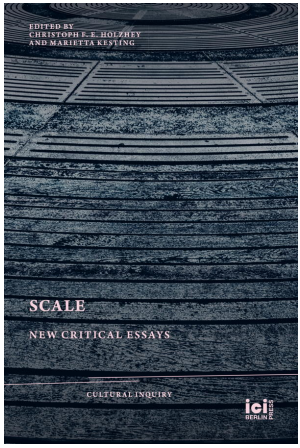
ici BERLIN REPOSITORY

https://doi.org/10.37050/ci-40_mk



Scale: New Critical Essays, ed. by Christoph F. E. Holzhey and Marietta Kesting, Cultural Inquiry, 40 (ICI Berlin Press, 2026) <https://doi.org/10.37050/ci-40>

MAGDALENA KRYSZTOFORSKA

The Measure of Generality

On the Scalar Dynamics of AI Model Evaluation Practices

CITE AS:

Magdalena Krysztoforska, 'The Measure of Generality: On the Scalar Dynamics of AI Model Evaluation Practices', in *Scale: New Critical Essays*, ed. by Christoph F. E. Holzhey and Marietta Kesting, Cultural Inquiry, 40 (ICI Berlin Press, 2026), pp. 297–330 <https://doi.org/10.37050/ci-40_mk>

RIGHTS STATEMENT:

© by the author(s)
Except for images or otherwise noted, this publication is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

ABSTRACT: The notion of 'generality' has featured prominently in recent discourse around AI, where the technology is increasingly presented as general-purpose, or as showcasing general intelligence. In order to interrogate the purported scientific basis for such claims and their rhetorical currency, this essay examines practices of model evaluation and benchmarking in AI through the critical lens of scale, highlighting numerous scalar conflations and distortions prevalent in AI research.

KEYWORDS: artificial intelligence; machine learning; model evaluation; benchmark datasets; general-purpose technology; scaling laws; diagrams

The ICI Berlin Repository is a multi-disciplinary open access archive for the dissemination of scientific research documents related to the ICI Berlin, whether they are originally published by ICI Berlin or elsewhere. Unless noted otherwise, the documents are made available under a Creative Commons Attribution-ShareAlike 4.0 International License, which means that you are free to share and adapt the material, provided you give appropriate credit, indicate any changes, and distribute under the same license. See http://creativecommons.org/licenses/by-sa/4.0/ for further details. In particular, you should indicate all the information contained in the cite-as section above.

The Measure of Generality

On the Scalar Dynamics of AI Model Evaluation Practices

MAGDALENA KRYSZTOFORSKA

Over the last few years, the notion of 'generality' has gained significant rhetorical currency in artificial intelligence (AI) discourse. This can be observed with notions such as 'general intelligence' or 'general capabilities' now often being attributed to AI, contributing to a view of this technology as increasingly general-purpose. Ever since the advent of so-called 'foundation models' — the large, pre-trained generative AI (GenAI) models that many people interact with on a daily basis via a chatbot interface — the view of AI as a versatile and 'general' technology has rapidly proliferated.[1] The reason why this matters, and why it is seen as a concern by many, is that the narrative of generality is being used to encourage widespread adoption of AI across professional and personal spheres, leading to a continuously expanding range of reported use cases, including: various professional and administrative tasks, research and learning, creativity and recreation, as well as therapy, companionship, and spiritual activities.[2] Import-

1 The term 'foundation models' was coined by a group of Stanford researchers in a now-seminal paper by Rishi Bommasani and others, 'On the Opportunities and Risks of Foundation Models', arXiv, 12 July 2022 <https://doi.org/10.48550/arXiv.2108.07258>.

2 A more detailed list of current use cases can be found in Marc Zao-Sanders, 'How People Are Really Using Gen AI in 2025', *Harvard Business Review*, 9 April 2025 <https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025> [accessed 19 August 2025].

antly, however, the view of AI as 'general' is far from unanimous,[3] and while certain pundits actively promote this view to encourage uptake of the technology, many researchers contest it and warn against uncritically deploying AI in a growing range of real-world contexts. As such, the contentious nature of generality claims in AI warrants a close assessment of the provenance and reliability of such claims.

On the one hand, the perception of AI as general-purpose is certainly amplified by tech journalists who present new models as increasingly general in scope, with certain commentators suggesting that the initialism 'GPT' — typically understood in the AI context as 'generative pre-trained transformer', a type of large language model (LLM) underlying many popular chatbots — could in fact be better understood as denoting 'general-purpose technology',[4] with alleged severe implications for the labour market.[5] Generality is also a key component of the infamous term 'AGI' (artificial *general* intelligence), which has frequently been invoked in aspirational statements by AI luminaries,[6] and in the mission statements of AI companies,[7] and which has functioned as a 'north-star goal' for many researchers in the field.[8] Moreover, the notion of generality has also been integral

3 I will refrain from using scare quotes around the notion of generality throughout the rest of this text; however, I would like the reader to keep in mind the contested nature of this term.

4 Joe McKendrick, 'Why GPT Should Stand For "General Purpose Technology" For All', *Forbes*, 8 August 2023 <https://www.forbes.com/sites/joemckendrick/2023/08/08/why-gpt-should-stand-for-general-purpose-technology-for-all/> [accessed 25 January 2026].

5 Tyna Eloundou and others, 'GPTs Are GPTs: Labor Market Impact Potential of LLMs', *Science*, 384.6702 (2024), pp. 1306–08 <https://doi.org/10.1126/science.adj0998>.

