Who Funds Frontier AI Evaluations? A Preliminary Mapping of the Third-Party Evaluation Ecosystem

Leon Staufer [1]

Abstract

Pre-deployment evaluations increasingly shape frontier model release decisions, yet the organizations behind them remain poorly understood. We map every evaluation cited in 75 public frontier system cards to its evaluator and funders. Of the 519 mapped evaluations, 68% are internal, while the visible third-party layer is thin, recurs across developers, and is in some cases funded by the developers it helps evaluate. The funding behind it is concentrated among a small set of recurrent philanthropic, corporate, and government sources, at a concentration significantly greater than chance allocation alone would produce. The visible network reflects only what evaluators and funders disclose, but even so the findings strengthen the case for greater transparency, a wider funder base, and stronger bridging institutions.

1. Introduction

Pre-deployment evaluations now sit near the center of frontier AI governance. When Anthropic limited the release of Claude Mythos in April 2026, citing dangerous cyber capabilities identified in pre-deployment tests (Anthropic, 2026), the decision showed both how evaluations have become gatekeeping infrastructure and how much now rests on them. They structure responsible-scaling, preparedness, and other frontier-safety frameworks (Shevlane et al., 2023; Alaga et al., 2024; Stelling et al., 2025), anchor proposals for managing frontier-model risk (Anderljung et al., 2023), and are required under Article 55 of the EU AI Act for general-purpose AI models with systemic risk (European Union, 2024; Fernández Llorca et al., 2025). New institutions, including the UK AI Security Institute (UK AISI) and the US Center for AI Standards and Innovation (US CAISI), have been created in part to run or commission them.

[1]University of Cambridge, Cambridge, UK. Correspondence to: Leon Staufer <lets2@cam.ac.uk>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

Yet the ecosystem producing these evaluations is poorly understood. Benchmarks have been criticized on their own terms (Raji et al., 2021; Salaudeen et al., 2025; Yu et al., 2026; Singh et al., 2025; Wallach et al., 2025; McIntosh et al., 2025; Eriksson et al., 2025; Cheng et al., 2025), black-box access is often inadequate for rigorous audits (Casper et al., 2024), and current audit tooling still falls short of broader accountability needs (Ojewale et al., 2025; Staufer et al., 2025). Despite these critiques, there is still no systematic public account of who produces the evaluations cited in frontier system cards or who funds them. That gap matters because funding shapes research agendas and outcomes across domains (Lundh et al., 2017; Abdalla & Abdalla, 2021). If the visible third-party layer is concentrated and financially tied to the developers it evaluates, then it may be less independent than governance frameworks assume.

We ask: *who performs and funds the pre-deployment evaluations cited in frontier AI system cards, and what does that imply for evaluator independence?* Drawing on the complete set of public system and model cards from Anthropic, Google DeepMind, Microsoft, OpenAI, and xAI, this paper makes three contributions:

1. a preliminary mapping of system cards, cited evaluations, evaluators, and funders;
2. early evidence that the third-party evaluation layer is thin and in places entangled with developers;
3. governance implications centered on transparency, funding diversity, and stronger bridging institutions.

2. Method

We collect every publicly available system or model card published through April 2026 by the five frontier developers above (Epoch AI, 2026). We extract evaluations, identify evaluators, trace funders, and cross-check major funder records. Extraction and funding research are LLM-assisted through OpenAI's Responses API (`gpt-5.4-mini` for document-grounded extraction and profiling, `o4-mini-deep-research` for multi-source funding research). Each funding link is then classified for whether it substantively supports AI evaluation work, so that broad institutional or unrelated funding is excluded. Entities are merged by deterministic identifiers and a semantic pass

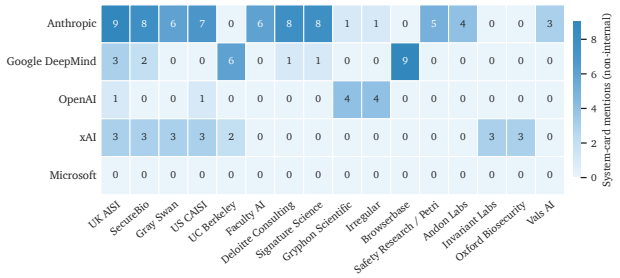

Figure 1. **A handful of external evaluators are cited by several developers, while most of the top evaluators work with just one.** Shading shows the number of non-internal system-card mentions between frontier developers and selected external evaluators.

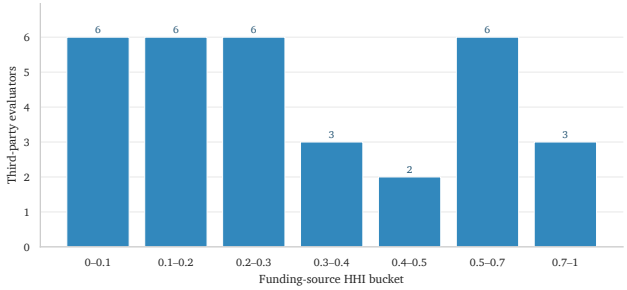

Figure 2. **Most third-party evaluators draw on several funders, with a concentrated minority.** Distribution of Herfindahl-Hirschman Index scores across the 32 third-party evaluators with disclosed non-investment funding. Values near 1 indicate reliance on a single funder.

that reconciles name variants such as Open Philanthropy and its successor Coefficient Giving (Coefficient Giving, 2025b). The full annotation taxonomy and pipeline details are given in Sections B and C. The mapping covers 5 developers, 75 system cards, 519 evaluations, 125 evaluators, and 351 funding sources.

To assess accuracy, we manually validated a stratified random sample of 115 extracted edges against their cited source: 78% were fully correct, with another 8% correct except for a minor detail. Extracting what each card cites is more reliable (89%) than the two attribution layers: which evaluator ran an evaluation (74%) and who funds that evaluator (73%). Funder counts in particular should therefore be read as indicative. See Section D for details.

3. Findings

Most cited evaluations are internal. Of the mapped evaluations, 68.2% are internal, 16.0% mixed,[1] 14.7% third-party, and 1.1% unknown. Internal evaluation dominates for every developer except xAI (49.1% internal, Figure 11). Even benchmarks that are externally *designed* are frequently *executed* internally, so readers must trust the developer's own compute, prompting, and reporting. Reported counts also vary sharply, from 21.2 evaluations per card at Anthropic to 3.3 at Microsoft (Table 2), though this partly reflects reporting granularity rather than comprehensiveness.

A small set of external evaluators bridges most developers. Among non-internal evaluators, SecureBio is cited across four of five developers in 17 system cards, UK AISI across four developers in 16 cards, Gray Swan across four developers in 15 cards, and US CAISI across three developers in 14 cards (Figure 1). Government institutions are among the main bridges in an otherwise fragmented evaluation ecosystem.

Some evaluators work with only a single developer rather than spreading across the market. Browserbase's 9 system-card appearances are all for Google DeepMind, and Deloitte Consulting, Signature Science, Gryphon Scientific, and Irregular each do 80% or more of their cited evaluations for a single developer (Table 6). Each developer reuses a characteristic set of evaluators that rarely appear for any other (Figure 9), indicating the lack of a competitive market in which evaluators would be interchangeable. A permutation test confirms this clustering is statistically significant ($p < 0.0001$ over 10,000 permutations, Section D).

Beyond these, external evaluation is sparse and weakly attributed. Evaluator citations are heavy-tailed, with a few evaluators recurring across many cards and a long tail cited only once or twice (Figure 10). Red team evaluations are particularly opaque, with the generic "external partners" label among the most-cited evaluator entries (Table 5 and Figure 12), aggregating potentially distinct organizations that should not be read as one.

