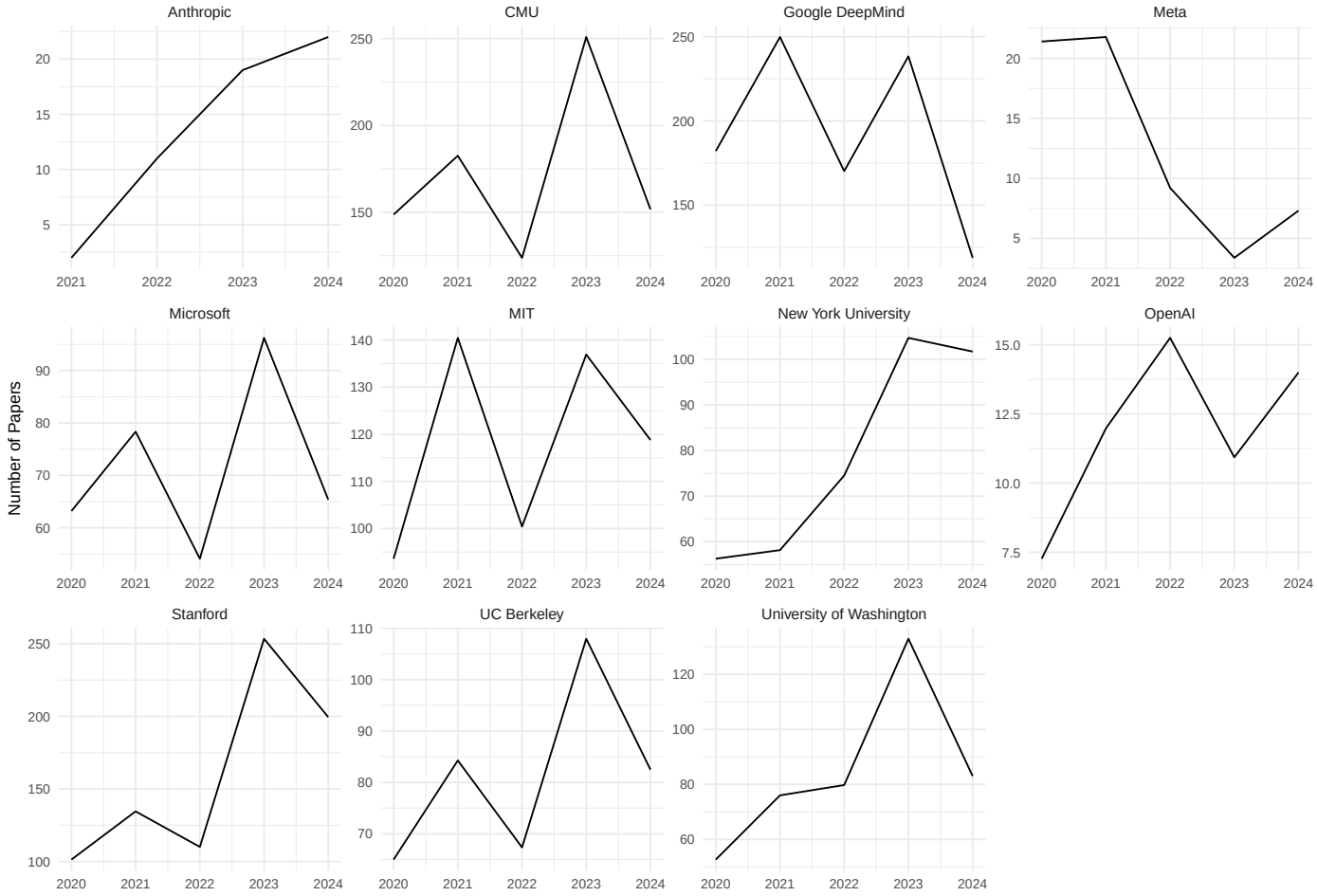
6.1 Additional Analysis

Table 4. Dataset Adjusted for Authorship: Institutional relative contributions

	Academic AI	**Corporate AI**
Safety & Reliability	438	255
All Generative AI	3,140	1,272

Note: Fractionally adjusted to account for each institution's relative contribution to each paper by number of authors relative to total authors and institutions. Divided into 'safety & reliability' and all generative AI research, January 2020 till March 31 2025. OpenAlex and scraped data.