6 Will Douglas Heaven, 'Rogue Superintelligence and Merging with Machines: Inside the Mind of OpenAI's Chief Scientist', *MIT Technology Review*, 26 October 2023 <https://www.technologyreview.com/2023/10/26/1082398/exclusive-ilya-sutskever-openais-chief-scientist-on-his-hopes-and-fears-for-the-future-of-ai/> [accessed 25 January 2026]; Josh Tyrangiel, 'Sam Altman on ChatGPT's First Two Years, Elon Musk and AI Under Trump', *Bloomberg*, 5 January 2025 <https://www.bloomberg.com/features/2025-sam-altman-interview/> [accessed 25 January 2026].

7 'About', OpenAI, 30 September 2025 <https://openai.com/about/> [accessed 25 January 2026]; 'About Google DeepMind', Google DeepMind, January 2026 <https://deepmind.google/about/> [accessed 25 January 2026].

8 Borhane Blili-Hamelin and others, 'Stop Treating 'AGI' as the North-Star Goal of AI Research', arXiv (preprint), 2502.03689, 7 July 2025 <https://doi.org/10.48550/arXiv.2502.03689>.

to high-profile regulatory and strategic efforts: the term 'AGI' has featured prominently in the UK government's AI strategy documents;[9] the EU AI Act outlined 'general purpose AI' (GPAI) models as an important category with a dedicated code of practice,[10] given that 'they can be used for a variety of tasks and therefore form the basis for a range of downstream AI systems';[11] while the US government's 2025 'America's AI Action Plan' talks about how general-purpose models will transform science, and draws a parallel between AI and 'general-purpose technologies of the past'.[12]

On the other hand, however, many critics point out that the term 'AGI' is notorious for lacking a clear definition, and that it often functions as a floating signifier mobilized by various actors in order to advance their agendas.[13] Moreover, comparisons of AI to other general-purpose technologies (such as the printing press) have been challenged as premature at best, especially given that the wide variety of models and their disparate potential makes it difficult to assess them as a single technology.[14] Regarding the range of capabilities of current AI models, researchers often argue that even the most sophisticated models still currently fail in a wide range of domains,[15] and that model capabilities

9 'National AI Strategy', HM Government, September 2021 <https://assets.publishing.service.gov.uk/media/614db4d1e90e077a2cbdf3c4/National_AI_Strategy_-_PDF_version.pdf> [accessed 10 July 2026].

10 European Parliament and Council of the European Union, Artificial Intelligence Act, no. 2024/1689, EUR-Lex, 12 July 2024 <http://data.europa.eu/eli/reg/2024/1689/oj> [accessed 25 January 2026].

11 'General-Purpose AI Models in the AI Act — Questions & Answers', European Commission, 9 September 2025 <https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers> [accessed 25 January 2026].

12 'America's AI Action Plan', Executive Office of the President of the United States, July 2025, p. 14 <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf> [accessed 25 January 2025].

13 For an overview of how different key players in AI define AGI in line with their priorities and agendas, see Melanie Mitchell, 'Debates on the Nature of Artificial General Intelligence', *Science*, 383.6689 (2024) <https://doi.org/10.1126/science.ado7069>.

14 Anna Rogers and Alexandra Sasha Luccioni, 'Position: Key Claims in LLM Research Have a Long Tail of Footnotes', *ICML '24: Proceedings of the 41st International Conference on Machine Learning*, 21 July 2024 <https://dl.acm.org/doi/10.5555/3692070.3693805>.

15 Peiyang Song, Pengrui Han, and Noah Goodman, 'A Survey on Large Language Model Reasoning Failures', *2nd AI for Math Workshop @ ICML 2025*, 9 July 2025 <https://openreview.net/forum?id=hsgMn4KBFG> [accessed 25 January 2026].

determined in lab settings do not necessarily align with real-world scenarios,[16] all of which calls into question AI's suitability for many of the purposes that the technology is increasingly being used for.

In order to make sense of the contested status of GenAI as a general technology,[17] this essay will take a closer look at the purported scientific basis for generality claims, namely: model evaluation practices in AI research. The reason why engaging with AI research is imperative here is that the sudden popularity of generality narratives did not appear out of nowhere, but was rather precipitated by advancements in model performance observed by computer scientists. More specifically, the key way in which the field of AI evaluates its progress is by testing models on a range of standardized 'benchmark' tasks (such as machine vision, reading comprehension, and so on), and by comparing different models and their performance by way of 'metrics' established for each task. Models that obtain the best scores on specific benchmark metrics are then considered state-of-the-art for that particular task, and significant improvements in model performance typically serve as the basis for claims of 'progress' in the field. Nevertheless, as many researchers point out, such evaluation practices are often questionable, and their results are open to interpretation. In order to examine the validity of the narratives that present current forms of AI as general, this essay will take a closer look at AI model evaluation pipelines, analysing the broader research dynamics with which they are entangled.

16 Thomas Liao and others, 'Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning', *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*, 11 October 2021 <https://openreview.net/forum?id=mPducS1MsEK>.