The evaluation funding is concentrated. The 351 funding sources generate 757 funding links, with Coefficient Giving as the most recurrent named source (25 evaluators, 58 links). Anthropic appears as a funder of 11 other evaluators and OpenAI as a funder of 10, a developer-as-funder pattern made explicit by OpenAI's recent USD 7.5M commitment to the Alignment Project (OpenAI, 2026). The Frontier Model Forum and its AI Safety Fund, financed jointly by frontier developers and philanthropic partners (Frontier Model Forum, 2023; 2025), are together among the most recurrent funders (6 and 5 evaluators). By source type, philanthropy is the largest channel, with 221 links (29%) across 54 evaluators, ahead of corporate funding at 188 links (25%) across 51 evaluators and venture capital at 149 links (20%) across 24 evaluators. Table 4 gives the full breakdown.

The funder base behind each evaluator is narrow. For each evaluator we compute the Herfindahl-Hirschman Index (HHI) (Rhoades, 1993) over the number of distinct non-investment funding sources. Most evaluators draw on

[1] Mixed refers to evaluations externally designed but internally executed, or run by an external organization using developer staff or infrastructure. See Section C.

several funders, but the base is thin: of the 32 evaluators with disclosed funding, 9 have HHI scores above 0.5 and 2 rely on a single funder (Figure 2). Across 30 evaluators with at least two disclosed funders, the mean HHI is 0.284, which is statistically significant (null mean 0.205, $p < 0.0001$; Section D) and sits above the 0.18 threshold that antitrust guidelines treat as highly concentrated (U.S. Department of Justice & Federal Trade Commission, 2023). A dollar-weighted index, computed on the disclosed funding amounts and scorable only for the 20 evaluators that disclose at least one funding amount, shows higher concentration (Figure 8). Table 3 lists the top funders. The evaluator base is also geographically narrow: 77.6% of the 76 evaluators with a recorded country are US- or UK-based (Table 7 and Figure 13).

4. Discussion

Being third-party does not make an evaluator independent. Governance frameworks treat external evaluation as a check on the developer, but the mapping shows that organizational separation and independence come apart. An evaluator can be a distinct legal entity and still depend on the developer it assesses for its funding, its model access, and its next engagement. Independence is therefore not settled by who runs an evaluation, which is what system cards record, but by the money and access behind it, which they rarely record.

Concentrated funding shapes which risks get tested. Funding is not neutral with respect to what gets studied. Across domains from pharmaceutical trials to commercialized academic science, funding sources shape which questions are asked and which findings are produced, even when researchers act in good faith (Lundh et al., 2017; Mirowski, 2011; Jukola, 2021; Franssen et al., 2018). This can occur through agenda-setting where funders have characteristic priorities, and the aggregate output of a field comes to reflect the distribution of its funders. Abdalla & Abdalla (2021) document the reach of industry funding in AI ethics research, and Birhane et al. (2026) extend it to AI governance more broadly. The analysis here shows concentration even among largely good-faith philanthropic evaluation funders. For example, Coefficient Giving has acknowledged holding a concentrated share of AI-safety philanthropic funding and urged new funders to enter (Coefficient Giving, 2025a). However, this concentration is likely to deepen. As frontier developers pursue initial public offerings, far more money is expected to enter the same AI-safety philanthropy base that already funds much of the third-party evaluation studied here (Ford, 2026). This may mean that a small set of funders carries a particular view of which risks matter, and a concentrated evaluation ecosystem will test for those risks and leave others under-covered.

Due to limitations of the underlying data, the mapping and analysis should be understood as preliminary. It tracks only named evaluations and the relationships that evaluators and funders choose to reveal. It is a network of visible relationships, not a complete financial account. Because the most transparent funders are the easiest to trace, the visible concentration may overemphasize those who disclose. Some links attach to an evaluator's overall funding rather than money earmarked for a given evaluation. Funder counts are therefore best read as an upper bound, and the data show concentration rather than causation. A further limitation lies in the LLM-assisted and only partially validated pipeline (Sections 2 and D). Its weakest stage, funder attribution, errs mostly by adding over-broad links that dilute rather than inflate concentration, so those findings are conservative. Finally, pre-deployment testing favors risks that operationalize as discrete tasks, skewing attention toward acute and catastrophic harms and away from slower societal and systemic risks (Weidinger et al., 2023; Solaiman et al., 2023).

A robust third-party evaluation ecosystem requires transparency, independence, and funder plurality. *Transparency.* Unattributed red teams and the recurring generic "external partners" entry show that system cards often obscure who evaluated a model and on what terms (Table 5). For each evaluation, cards could report who designed it, who ran it, what access was granted, and any funding or contractual tie between developer and evaluator (Staufer et al., 2025). *Independence by distance.* Because the same developers fund, contract, and read the work (Tables 3 and 6), declared independence is fragile. What matters is not an evaluator's client diversity but distance from the evaluated developer's money and access, which evaluators guard only unevenly and informally (METR, 2026; Apollo Research, 2026; Ford, 2026). In clinical research, a levy on pharmaceutical companies funds a public program of independent drug research (Italian Medicines Agency, 2010). An analog would have frontier developers pay into a public pool from which government regulators commission evaluators. *Funder plurality.* A wider base, including national research councils, civil-society foundations, and non-Anglophone public bodies, would open the choice of which risks to prioritize rather than leaving it to a small and geographically narrow set of funders (Tables 3 and 7). These suggestions come with tradeoffs, since a public commissioner carries its own priorities, weighted toward national security and competitiveness, but public bodies are at least more democratically answerable for which risks get tested than private developers or a few philanthropies. As frontier release decisions increasingly rest on the third-party evaluation layer, and as long as that layer is thin, concentrated, and largely self-disclosed, greater transparency, independence, and funder plurality are needed to let evaluations bear that weight.

Acknowledgments

This research was supported by the MATS Research program, which provided funding for L.S. through research stipends. MATS is funded in part by Coefficient Giving, among others.

We thank MATS and our research manager Keivan Navaie for their organizational assistance and research support. We are grateful to Giulio Corsi and Stephen Casper for their valuable feedback on earlier drafts of this paper. Thanks also to participants of the CHIA annual conference as well as the anonymous reviewers here and at the University of Cambridge's Centre for the Future of Intelligence for their constructive comments and suggestions.

Impact Statement

This paper maps who produces and funds pre-deployment evaluations in order to inform debates about evaluator independence. The mapping remains descriptive and incomplete. Concentration in the visible record partly reflects uneven disclosure, and the results should not be read as evidence that any funder or evaluator acts in bad faith, nor as a complete financial account.

Positionality: L.S. has received research support connected to the same philanthropic ecosystem studied here, including funding linked to Coefficient Giving, the most-named funder in this analysis. This brings familiarity with the ecosystem but also a potential conflict of interest. The support came with no further conditions.

LLM Usage Statement

LLMs were used in two roles. Within the research pipeline, OpenAI models assisted with document collection and evaluation extraction (`gpt-5.4-mini`), evaluator profiling and funding research (`o4-mini-deep-research`), and entity resolution and deduplication (`gpt-5.5`), as described in Section 2 and Section B. Model outputs were spot-checked against primary sources. Separately, Anthropic's `Claude Opus 4.8` was used to assist with editing and formatting the manuscript. The author takes responsibility for all claims and any remaining errors.

References

Abdalla, M. and Abdalla, M. The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. ACM, 2021. doi: 10.1145/3461702.3462563. URL https://dl.acm.org/doi/10.1145/3461702.3462563.

Alaga, J., Schuett, J., and Anderljung, M. A grading rubric for AI safety frameworks, 2024. URL https://arxiv.org/abs/2409.08751.

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., Schuett, J., Shavit, Y., Siddarth, D., Trager, R., and Wolf, K. Frontier AI regulation: Managing emerging risks to public safety, 2023. URL https://arxiv.org/abs/2307.03718.

Anthropic. Claude mythos preview system card. Anthropic system card, April 2026. URL https://red.anthropic.com/2026/mythos-preview/.