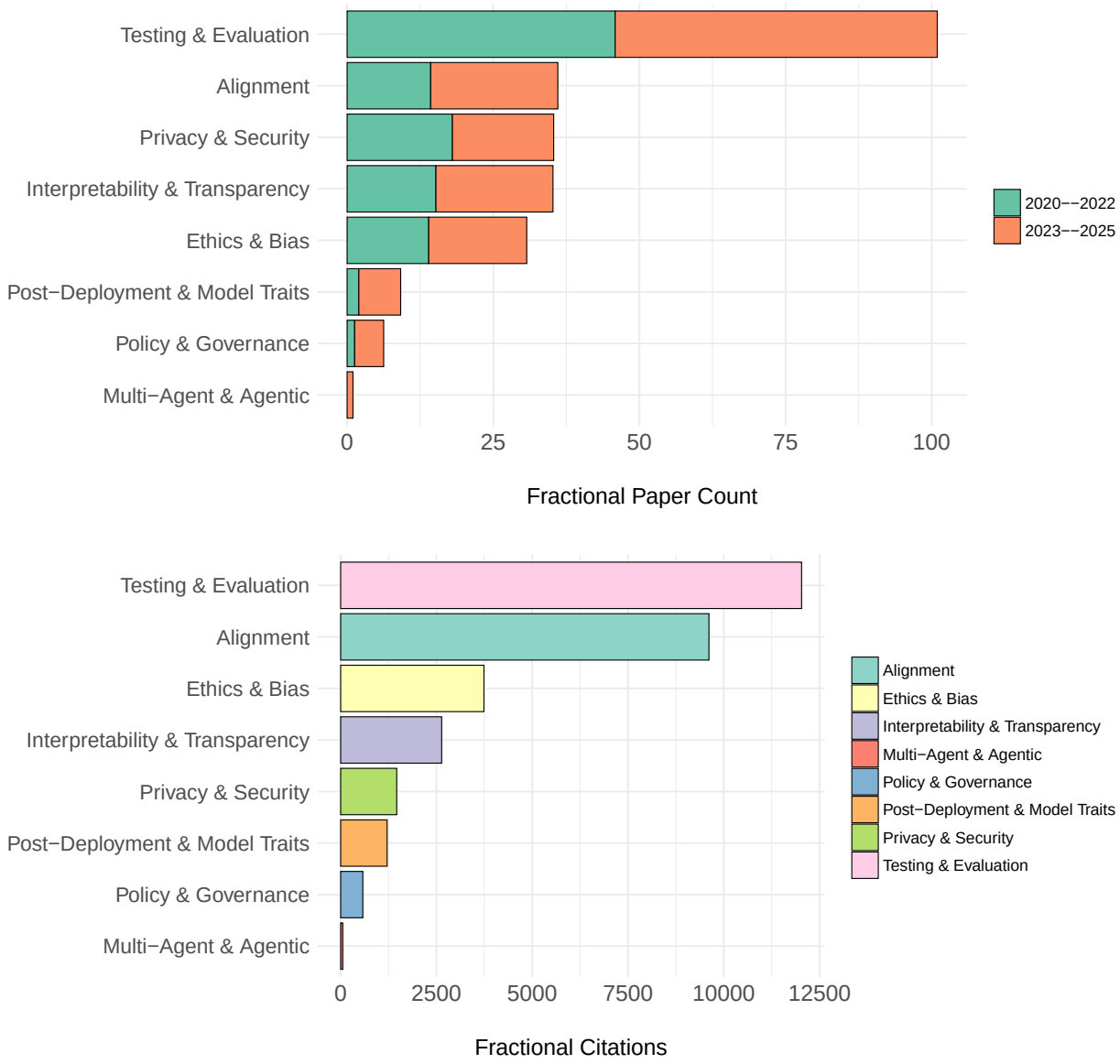
Figure 3. All Generative AI Publications by Institution (2020-2024)



Note: Y-scale differs by entity. DeepMind shows the largest absolute and relative decline (putting aside Meta for now). But Microsoft, CMU, UC Berkeley, and University of Washington also show notable declines. Fractionally adjusted to account for each institution's relative contribution to each paper by number of authors relative to total authors and institutions

Corporate AI's research focus and impact broken down into our eight AI 'safety & reliability categories' is more clearly shown in Figure 4, showing considerable concentration in testing & evaluation, and alignment work.

Figure 4. AI Governance Areas by Total Paper Numbers (by Year) - Top Graph; and by Total Citations (Fractionally Adjusted) - Bottom Graph.



Note: Adjusted for each institution's relative contribution to the paper by authorship. Data is for 2020- March 2025.