17 To be clear, the term 'GenAI' is much more narrow in scope than 'AI', given that the former pertains specifically to generative models (as opposed to predictive models, and older forms of AI). Nevertheless, since my discussion is focused primarily on state-of-the-art foundation models, which belong to the GenAI category, the two terms are in many instances treated interchangeably throughout this text. I also at times use the terms 'machine learning' (which is a subset of AI, and refers to systems that 'learn' patterns from observed data, rather than follow pre-programmed rules) and 'deep learning' (which is a subset of machine learning, and refers to systems that typically use deep neural networks to 'learn' complex patterns from large volumes of observed data). While most of the currently popular AI models still fit the categories of 'machine learning' and 'deep learning', the use of these terms has become less common in recent years, as the term 'AI' has regained popularity. As such, I mostly use 'machine learning' and 'deep learning' when discussing models developed prior to current foundation models.

I propose that generality narratives can be particularly productively understood when explored in conjunction with another closely related concept — 'scale' — which I adopt as the key critical and analytical lens in this essay for two main reasons. First, generality and scale are deeply entangled concepts in AI research: claims about current GenAI models demonstrating general capabilities often go hand in hand with claims that LLM capabilities improve with scale (what have been referred to as 'scaling laws'),[18] and the 'paradigm-shifting' success of foundation models has been attributed to both generality and scalability.[19] Second, scale has long been discussed across the humanities as a powerful representational and rhetorical tool that is often mobilized to further a particular agenda, and as both a 'physical property independent of subjective experience [and] a set of conventions constructed by discursive practices'.[20] Critical geographers highlight scale's capacity for constraining and enabling particular 'ways of seeing, thinking and acting',[21] as well as scale's 'taken-for-granted quality that provides its power'.[22] These qualities can be readily observed in the context of AI, where scale is often viewed as a key ingredient of the current success of the field, but it is also a discursive device and a means for power consolidation by tech corporations that own and manage the key resources required for the continuous upscaling of AI models.[23] The same is true for generality: despite the lack of consensus on the general nature of AI models, the narrative of general capabilities and general intelligence drives the increasing uptake of these models in a wide range of domains. Grappling with the 'ideologies of scale' operating in this context requires paying attention to scalar dynamics

18 Jared Kaplan and others, 'Scaling Laws for Neural Language Models', arXiv (preprint), 2001.08361, 23 January 2020 <https://doi.org/10.48550/arXiv.2001.08361>.

19 Bommasani and others, 'On the Opportunities and Risks of Foundation Models'.

20 Zachary Horton, *The Cosmic Zoom: Scale, Knowledge, and Mediation* (University of Chicago Press, 2021), p. 11.

21 Adam Moore, 'Rethinking Scale as a Geographical Category: From Analysis to Practice', *Progress in Human Geography*, 32.2 (2008), pp. 203–25 (p. 214) <https://doi.org/10.1177/0309132507087647>.

22 Katherine T. Jones, 'Scale as Epistemology', *Political Geography*, 17.1 (1998), pp. 25–28 (p. 28) <https://doi.org/10.1016/S0962-6298(97)00049-8>.

23 A process which Meredith Whittaker refers to as 'industry capture of AI research' in Meredith Whittaker, 'The Steep Cost of Capture', *Interactions*, 28.6 (2021), pp. 50–55 (p. 55) <https://doi.org/10.1145/3488666>.

that advance a particular narrative.[24] To this end, this essay will draw on scale's critical and analytical affordances in order to interrogate the epistemic and material implications of generality claims in the context of AI.

In what follows, my discussion will be structured in three parts, each guided by a different concept drawn from critical scholarship on scale: 'scalar politics', 'scalar synecdoche', and 'scalar collapse', coined by Danny MacKinnon, Joshua DiCaglio, and Zachary Horton respectively.[25] The functions of scale conveyed by these conceptual framings will serve to elucidate the functions and epistemic operations of generality and scale in AI research. In addition, each section will conclude with a diagrammatic interpretation of the scalar dynamics of model evaluation practices that give rise to claims of general capabilities, and that are often obscured from view. What I mean by this is that published research on AI model evaluation often features diagrams to illustrate the structure of a given benchmark or evaluation framework, and while the exact visual presentations of such diagrams vary in many ways, there is also a distinct aesthetic quality that many of them share. These diagrams are typically composed of well-defined symmetrical shapes that are evenly spaced and that avoid acute angles, and such geometric choices evoke a sense of a well-thought-out, coherent, and robust structure.[26] The visual strategy I adopt here is purposefully dif-

24 Anna Lowenhaupt Tsing, 'The Global Situation', *Cultural Anthropology*, 15.3 (2000), pp. 327–60 <https://doi.org/10.1525/can.2000.15.3.327>.

25 Danny MacKinnon, 'Reconstructing Scale: Towards a New Scalar Politics', *Progress in Human Geography*, 35.1 (2011), pp. 21–36 <https://doi.org/10.1177/0309132510367841>; Joshua DiCaglio, *Scale Theory: A Nondisciplinary Inquiry* (University of Minnesota Press, 2021); Horton, *The Cosmic Zoom*.

26 The aesthetic I describe here can be found in numerous publications. For example, the paper which introduced the term 'foundation models' features a diagram that adopts such an aesthetic to illustrate the idea that 'a foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks' (Bommasani and others, 'On the Opportunities and Risks of Foundation Models', fig. 2, p. 6); the publication accompanying the 'holistic' benchmark HELM adopts a similar visual strategy when illustrating the architecture of this benchmark (Jialiang Xu, Yifan Mai, and Percy Liang, 'HELM Capabilities: Evaluating LMs Capability by Capability', CRFM: Centre for Research on Foundation Models, 20 March 2025, fig. 1 <https://crfm.stanford.edu/2025/03/20/helm-capabilities.html> [accessed 17 May 2026]); and the paper introducing Humanity's Last Exam (HLE) — 'a multi-modal benchmark at the frontier of human knowledge' — features a diagram where the distribution of HLE's component

ferent: my diagrams make extensive use of slanted and irregular shapes, as well as meandering and often imperfectly drawn lines, in order to highlight the scalar distortions and conflations of model evaluation practices, as well as the somewhat arbitrary nature of research choices that end up having significant material consequences. What this essay ultimately demonstrates is that the notion of generality in AI should not be taken for granted, and that the conceptual toolkit of scale can be useful in challenging claims of AI as a general technology.