Apollo Research. Our norms on conflicts of interest, security, and science communication, 2026. URL https://www.apolloresearch.ai/blog/our-norms-coi-security-science-communication/. Accessed June 2026.

Birhane, A., Angius, R., Agnew, W., Pandit, H. J., Mitra, B., Dobbe, R., and Talat, Z. Big AI's regulatory capture: Mapping industry interference and government complicity. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26)*, 2026. doi: 10.1145/3805689.3806740. URL https://doi.org/10.1145/3805689.3806740.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Von Hagen, M., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., and Hadfield-Menell, D. Black-box access is insufficient for rigorous AI audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, pp. 2254–2272, 2024. doi: 10.1145/3630106.3659037. URL https://doi.org/10.1145/3630106.3659037.

Cheng, Z., Wohnig, S., Gupta, R., Alam, S., Abdullahi, T., Ribeiro, J. A., Nielsen-Garcia, C., Mir, S., Li, S., Orender, J., Bahrainian, S. A., Kirste, D., Gokaslan, A., Glinka, M., Eickhoff, C., and Wolff, R. Benchmarking is broken – don't let AI be its own judge, 2025. URL https://arxiv.org/abs/2510.07575.

Coefficient Giving. AI safety and security need more funders. Coefficient Giving research note, October 2025a. URL https://coefficientgiving.org/research/ai-safety-and-security-need-more-funders/.

Coefficient Giving. Open philanthropy is now coefficient giving. Coefficient Giving blog post, November 2025b.

URL https://coefficientgiving.org/research/open-philanthropy-is-now-coefficient-giving/.

Epoch AI. Ai companies. Epoch AI frontier-company tracker, 2026. URL https://epoch.ai/data/ai-companies?view=table&tab=companies.

Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., and Fernandez-Llorca, D. Can we trust AI benchmarks? an interdisciplinary review of current issues in AI evaluation, 2025. URL https://arxiv.org/abs/2502.06559.

European Union. Article 55: Obligations for providers of general-purpose AI models with systemic risk. Regulation (EU) 2024/1689, 2024. URL https://artificialintelligenceact.eu/article/55/.

Fernández Llorca, D., Eriksson, M., Gómez, E., and Vanschoren, J. The role of AI safety benchmarks in evaluating systemic risks in general-purpose AI models. European Commission Joint Research Centre, 2025. URL https://data.europa.eu/doi/10.2760/1807342.

Ford, C. Anthropic employees say they'll give away billions. where will it go? Transformer, 2026. URL https://www.transformernews.ai/p/anthropic-employees-philanthropy-billions-donations-effective-altruism-coefficient-giving-ai-safety. Published 12 March 2026; accessed June 2026.

Franssen, T., Scholten, W., Hessels, L. K., and de Rijcke, S. The drawbacks of project funding for epistemic innovation: Comparing institutional affordances and constraints of different types of research funding. *Minerva*, 56:11–33, 2018. doi: 10.1007/s11024-017-9338-9. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5807497/.

Frontier Model Forum. AI Safety Fund. Frontier Model Forum, October 2023. URL https://www.frontiermodelforum.org/ai-safety-fund/.

Frontier Model Forum. Announcement of new AI safety fund grantees. Frontier Model Forum update, December 2025. URL https://www.frontiermodelforum.org/updates/announcement-of-new-ai-safety-fund-grantees/.

Italian Medicines Agency. Feasibility and challenges of independent research on drugs: the italian medicines agency (AIFA) experience. *European Journal of Clinical Investigation*, 40(1):69–86, 2010. doi: 10.1111/j.1365-2362.2009.02226.x.

Jukola, S. Commercial interests, agenda setting, and the epistemic trustworthiness of nutrition science. *Synthese*, 198:S2629–S2646, 2021. doi: 10.1007/s11229-019-02228-3. URL https://doi.org/10.1007/s11229-019-02228-3.

Lundh, A., Lexchin, J., Mintzes, B., Schroll, J. B., and Bero, L. Industry sponsorship and research outcome. *Cochrane Database of Systematic Reviews*, (2):MR000033, 2017. doi: 10.1002/14651858.MR000033.pub3. URL https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.MR000033.pub3/full.

McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Xu, D., Watters, P., and Halgamuge, M. N. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Transactions on Artificial Intelligence*, 7(1):22–39, 2025. doi: 10.1109/TAI.2025.3569516. URL https://ieeexplore.ieee.org/document/11002710.

METR. About, 2026. URL https://metr.org/about. Accessed June 2026.

Mirowski, P. *Science-Mart: Privatizing American Science*. Harvard University Press, Cambridge, MA, 2011.

Ojewale, V., Steed, R., Vecchione, B., Birhane, A., and Raji, I. D. Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025. doi: 10.1145/3706598.3713301. URL https://arxiv.org/abs/2402.17861.

OpenAI. Advancing independent research on AI alignment. OpenAI announcement, March 2026. URL https://openai.com/index/advancing-independent-research-ai-alignment/.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. Ai and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2111.15366.

Rhoades, S. A. The herfindahl-hirschman index. *Federal Reserve Bulletin*, 79(3):188–189, 1993.

Salaudeen, O., Reuel, A., Ahmed, A., Bedi, S., Robertson, Z., Sundar, S., Domingue, B., Wang, A., and Koyejo, S. Measurement to meaning: a validity-centered framework for AI evaluation, 2025. URL https://arxiv.org/abs/2505.10573.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S.,

Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., Christiano, P., and Dafoe, A. Model evaluation for extreme risks, 2023. URL https://arxiv.org/abs/2305.15324.

Singh, S., Nan, Y., Wang, A., D'Souza, D., Kapoor, S., Üstün, A., Koyejo, S., Deng, Y., Longpre, S., Smith, N. A., Ermis, B., Fadaee, M., and Hooker, S. The leaderboard illusion, 2025. URL https://arxiv.org/abs/2504.20879.

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Chen, C., Daumé III, H., Dodge, J., Evans, E., Friedrich, F., Gohar, U., Hooker, S., Jernite, Y., Luccioni, A., Lusoli, A., Mickel, J., Mitchell, M., Newman, J., Png, M.-T., Strait, A., Vassilev, A. T., and Subramonian, A. Evaluating the social impact of generative AI systems in systems and society, 2023. URL https://arxiv.org/abs/2306.05949.

Staufer, L., Yang, M., Reuel, A., and Casper, S. Audit cards: Contextualizing AI evaluations, 2025. URL https://arxiv.org/abs/2504.13839.

Stelling, L., Murray, M., Galizzi, B., Schaffelder, M., Campos, S., and Papadatos, H. Evaluating AI providers' frontier safety frameworks, 2025. URL https://arxiv.org/abs/2512.01166.

U.S. Department of Justice and Federal Trade Commission. Merger guidelines, 2023. URL https://www.ftc.gov/system/files/ftc_gov/pdf/2023_merger_guidelines_final_12.18.2023.pdf. Accessed June 2026.

Wallach, H., Desai, M., Cooper, A. F., Wang, A., Atalla, C., Barocas, S., Blodgett, S. L., Chouldechova, A., Corvi, E., Dow, P. A., Garcia-Gathright, J., Olteanu, A., Pangakis, N. J., Reed, S., Sheng, E., Vann, D., Vaughan, J. W., Vogel, M., Washington, H., and Jacobs, A. Z. Position: Evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 82232–82251, 2025. URL https://proceedings.mlr.press/v267/wallach25a.html.

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., and Isaac, W. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*, 2023. doi: 10.48550/arXiv.2310.11986. URL https://arxiv.org/abs/2310.11986.

Yu, J., Engelmann, S., Cao, J., Ali, L., and Papakyriakopoulos, O. How should AI safety benchmarks benchmark safety?, 2026. URL https://arxiv.org/abs/2601.23112.

A. Network Graph

Figure 3 gives a visual overview of the mapped network linking system cards, evaluations, evaluators, and funding sources.