Among AI corporations, Figure 4 shows that policy & governance, as well as post-deployment risks and model traits, have consistently had a low research priority. Agentic safety & reliability research is also notably absent, despite the boom in applications in this area more recently. Several behavioral risk papers on model sycophancy (being overly

agreeable) and persuasiveness – including relatively well cited papers – were classified in alignment and other categories, so we break out these papers separately in Table 3 in the main paper. Figure 4 highlights the notable acceleration in model alignment and testing & evaluation research.

6.2 Research Dataset Construction

We rely primarily on the OpenAlex database (via the R package openalexR). We focus on a specific set of institutions — *academic* (Carnegie Mellon University (CMU), Massachusetts Institute of Technology (MIT), New York University (NYU), Stanford University, University of California Berkeley (UC Berkeley), and University of Washington) and *corporate* (Anthropic, Google DeepMind, Meta, Microsoft, and OpenAI) — by specifying each entity's ROR ID.

We retrieve from OpenAlex papers published from January 2020 through March 2025, searching for works whose titles or abstracts reference large language models and generative AI research. Our keyword filter for: "language model*" OR "large language model*" OR "LLM*" OR "GPT" OR "BERT" OR "transformer" OR "generative model*" OR "foundation model*" with wildcard operators to capture lexical variations (e.g., "models", "LLMs").

Deduplication and Filtering of Publication Types. We restricted the dataset to standard research outputs (e.g., articles, book chapters, preprints) by filtering out items like editorials, retractions, errata, letters, and purely supplementary materials. We also ensured that titles appearing multiple times in different forms (e.g., both a preprint and a published version) were deduplicated, generally favoring the peer-reviewed publication type over alternatives.

Supplementing Anthropic and OpenAI Data. Because OpenAI and Anthropic publications can sometimes be sparse in OpenAlex, we merged in additional CSV files containing each company's publication data that we scraped from their websites, combined with the scrape from Delaney, Guest, and Williams (2024) – but excluding their DeepMind scrape. After ensuring consistent columns, we appended these records, matched them to ROR IDs

for correct attribution, and again removed duplicates at the title level.[18]

Missing abstract and citation data. For entries missing a DOI, we use the OpenAlex API using the publication's OpenAlex ID to retrieve the DOI. Once DOIs are obtained, we employ multiple strategies to fetch abstracts. For general entries, we use the Crossref API to retrieve abstracts in a standardized XML format and processes the content to extract plain text. For entries published by specific organizations like Springer Nature, Elsevier, or Nature Portfolio, we use their respective APIs or webpage scraping methods tailored to each publisher's content structure. For Springer and Elsevier, valid API keys are used to authenticate requests and fetch metadata. If API access fails or isn't available, web scraping via BeautifulSoup is used as a fallback to extract abstract text directly from publisher websites.

We assign citation counts using Google Scholar data via the SerpApi service. Initially, we attempt a direct title-based search to extract citation data from the first relevant result. We then progress to more sophisticated approaches that include exact title matching and fuzzy string matching (via the fuzzywuzzy library), which allows us to better handle variations in how article titles are listed on Google Scholar. Our final dataset has 92 missing abstracts and 43 missing citation counts.

Fractional contribution. For multi-author papers, we computed each institution's fractional contribution based on the number of authors affiliated with that institution versus total authors on the paper (e.g., if an institution had 2 authors on a 10-author paper, it received a fraction of 0.20 for that paper). We retained only the distinct (paper, institution) pairs for our final dataset, ensuring one affiliation per author.

This approach does not distinguish among first authors, last authors, or any hierarchical authorship order; every co-author is given equal weight. *In effect, it ensures each author is credited exactly once to a single institution*. By summing these fractional shares across all authors, we can then calculate each institution's share of total authorship for each paper, summed over all papers.

When authors listed multiple institutional affiliations, we assigned each author to one

[18] See: https://github.com/Oscar-Delaney/safe_AI_papers.

institution for fractional counting. Specifically, we checked whether the author had any affiliation in our set of target ROR IDs (i.e., the academic or corporate AI institutions we tracked). If so, we took that affiliation as the author's "primary" affiliation for this study. Otherwise, we fell back to whichever affiliation appeared first in the metadata. By doing so, we avoid double-counting an author's fractional credit across multiple institutions.