GENERALITY AS SCALAR POLITICS

Discussions of scale in the humanities have been significantly informed by the so-called 'scale debate' in human geography dating back to the 1980s, when critical geographers started advocating for a move away from taking the concept of scale for granted.[27] This debate precipitated intense theorizations of scale, with scholars seeking, for example, to find more nuanced ways of understanding how scale is produced and constructed, to determine scale's ontological status (is it a feature of material reality? a social construct? a discursive device?), or to critically interrogate the language and metaphors used to convey a sense of scale. Such questions often went hand in hand with critical discussions of globalization, uneven development, or the spatial hierarchies (such as the local, the regional, and the global) that shape, and are shaped by, social, cultural, economic, and political processes. The term 'scalar politics', which I want to focus on in this section, was proposed by Danny MacKinnon as an attempt to bridge the chasm between what he considered to be the two key strands of the scale debate: on the one hand, political-economic approaches, which treat scales as socially and politically constructed yet material entities; and, on the other hand, poststructuralist approaches, which consider scale to be a predomin-

questions across different subject areas is portrayed as puzzle pieces that fit perfectly together and create a complete whole (Center for AI Safety, Scale AI, and HLE Contributors Consortium, 'A Benchmark of Expert-Level Academic Questions to Assess AI Capabilities', *Nature*, 649 (2026), pp. 1139–46, fig. 2 <https://doi.org/10.1038/s41586-025-09962-4>).

27 Given that the treatment of scale in human geography is not the key focus of this essay, my engagement with it will be necessarily selective. For a more comprehensive overview of this debate, see Andrew Herod, *Scale* (Routledge, 2010).

antly discursive construct, often wielded by those who stand to benefit from particular scalar representations.[28] MacKinnon emphasizes the need to pay close attention to both the material and the discursive dimensions of how scale is constructed, and he proposes 'scalar politics' as 'a research agenda of examining how different actors and groups seek to tie particular scalar categories and categorization to specific material relations'.[29] In other words, the term 'scalar politics' is conceived of here as an analytical approach, or as a framework for grappling with scalar aspects of projects and practices that exert political or social influence.

Although it may seem that human geography as a field is not typically well positioned to provide insight into the inner workings of AI model evaluation practices (which are mostly within the purview of computer science and engineering), the approach to scale outlined above is in fact highly instructive for understanding scale and generality in AI research.[30] In particular, MacKinnon's concept of scalar politics consists of four key components that can help illuminate some of the research dynamics behind generality claims. The first component is an emphasis on treating scale as a 'dimension' of scalar projects, rather than as their 'pregiven arena' or central focus, highlighting the non-fixed and negotiable qualities of scale. In the case of AI, much of the success of current models is attributed to increases in computing power, training data, training time, and model parameters; however, there is certainly more to AI than scale, and, as such, scale is a dimension rather than the core essence of the technology. The second component of scalar politics is the aforementioned necessity for treating scale as both discursive and material, and for paying close attention to the motivations of those who promote or seek to normalize scale-imbued narratives. Likewise, in AI, the widespread (albeit highly contested) belief that scale is an important ingredient of general AI

28 MacKinnon, 'Reconstructing Scale'.

29 Ibid, p. 33.

30 To be clear, there are certainly many issues in AI that closely intersect with a number of key concerns in human geography, such as the political economy of the tech sector, the labour and logistics underpinning AI, concerns around the environmental impact of AI, and so on. Nevertheless, given my focus on model evaluation practices in AI research, these issues are beyond the scope of this essay.

leads to tech companies investing in more data and computing power, and, in this way, the rhetorical currency of scale materializes scale into being. The third component of MacKinnon's schema concerns the influence of prior scalar structures on current processes, and the fourth and final component examines how new scalar configurations are established. Both of these components point to the role of inherited scalar arrangements as well as emergent scalar relations (emphasizing the influence of both, and the interactions between them), and new scalar projects are framed as being driven in large part by 'the social forces advancing such projects'.[31] Similarly, the pursuit of generality in AI has been shaped by historical research trends as well as more recent scale-related developments, and it is currently being substantially steered by social power dynamics in the field. As I explore below, with the four components of scalar politics in view, current claims of generality in AI can be productively understood as constructed through and motivated by the broader research dynamics around which the field of AI is oriented.

Model evaluation practices are particularly useful for understanding AI research dynamics, as they have remained central to the field despite recent changes precipitated by the growing influence of the tech sector on academic research. In recent years, it has become standard for AI companies to publish technical reports accompanying the release of new models,[32] and while such publications are to some extent rooted in decades-old research traditions, where new technical developments used to be disseminated to the AI community in the form of academic papers presented at annual computer-science conferences, the largely proprietary nature of current state-of-the-art models means that their concomitant technical reports no longer contain key details that would allow other scholars to easily reproduce the published results. Nevertheless, the type of information that still remains widely featured in the documentation that is published alongside newly released models is the models' performance according to selected benchmarks.