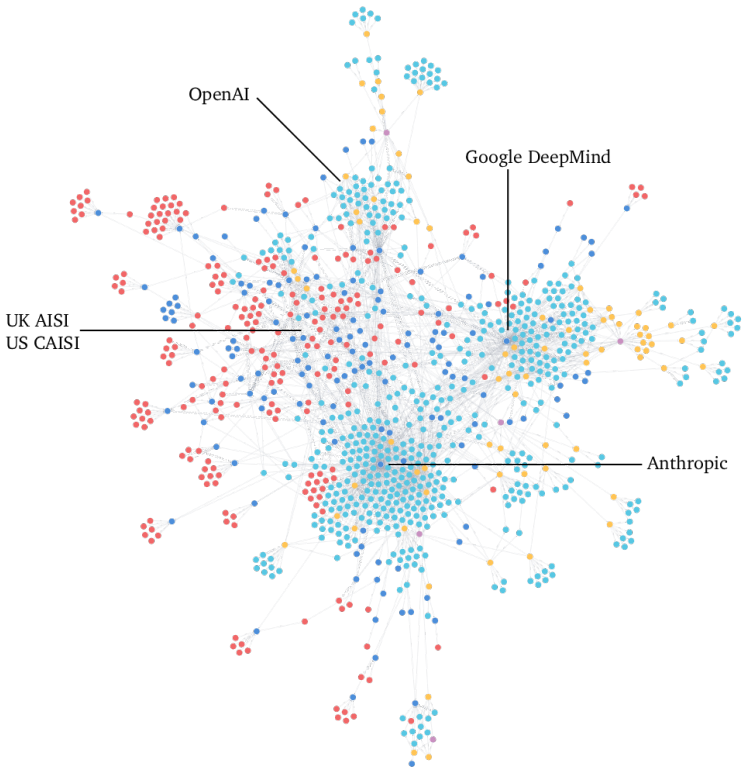

Figure 3. **The mapping forms one connected graph running from system cards through evaluations and evaluators to funders.** Network view of the mapping. Node colors indicate type: yellow = system cards, cyan = evaluations, blue = evaluators, red = funding sources, purple = developers, and light purple = high-degree hub or aggregate nodes marked with an asterisk. The figure is a high-level visual companion to the appendix taxonomy and tables.

B. Method Details

Collection proceeds in five stages, each using the research mode best suited to its task. Figure 4 summarizes this pipeline.

(1) **Card collection.** System and model cards are gathered through web search and developer documentation pages.
(2) **Evaluation extraction.** Evaluations cited in each card are extracted through a document-grounded LLM-assisted workflow that pairs the source document with structured extraction prompts.
(3) **Evaluator profiling.** Producing organizations are profiled using web search against the evaluator's website, public outputs, and corporate or charity filings.
(4) **Funding research.** Funding for each evaluator is researched using a deep-research workflow (OpenAI's `o4-mini-deep-research` via the Responses API, not an interactive chatbot feature) that draws on the evaluator's own disclosures (papers, transparency pages, acknowledgments), public filing databases (UK Companies House, US IRS Form 990 returns via ProPublica's Nonprofit Explorer), and funding signals implicit in the system card itself. Each recovered link is classified by whether it substantively supports AI evaluation work, separating direct evaluation or AI-safety support from broad institutional or unrelated funding, and only substantively relevant links enter the funding analysis.
(5) **Reverse lookup.** A curated list of major funders (Coefficient Giving, Survival and Flourishing Fund, Frontier Model Forum AI Safety Fund, Schmidt Sciences, Longview Philanthropy, Knight Foundation, Omidyar Network) is cross-checked against the evaluators already mapped.

Figure 4. **The pipeline runs in five stages, each using the research mode best suited to its task.** Overview of the data-collection pipeline. The reverse-lookup stage cross-checks funder records against the evaluators already mapped rather than creating a parallel analytical entity.

C. Annotation Taxonomy

Each node and edge in the map is classified by type and by where its supporting evidence was located, recorded in the funding-signal and funding-disclosure-location fields below.

- **Document type.** System card, model card, safety report, technical report, or other.
- **Evaluation type.** Benchmark, red team, external testing, internal testing, expert review, audit, or other.
- **Evaluation scope.** Internal (run by the developer), third-party (run by an external organization), or mixed (an external organization working with developer staff or infrastructure, or a benchmark externally designed but internally executed).
- **Funding signal in source.** None, implicit (e.g. a contracted engagement), or explicit (a named funder or amount).
- **Evaluator type.** Academic lab, for-profit company, non-profit, government, individual, or mixed.
- **Evaluator country.** Country of primary operation, or *unknown*.
- **Evaluator attribution.** Whether the evaluator is named in the source document (yes/no).
- **Funding-disclosure location.** System card, evaluation paper or documentation, evaluator website or transparency page, public filings, funder grant list, FOI response, or other.
- **Funding source type.** Philanthropy, venture capital, corporate, government, university, donor-advised fund, individual, or other.
- **Funding relationship type.** Grant, contract, investment, sponsorship, parent-organization support, donation, or other.

D. Validation and Review

Each stage writes structured records with the source document and locator that justify it, and these are inspected through a purpose-built review tool (Figure 5) that lets a human browse every system card, evaluation, evaluator, and funding link alongside its source URL and supporting excerpt. We used this tool to construct and curate the graph, including the entity-resolution pass that merges name variants, and to draw and adjudicate the 115-edge spot-check reported in Section 2. Each sampled edge was checked against the cited source and marked correct, partially correct, or wrong.

Validation has two parts: a manual spot-check of extraction accuracy, and permutation tests of the two concentration findings. The spot-check draws a stratified random sample of 115 edges across the three attribution layers (card-to-evaluation, evaluation-to-evaluator, evaluator-to-funder) and codes each against its cited source, as detailed below.

Spot-check coding. Each sampled edge was coded against its cited source on a three-point scale.

- *Fully correct*: the entity and every checked detail match the source.
- *Partially correct*: the right entity or relationship is identified, but a non-core detail is wrong or incomplete in a way that does not invalidate the core claim. Examples are a third-party benchmark run in-house labeled third-party rather than as an internal run (card-to-evaluation), a role such as designer or maintainer overstating loose involvement (evaluation-to-evaluator), or a correct funder paired with the wrong amount, date, or relationship type (evaluator-to-funder).
- *Wrong*: the claimed entity or relationship is unsupported by the source, such as a phantom funding edge or a misidentified organization.

Of the 115 sampled edges, 90 were fully correct, 9 partially correct, and 16 wrong. Strict agreement counts only the fully correct edges (78%), and the lenient rate gives partials half credit (82%). Accuracy varies by edge type (Table 1): extracting what each card cites is reliable, while the two attribution layers, which evaluator ran an evaluation and who funds that evaluator, are weaker, so funder counts in particular should be read as indicative.