AI safety & reliability classification: Two stages. We identified papers related to AI safety & reliability research in two stages. First, using a comprehensive keyword approach, scanning titles and abstracts for: safety, control, security, privacy, bias, fairness, explainability, interpretability, transparency, governance, risk, mitigation, evaluation, benchmarking, testing, alignment, ethics, responsibility, accountability, oversight, robustness, trust, and value alignment. Each paper containing at least one of these words was labeled "AI safety & reliability". This roughly halved our dataset. Next, we used GPT o4-mini to see if it agreed with these classifications. This reduced the dataset size substantially (by around two-thirds) to 1,178 papers.

6.3 Classification Process: Categories

OpenAI's o3 mini model used to classify AI research papers into eight categories. It first checked if the paper related to AI safety & reliability. After which the model was asked to classify each paper in to one of eight of the below categories, on the basis of the paper's title and abstract, given the category descriptions below. It provided a justification for each of its classifications. Each paper was only permitted to have a single classification.

AI safety definition. AI safety research covers the entire model life-cycle (pre-deployment or post-deployment) and involves reducing or identifying harms and implementing measures to make models safer and more reliable.

Eight AI Safety & Reliability Research Clusters

Testing and Evaluation. Testing, performance benchmarking ("bench" and "evals"), and auditing models to assess model capabilities, risks, behaviors, and flaws. Ensuring models

are robust to minor changes.

Alignment (Pre-Deployment). Ensuring AI systems behave in ways that are congruent with human values, expectations, and intents. This includes making AI systems functional, helpful, and harmless for humans and/or users, while avoiding behavior that diverges from intended goals or causes harm. Model deception, including any power-seeking tendencies, is included here, along with reward hacking.

Post-Deployment Risks and Model Traits. Societal impacts from AI products' applications and behavioral traits, as deployed in the marketplace, including addictiveness, persuasiveness, and model sycophancy (excessive agreement or manipulation to align with user preferences). Covers how corporate commercial incentives may be coded into AI models and products to prioritize engagement, advertising, and profit-seeking — including through the use of these behavioral traits. Includes misuse of models for ransomware, phishing, or spreading misinformation for commercial gain.

Ethics and Bias. Combating systemic biases embedded in AI models (in data, training, and alignment) and ensuring ethical decision-making. Focuses on mitigating harms to marginalized groups, addressing structural inequalities, and ensuring AI promotes justice and inclusion.

Multi-Agent and Agentic Safety. Safety issues specific to AI agents, including single-agent autonomy and multi-agent interactions. Covers coordination problems, emergent behaviors, incentive misalignment, and prevention of conflicts or unintended consequences in agentic systems and from autonomous agents.

Interpretability and Transparency. Making AI systems more understandable and accountable. Includes methods for explaining model behavior, clarifying decision-making processes, and enhancing trust by reducing the "black box" nature of AI systems.

Policy and Governance. Approaching AI safety as a challenge that extends beyond technical fixes, requiring legal and policy frameworks. Involves collaboration among policymakers, industry, civil society, and researchers to develop standards that guide safe AI development and deployment. Includes institutional governance, corporate transparency, technical

disclosures, and standards promoting interoperability, equity, and reliability.

Privacy and Security. Protecting AI systems from malicious use, adversarial attacks, and misuse by bad actors, along with addressing privacy violations and developing privacy-preserving methods. Includes vulnerabilities from adversarial inputs, data poisoning, misuse in surveillance, and theft of model weights.

6.4 Selective Behavioral Impact Papers

Sycophancy Papers:

- Sharma et al. (2023): Found that models tend to favor well-written agreeable ("sycophantic") responses over higher quality ones likely due to incorporating human feedback (since humans and preference models "prefer convincingly-written sycophantic responses over correct ones").
- Denison et al. (2024): Notes that sycophantic behavior is a form of specification gaming when AI systems learn undesired behaviors that are highly rewarded due to misspecified training goals.
- Perez et al. (2022): Highlights that user preferences tend to favor sycophantic answers and more reinforcement learning can lead to worse outcomes (such as stronger political views).

Persuasion Papers:

- Phuong et al. (2024): Introduces persuasion and deception as part of evaluations for frontier models, scoring persuasion as the highest risk among self-reasoning, self-proliferation, and cyber-security.

Deception Papers:

- Weidinger et al. (2021): A widely cited paper that structures the risk landscape from LLMs into six areas, including misinformation harms and human-computer interaction harms.

- Ngo, Chan, and Mindermann (2022): Reviews evidence on deception as a learned behavior during fine-tuning that can generalize beyond training contexts.

AI DISCLOSURES PROJECT

SOCIAL SCIENCE RESEARCH COUNCIL

Social Science Research Council
300 Cadman Plaza West, 15th Floor
Brooklyn, NY 11201, USA