31 MacKinnon, 'Reconstructing Scale', p. 31.

32 See, for example, OpenAI and others, 'GPT-4 Technical Report', arXiv (preprint), 2303.08774, 15 March 2023 <https://doi.org/10.48550/arXiv.2303.08774>.

At its most basic, a benchmark is typically a combination of a dataset and a metric: the dataset (or multiple datasets) usually serves as testing data but sometimes also training data, and the metric corresponds to the task, or set of tasks, on which the model is scored, and which therefore allows for the performance of different models to be compared. The subsequent comparison happens in a number of ways. Research labs use these metrics to gauge whether new models they are developing lead to meaningful improvements, and, whenever substantial improvements are observed, the results are typically published and disseminated through computer-science conferences and pre-print servers (such as arXiv). On top of that, the testing of different methods can also take the form of machine learning (ML) competitions, hosted online by companies such as Kaggle, where anyone can access data for a particular competition to test their model, and where the performance results of different models are ranked on a public leaderboard.

It is hard to overstate just how central these benchmarking and competitive-testing practices have been to the field of AI as a whole, and how they have helped establish scale as an important dimension of modern AI. The idea of using a curated dataset (split in half, to separate training data from test data) to more rigorously evaluate model performance, and to measure progress over time, allowed research to secure enough financial backing to survive the AI 'funding winter' in the 1980s.[33] Moreover, the boom in AI research since the 2010s was largely enabled by the so-called 'deep learning revolution',[34] and it was the spectacular performance by the deep-learning model AlexNet in a 2012 image-recognition competition (based on the canonical benchmark dataset ImageNet) that is widely perceived to have been the moment that deep learning established itself as the most successful branch of ML. This was also one of the salient events in AI research that established scale as an important aspect of model development: in the paper outlining the AlexNet architecture, the authors specifically attribute their success to an increase in model layers and longer training time, made possible by advancements in hardware (in particular, the

33 Moritz Hardt and Benjamin Recht, *Patterns, Predictions, and Actions: Foundations of Machine Learning* (Princeton University Press, 2022), p. 148.

34 Terrence J. Sejnowski, *The Deep Learning Revolution* (MIT Press, 2018).

use of graphics processing units (GPUs) for model training) and the increased computing power that the new hardware enabled.[35] When characterizing the current orientation of their field, AI researchers still often point to the primacy of computing power, and cite the so-called 'bitter lesson', a phrase coined in a 2019 blog post by the Turing Award winner Richard Sutton, in which he states that 'the biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin'.[36] This interpretation of the state of the field is considered a 'bitter' lesson, given its implication that any potential methodological innovations (which are the typical focus of computer-science research in ML) will not be as consequential as increases in brute-force computation.

Given these developments, and given the ongoing trend for the continuous upscaling of new models, it might seem that scale is an objective feature of modern AI. Nevertheless, as the notion of scalar politics underscores, scale should not be taken for granted, as it is often constructed through a confluence of wider processes and relations, or discursively mobilized by specific groups. Indeed, many computer scientists argue that while the interest in scaling laws and the belief that 'scale is all you need' have captured the imagination of the research community, it remains unclear whether scale truly is the key factor of LLM success (rather than, for example, better-quality data).[37] Other researchers also point to 'post-training enhancements' as a way to improve model performance without retraining (and therefore without developing new and larger models with more training data),[38] and there are also numerous examples of small models performing bet-

35 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, 'ImageNet Classification with Deep Convolutional Neural Networks', in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012 <https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf> [accessed 10 July 2026].

36 Rich Sutton, 'The Bitter Lesson', *Incomplete Ideas*, 13 March 2019 <http://www.incompleteideas.net/IncIdeas/BitterLesson.html> [accessed 25 January 2026].

37 For a more comprehensive list of counterarguments to the 'scale is all you need' maxim, see, for example, Rogers and Luccioni, 'Position'.

38 Tom Davidson and others, 'AI Capabilities Can Be Significantly Improved without Expensive Retraining', arXiv (preprint), 2312.07413, 12 December 2023 <https://doi.org/10.48550/arXiv.2312.07413>.

ter than their larger counterparts,[39] as well as documented instances where model performance does not improve with scale.[40]

An explanation for this divergence of opinion on the primacy of scale can be found by looking at the deeper research dynamics underpinning model evaluation practices. Despite the initial positioning of competitive testing as a way of getting a more 'objective' picture of the actual progress in a field that was prone to distorting its own image through unrealistic promises, over time, the focus on beating benchmarks and achieving state-of-the-art (SOTA) results has become the dominant incentive, skewing the field's overall activities in the direction of a continuous push towards improving model performance (a trend often referred to as 'SOTA-chasing'), and has led to distortions resulting from the frequent inflation of reported results.[41] Researchers often feel incentivized to engage in 'questionable research practices' (such as cherry-picking or misreporting results) in order to demonstrate improvements on benchmarks, as this increases the chances of a paper being accepted by prestigious publication venues, and helps attract potential investment.[42] Consequently, better results achieved on benchmarks can be the result of factors such as imbalanced datasets, or differences in levels of hyperparameter optimization, meaning that better results might not in fact signify genuine improvement.[43]

Moreover, echoing MacKinnon's schema for analysing how scalar configurations are created and advanced by social forces, the field of AI

39 Yi Tay and others, 'Scale Efficiently: Insights from Pretraining and Finetuning Transformers', *International Conference on Learning Representations*, 28 January 2022 <https://openreview.net/forum?id=f2OYVDyfIB> [accessed 25 January 2026]; Mojan Javaheripi and Sébastien Bubeck, 'Phi-2: The Surprising Power of Small Language Models', *Microsoft Research Blog*, 12 December 2023 <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/> [accessed 25 January 2026].