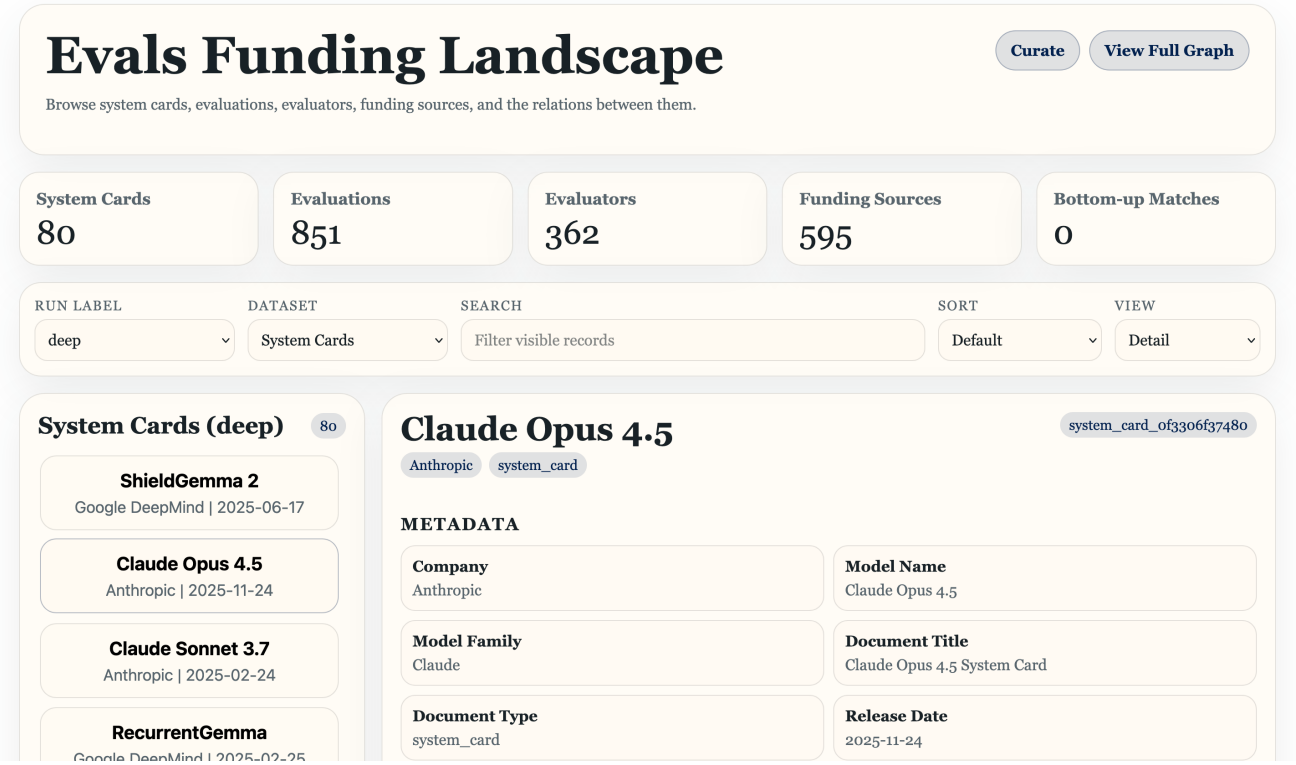

Figure 5. **The purpose-built review tool allowed us to check every record against its source.** Each record (here a system card) is shown with its metadata, source URL, deduplication key, and cited evaluations, so an entry can be checked against its source.

Table 1. **Extracting what each card cites is reliable, while the two attribution layers are weaker.** Spot-check accuracy by edge type over the stratified random sample of 115 edges. *Strict* counts only fully correct edges; *lenient* gives partially correct edges half credit.

Edge type	Correct	Partial	Wrong	n	Strict	Lenient
Card → evaluation	31	3	1	35	88.6%	92.9%
Evaluation → evaluator	26	4	5	35	74.3%	80.0%
Evaluator → funder	33	2	10	45	73.3%	75.6%
Overall	90	9	16	115	78.3%	82.2%

Significance tests. Some evaluator-developer clustering is mechanical: a few developers publish most system cards and a few funders make most grants, so any evaluator would look somewhat concentrated by chance. To separate genuine clustering from this marginal imbalance, we test both concentration findings against a *degree-preserving* null, a bipartite configuration model that reshuffles the observed edges 10,000 times while holding both margins fixed, so every evaluator keeps its exact number of edges and every developer or funder keeps its total. The resulting distribution is the concentration that natural imbalance alone would produce. We report a one-sided p as the fraction of permutations whose statistic meets or exceeds the observed value. This approach avoids the trivial dominance of Anthropic's high card count. Both tests measure concentration given what is disclosed and do not correct for uneven disclosure.

For the bilateral clustering, the population is the 44 non-internal evaluators with at least three distinct system-card mentions, excluding the five developers as evaluators and individual-typed evaluators, giving 300 mention-edges across four developers. The statistic is the mean dominant-developer share, each evaluator's largest fraction of mentions with a single developer. The observed 0.754 far exceeds the null (mean 0.574, 95% interval [0.541, 0.609], $p < 0.0001$), shown in Figure 6. A multinomial null gives the same result (mean 0.576, $p < 0.0001$), and the share of evaluators tied entirely to one developer is 0.25 (11 of 44) against a null of 0.05 ($p < 0.0001$).

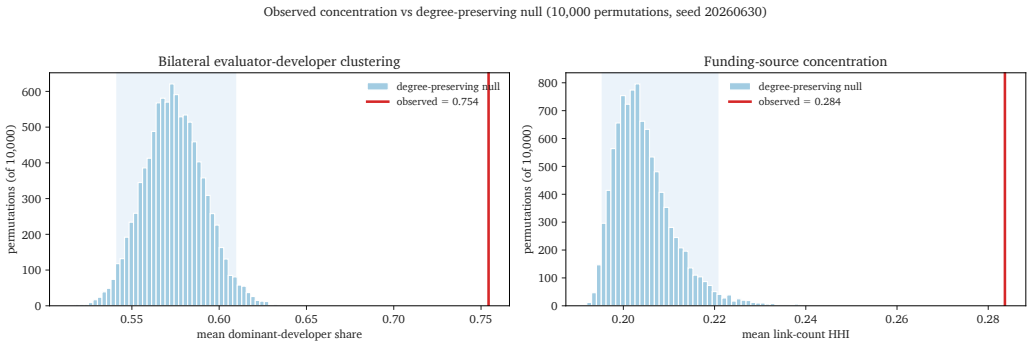

Figure 6. **Evaluator-developer clustering far exceeds what chance allocation would produce.** Observed mean dominant-developer share against a degree-preserving null in which each evaluator's number of cited appearances and each developer's total are held fixed and mentions are reshuffled 10,000 times. The observed value (red) lies far above the null distribution (blue), so the clustering is not an artifact of a few developers publishing most system cards.

For funding concentration, the population is the 30 third-party evaluators with at least two non-investment funding links, spanning 259 links across 141 funders. The statistic is the mean link-count HHI. The observed 0.284 exceeds the null (mean 0.205, 95% interval [0.195, 0.221], $p < 0.0001$).

E. Ecosystem Diagram

This appendix supplements the method summary with a stylized ecosystem diagram. Figure 7 sketches the structure of the evaluation ecosystem in scope.