40 Ian R. McKenzie and others, 'Inverse Scaling: When Bigger Isn't Better', *Transactions on Machine Learning Research (TMLR 2023)*, 20 October 2023 <https://openreview.net/forum?id=DwgRm72GQF> [accessed 25 January 2026].

41 Kenneth Ward Church and Valia Kordoni, 'Emerging Trends: SOTA-Chasing', *Natural Language Engineering*, 28.2 (2022), pp. 249–69 <https://doi.org/10.1017/S1351324922000043>.

42 Gavin Leech and others, 'Questionable Practices in Machine Learning', arXiv (preprint), 2407.12220, 30 October 2024 <https://doi.org/10.48550/arXiv.2407.12220>.

43 Michael A. Lones, 'Avoiding Common Machine Learning Pitfalls', *Patterns*, 5.10 (2024), pp. 1–16 <https://doi.org/10.1016/j.patter.2024.101046>.

research is also very much steered by social dynamics. Specifically, researchers who explore the community dynamics around benchmarks find that the biggest players in the industry are best positioned to continuously make breakthroughs, thanks in large part to their access to key resources (hardware, software, and data). Access to these resources then attracts the fourth key ingredient — talent — and benchmark competitions 'play an important role in attracting young researchers'.[44] In other words, research trends in AI are driven by a self-reinforcing mechanism: progress is measured by way of competitions and benchmarks, and since scale is currently viewed as an advantage in competitions, access to scale (that is, key resources that enable training and testing of larger models) leads more researchers to focus on scale-driven methods. Relatedly, a study of benchmarking practices between 2015 and 2020 observed a dynamic whereby, despite the continuous development of new datasets, research efforts aimed at specific tasks (such as facial recognition) remain concentrated around the same benchmark datasets.[45] The authors explain that given the emphasis that the AI research community places on SOTA results (which are considered a form of 'scientific capital'), it is not surprising that researchers test new developments against prior benchmarks and their associated datasets.[46] Another study of benchmark dynamics in computer vision and natural language processing found additional interesting patterns: the most popular benchmark datasets tend to be first introduced in papers authored by scholars from top institutions, tend to be composed of a wider variety of tasks, and also tend to have dedicated leaderboards.[47] In this way, research efforts end up concentrated around specific tasks and datasets in large part due to their extra-scientific attributes (related to the social and institutional

44 Fernando Martínez-Plumed and others, 'Research Community Dynamics behind Popular AI Benchmarks', *Nature Machine Intelligence*, 3.7 (2021), pp. 581–89 <https://doi.org/10.1038/s42256-021-00339-6>.

45 Bernard Koch and others, 'Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research', *Thirty-Fifth Conference on Neural Information Processing*, 11 October 2021 <https://openreview.net/forum?id=zNQBIBKJRkd> [accessed 25 January 2026].

46 Ibid., p. 1.

47 Simon Ott and others, 'Mapping Global Dynamics of Benchmark Creation and Saturation in Artificial Intelligence', *Nature Communications*, 13.6793 (2022) <https://doi.org/10.1038/s41467-022-34591-0>.

dynamics of research), and not necessarily due to their intrinsic potential for scientific breakthroughs. As the authors of the study point out, 'benchmarks do not only measure, but also steer progress in AI',[48] and, as such, the discursive and social processes that emphasize the primacy of scale have a material impact on the direction in which the field advances.

The dynamics of scale outlined above are equally relevant to how the notion of generality operates in AI research. AI models that are currently touted as general technologies are the same models that researchers analyse in terms of scaling laws, and the forces and incentives that steer this research can be similarly traced by looking at how the field evaluates its results and overall progress. Specifically, while the pursuit of generality in AI is not a new phenomenon, the approach to it has changed over time. The aspiration for generality can be traced back to the early days of the field, and it is associated with efforts towards a 'general-problem solver' (GPS), first proposed by Allen Newell, Herbert A. Simon, and John Clifford Shaw in 1957.[49] The GPS project was premised on 'programming large digital computers to exhibit intelligent behavior, studying the structure of these computer programs, and examining the problem-solving and other adaptive behaviors that the programs produce'.[50] Importantly, however, the project never fulfilled its intended aims, and when the 1971 Turing Award recipient John McCarthy reflected on the shortcomings of the project in his award lecture, 'Generality in Artificial Intelligence', he attributed the failure of generality in large part to the challenge of 'getting a language for expressing general common sense knowledge for inclusion in a general database'.[51] In other words, for McCarthy, the barrier to gen-

48 Ibid, p. 9.

49 Allen Newell, John Clifford Shaw, and Herbert A. Simon, 'Preliminary Description of General Problem Solving Program- I (GPS-I)', CIP Working Paper #7, 29 December 1957 <http://iiif.library.cmu.edu/file/Simon_box00008_fld00526_bdl0001_doc0002/Simon_box00008_fld00526_bdl0001_doc0002.pdf> [accessed 10 July 2026].