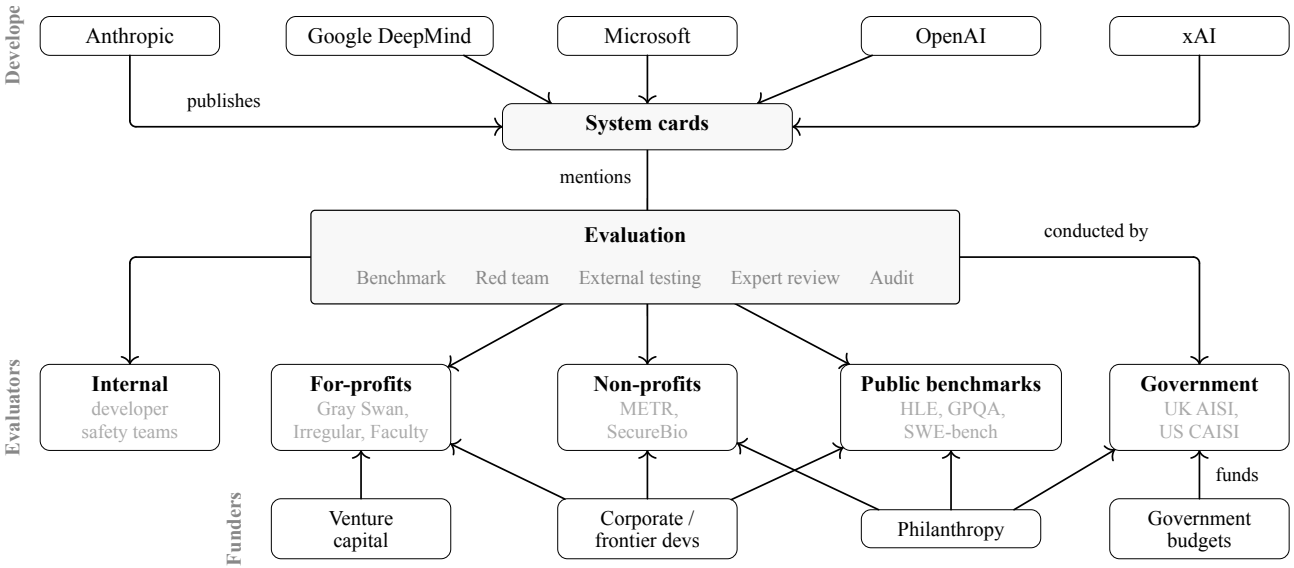

Figure 7. **Cited evaluations are produced internally or by one of four external categories, each drawing on a different funding mix.** Stylized view of the pre-deployment evaluation ecosystem in scope. Frontier developers cite evaluations in their system cards. Arrows show dominant flows rather than every observed link.

F. Additional Figures from Analysis

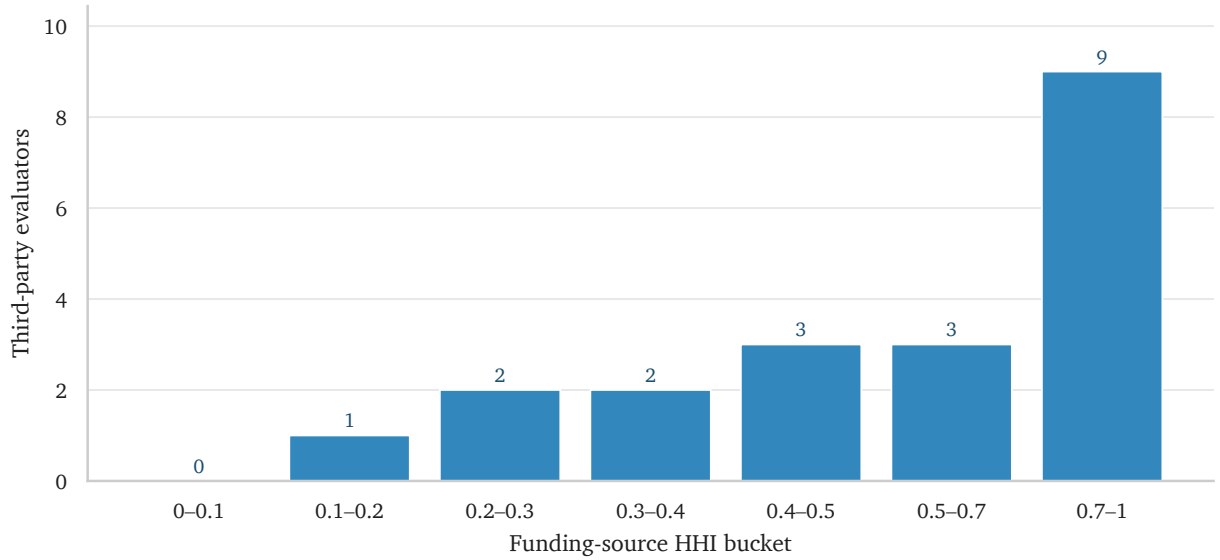

Figure 8. **Third-party evaluator HHI computed on disclosed funding amounts.** The evaluators from Figure 2 scored by their disclosed non-investment funding in USD rather than by number of distinct funders. Only the 20 evaluators that disclose at least one funding amount can be scored this way, since funders whose contributions are undisclosed do not enter the calculation. Values closer to 1 mean a single funder accounts for most of the evaluator's disclosed funding.

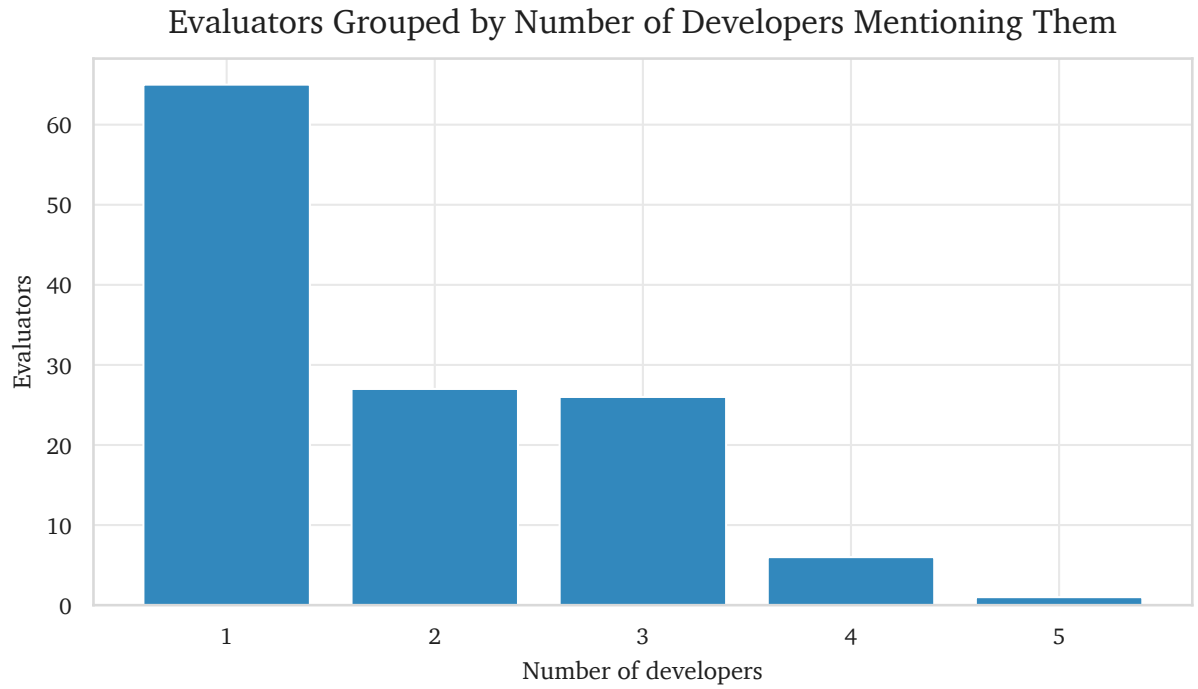

Figure 9. **Most external evaluators appear for a single developer.** Number of developers that mention each external evaluator. Only a handful (notably SecureBio, UK AISI, and Gray Swan) are cited by four or more developers.