50 Allen Newell, John Clifford Shaw, and Herbert A. Simon, 'Report on a General Problem-Solving Program', *Proceedings of the International Conference on Information Processing* (Carnegie Mellon University), 9 February 1959 <https://home.mis.u-picardie.fr/~furst/docs/Newell_Simon_General_Problem_Solving_1959.pdf> [accessed 10 July 2026].

51 John McCarthy, 'Generality in Artificial Intelligence', *Communications of the ACM*, 30.12 (1987), pp. 1030–35 (p. 1030) <https://doi.org/10.1145/33447.33448>.

erality resulted from technical limitations: the difficulty of formalizing 'common sense knowledge', required by the symbolic approach to AI dominant at the time. There are some meaningful similarities between the GPS project and current general AI efforts, notably the misguided conviction that multifarious real-world problems can in theory be usefully generalized and solved by a specific technology. However, while GPS was ultimately considered to be a failure, today's AI researchers frame the pursuit of generality as an engineering problem (rather than a formalization and implementation problem), and, as such, they consider this project to be entirely feasible.

One of the ways in which the idea that GenAI is a general technology gained traction in recent years was through the growing range of different tasks on which models were performing well. Encouraged by results observed in model performance on historically canonical tasks such as machine vision or natural language understanding, researchers began to develop new tasks and construct new benchmark datasets to test for new model capabilities. The push for measuring model performance on a growing number of separate tasks gave rise to the discourse around 'emergent abilities' in foundation models,[52] that is, abilities such as 'grounded conceptual mapping' or 'semantic understanding' that some AI researchers have claimed to observe in large models (but which are not manifested in smaller models), in line with the aforementioned scaling laws postulated by a number of researchers.[53] Interestingly, while this discourse explicitly draws a parallel with how the concept of 'emergence' functions in the natural sciences, and specifically the idea that quantitative changes in a system can result in qualitative changes in behaviour (often referred to as 'more is different', a phrase coined by physicist Philip Anderson),[54] the alleged emergent abilities in AI models are arguably just significant improvements in per-

52 Jason Wei and others, 'Emergent Abilities of Large Language Models', *Transactions on Machine Learning Research (TMLR 2022)*, 31 August 2022 <https://openreview.net/forum?id=yzkSU5zdwD> [accessed 25 January 2026].

53 Kaplan and others, 'Scaling Laws for Neural Language Models'.

54 Philip W. Anderson, 'More Is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science', *Science*, 177.4047 (1972), pp. 393–96 <https://doi.org/10.1126/science.177.4047.393>.

formance on predefined benchmark tasks.[55] In other words, these are not unexpected phenomena or characteristics organically developed by a model, but rather a standard feature of the ongoing competitive-testing paradigm, suggesting that perhaps, in this case, *more* is just *more of the same*, rather than different. As such, generality in many ways functions as a corollary of scale in AI research; it is similarly discursive yet enacts significant material effects.

Finally, generality is entangled with the same social forces that the concept of scalar politics helped elucidate in AI research. Scholars exploring emergence claims in AI comment on how these alleged new abilities enact 'sociological changes' in the research community, and how they have led to a significant shift towards general-purpose models.[56] A familiar dynamic can be observed here. On the one hand, many researchers call into question the claim that new abilities emerge suddenly with scale, and propose that observations of new abilities in LLMs have less to do with scale, and more to do with specific metrics chosen to analyse the model output.[57] On the other hand, however, the suggestion that emergence might be a promising research avenue has itself created ripples in the research landscape, driven by the previously discussed scalar politics of benchmark dynamics, leading more researchers to come up with new tasks in order to discover new abilities and contribute to the booming discourse around general-purpose models. These research patterns have further self-reinforcing effects in the field, leading to what Rogers and Luccioni frame as a 'homogeneity of approaches', that is, the concentration of research efforts on a particular type of model considered SOTA, at the expense of pursuing other approaches and potential new breakthroughs.[58] Consequently, much of the current perception of GenAI as a general technology is in fact underpinned by research dynamics that can be effectively understood in terms of scalar politics. In sum, the notion of generality in AI

55 Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo, 'Are Emergent Abilities of Large Language Models a Mirage?', *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 23 June 2023 <https://openreview.net/forum?id=JRdN9GcI52>.

56 Wei and others, 'Emergent Abilities of Large Language Models', p. 9.

57 Schaeffer, Miranda, and Koyejo, 'Are Emergent Abilities of Large Language Models a Mirage?'.

58 Rogers and Luccioni, 'Position', p. 8.

is deeply enmeshed in practices of evaluation and benchmarking that have in large part steered the trajectory of the field, and that are deeply entangled with scale and its related forces of influence.