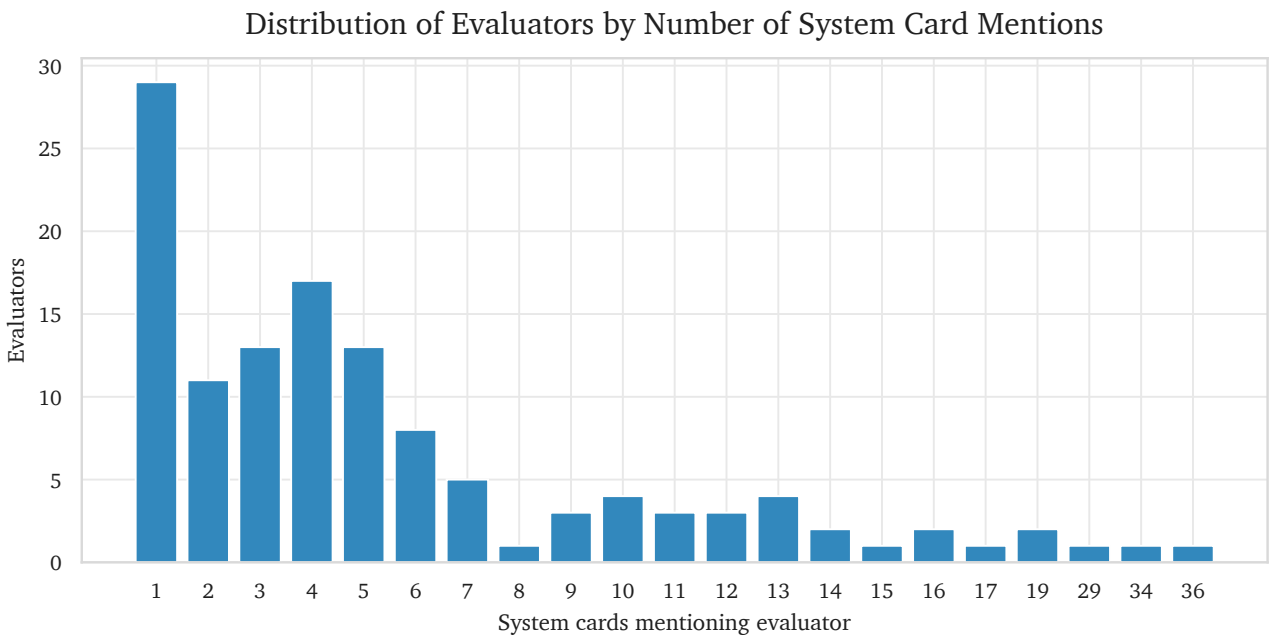

Figure 10. **Evaluator citations are heavy-tailed.** Number of system cards mentioning each evaluator, including internal evaluators run by the developers themselves. A long tail is cited only once or twice, while a small number recur across many system cards.

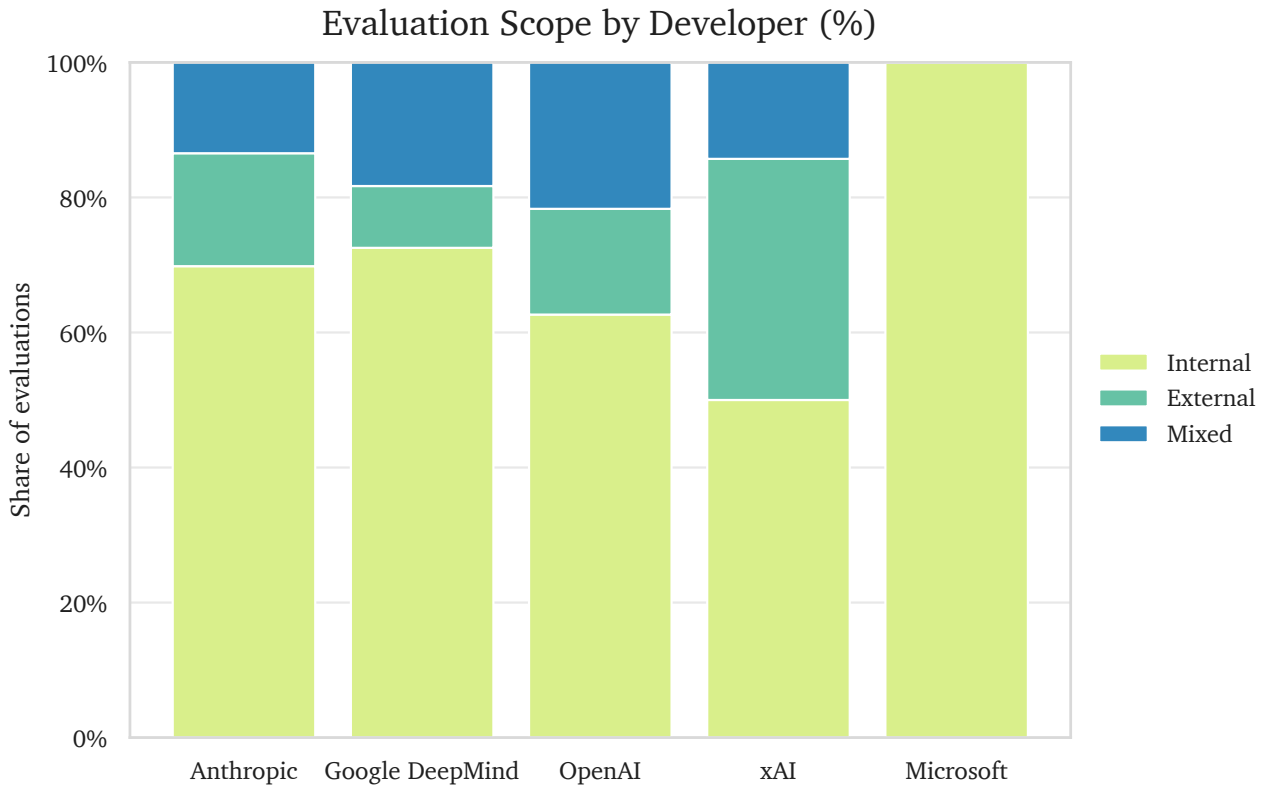

Figure 11. **Internal evaluation dominates for every developer except xAI.** Share of cited evaluations by scope (internal, mixed, third-party, unknown) for each developer.

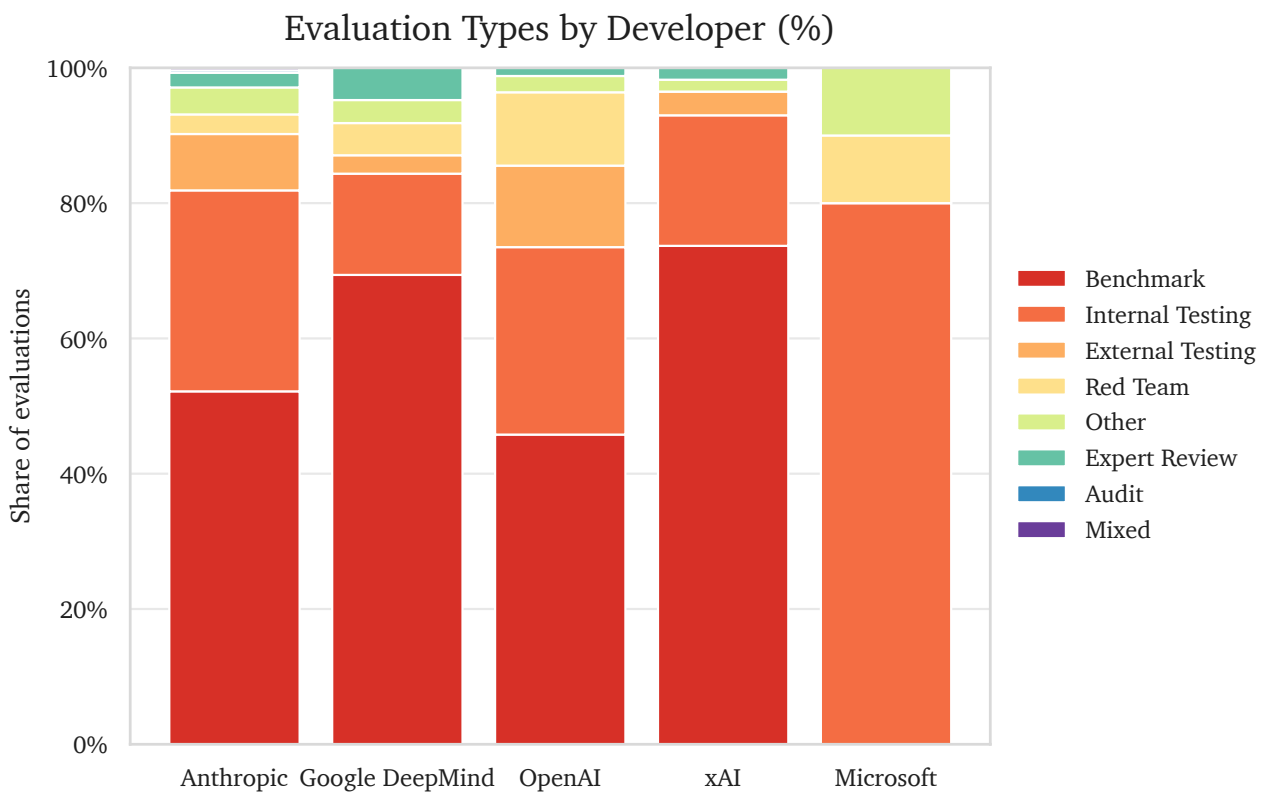

Figure 12. **Benchmarks are the largest evaluation category, and red-teaming is listed with less attribution.** Share of cited evaluations by type for each developer.

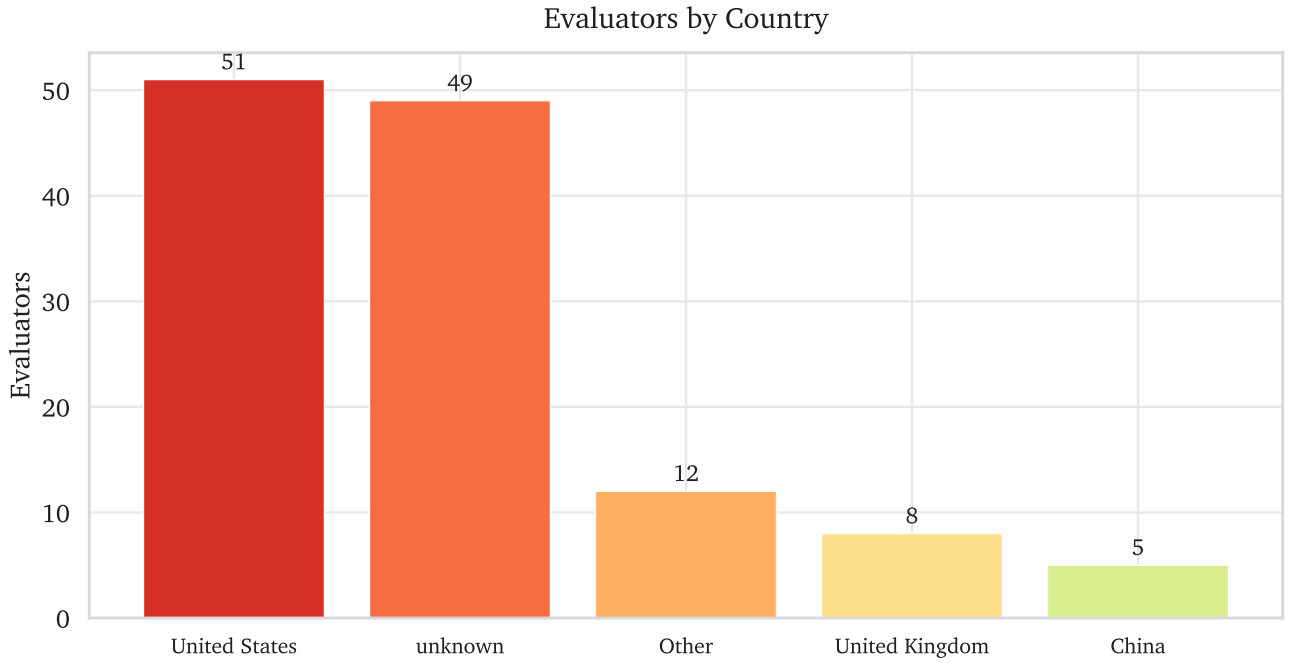

Figure 13. **The evaluator base is geographically narrow, and the country of many evaluators is unknown.** Country of primary operation for each evaluator in the mapping, including an explicit *unknown* category.

G. Data Tables

Table 2. **Cited evaluations per card vary sharply across developers.** Evaluation density per developer. Microsoft's lower count reflects its different document format, and its "platform cards" and "transparency notes" focus more on intended uses than on detailed evaluation results, which partly explains its 100% internal share.

Developer	System Cards	Min per Card	Avg per Card	Max per Card
Anthropic	13	1	21.2	92
Google DeepMind	32	0	4.6	40
OpenAI	15	1	5.5	23
xAI	12	1	4.8	14
Microsoft	3	3	3.3	5

Table 3. **A few recurrent funders, led by Coefficient Giving, back much of the external layer.** Top funders by number of funding links, with the count of distinct evaluators each supports. Parent-organization support between frontier developers and their own internal evaluator entries is excluded to avoid double-counting. The Frontier Model Forum and its grantmaking vehicle the AI Safety Fund appear as separate nodes.

Funding Source	Source Type	Funding Links	Evaluators
Coefficient Giving	Philanthropy	58	25
Google / Alphabet	Corporate	21	15
Schmidt Sciences	Philanthropy	20	14
NSF	Government	15	11
Amazon	Corporate	14	8
Gates Foundation	Philanthropy	12	6
Frontier Model Forum	Philanthropy	12	6
Anthropic (as funder)	Corporate	11	11
OpenAI (as funder)	Corporate	11	10
AI Safety Fund	Philanthropy	10	5

Table 4. **Philanthropy supplies the most funding links, ahead of corporate funding and venture capital.** Funding links by source type, aggregated across the full graph.

Source Type	Funders	Funding Links	Evaluators
Philanthropy	66	221	54
Corporate	80	188	51
Venture Capital	74	149	24
Government	55	90	33
Individual	58	71	24
Other	6	13	7
Donor-Advised Fund	5	13	6
University	7	12	7

Table 5. **The most-cited external evaluators are few and recur across developers.** Top non-internal evaluators by number of system cards citing them (excluding frontier developers appearing as evaluators of peer models).

Evaluator	Organization Type	Developers	System Cards	Evaluations
SecureBio	Non-Profit	4	17	10
UK AISI	Government	4	16	9
Gray Swan	For-Profit	4	15	7
Scale AI	For-Profit	3	19	6
US CAISI	Government	3	14	8
External Partners (Unnamed)	Unknown	3	19	15
Center for AI Safety	Non-Profit	3	13	8
FutureHouse	Non-Profit	3	13	7
Princeton University	Academic	3	14	4
Stanford University	Academic	3	12	4

Table 6. **Many external evaluators do almost all their cited work for one developer.** Evaluators clustered with one developer. *Share* is the proportion of the evaluator's visible non-internal system-card appearances that come from its single most frequent developer.

Evaluator	Organization Type	System Cards	Dominant Developer	Share
Browserbase	For-Profit	9	Google DeepMind	100.0%
Faculty AI	For-Profit	6	Anthropic	100.0%
Petri / Safety Research	Mixed	5	Anthropic	100.0%
Andon Labs	For-Profit	4	Anthropic	100.0%
VALS AI	For-Profit	3	Anthropic	100.0%
Invariant Labs	For-Profit	3	xAI	100.0%
Oxford Biosecurity	Non-Profit	3	xAI	100.0%
Deloitte Consulting	For-Profit	9	Anthropic	88.9%
Signature Science	For-Profit	9	Anthropic	88.9%
Gryphon Scientific	For-Profit	5	OpenAI	80.0%
Irregular	For-Profit	5	OpenAI	80.0%
UC Berkeley	Academic	8	Google DeepMind	75.0%

Table 7. **Classifiable evaluators are overwhelmingly US- or UK-based.** Country of primary operation for the 125 connected evaluators. The US and UK together account for 59 of the 76 evaluators with a classifiable country (77.6%), and 47.2% of all evaluators, while the country of nearly two in five is unknown.

Country	Evaluators
United States	51
Unknown	49
United Kingdom	8
China	5
Canada	3
Switzerland	3
Singapore	2
Hong Kong	2
Australia	1
India	1