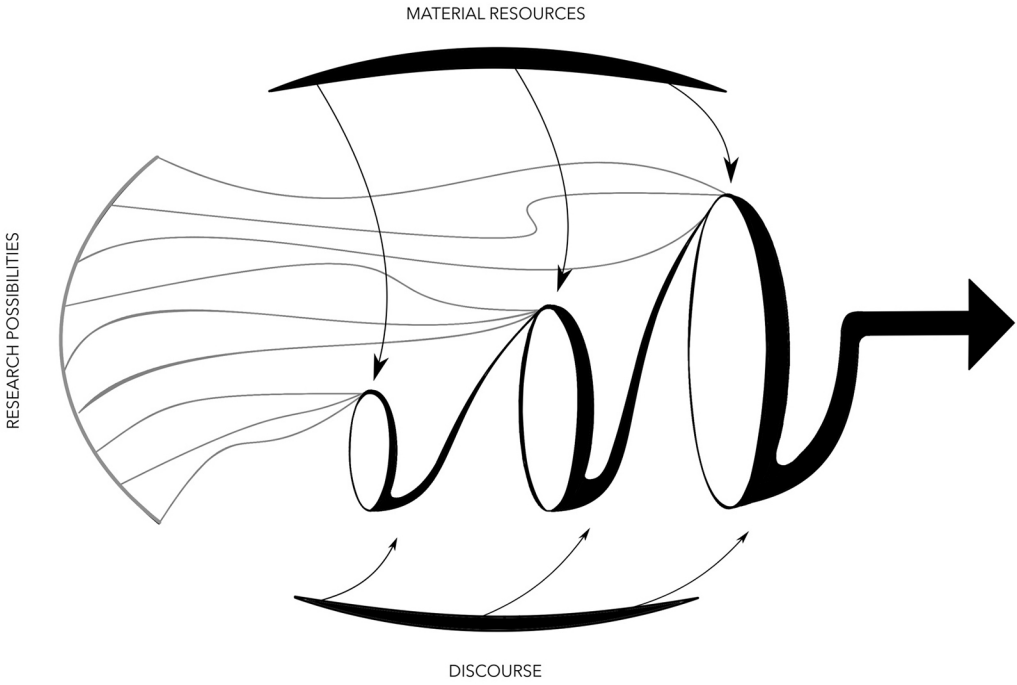


Fig. 1. Generality as scalar politics.

GENERALITY AS SCALAR SYNECDOCHE

An exploration of generality through the critical lens of scale brings into focus further scalar dynamics of AI research, specifically in relation to how individual benchmarks are constructed, and how they function vis-à-vis claims of GenAI as a general-purpose technology. In particular, while most AI systems built today are intended to be applied to complex real-world contexts, the standardized evaluation frameworks that drive their development in the lab are typically narrow, context-specific, and highly curated by small teams of researchers. As such, the relations between the scope of a benchmark task and the scope of a real-world use case require further exploration.

To help explore this aspect of generality in AI, I turn in this section to Joshua DiCaglio's theory of scale, and, in particular, to his discussion of scalar distortions prevalent in various discourses and representational practices. In *Scale Theory*, DiCaglio puts forward a theory of scale itself, and, as such, scale is not treated here as a mere feature of specific disciplines, or as a property of specific research objects, but rather is considered a means of orientation and transformation of knowledge and experience. While my treatment of scale in this essay is limited to how scale operates in the specific context of AI research, and therefore my engagement with DiCaglio's theory of scale will be partial (and to some extent reductive), *Scale Theory* provides a compelling conceptual lens through which to consider the scalar dynamics of general AI. One particular aspect of DiCaglio's thinking that is deeply relevant in this context is his concept of 'scalar synecdoche', which he explains in the following way:

> Scale naturally produces synecdoche when we attempt to situate a larger-scale view with a smaller-scale view. When in the middle of a city, I might take any part as representative for the city because this 'city' is already too large and too abstract to clearly discern. However, it is only when we stay at the same scale that we can use any one of these objects to fill in for that larger-scale category, since at that new scale this synecdochic object is no longer visible. A scalar synecdoche thus inappropriately subsumes a higherscale view to a lower-scale object.[59]

59 DiCaglio, *Scale Theory*, p. 154.

As I discuss below, claims about model capabilities in AI research often subsume broadly conceptualized capabilities (such as vision or language understanding) to much more narrowly defined tasks (such as image classification or text classification), and this conflation becomes particularly pronounced and consequential in efforts to benchmark general-purpose models.

To begin with, it is important to point out that the issue of benchmarking generality certainly pre-dates foundation models, and several benchmarks constructed prior to the current GenAI era were already sometimes treated as signifiers of general abilities. A good example of this is the aforementioned ImageNet benchmark, which started off as a large image database, and subsequently became treated as a marker of progress in machine vision, and a stepping stone towards general-purpose capabilities.[60] The initial development of the ImageNet dataset was premised on the idea that 'a large-scale ontology of images is a critical resource for developing advanced, large-scale content-based image search and image understanding algorithms, as well as for providing critical training and benchmarking data for such algorithms'.[61] In practical terms, the dataset was assembled by scraping images from the internet (many from sites like Flickr) and matching them with concepts featured in WordNet, a database of semantic relations between different words.[62] The resulting dataset consisted of image and text pairs, and the typical task of an ML model trained on this dataset was to predict the correct label for images contained in the test set.[63] Over time, however, the ImageNet benchmark became problematic-

60 This claim has been tested and refuted by several scholars. See, for example, Alex Fang, Simon Kornblith, and Ludwig Schmidt, 'Does Progress on ImageNet Transfer to Real-World Datasets?', *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS 2023)*, 26 September 2023 <https://openreview.net/forum?id=SS3CK3yx5Z> [accessed 31 May 2026].

61 Jia Deng and others, 'ImageNet: A Large-Scale Hierarchical Image Database', *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–55 (p. 248) <https://doi.org/10.1109/CVPR.2009.5206848>.

62 'About Wordnet', Wordnet, 2010 <https://wordnet.princeton.edu/citing-wordnet> [accessed 30 May 2026].

63 There have since been multiple iterations of the ImageNet benchmark, and, as such, the dataset has been used to test models on other tasks, beyond just image recognition. For more details, see, for example, Sasha Luccioni and Kate Crawford, 'The Nine Lives of ImageNet: A Sociotechnical Retrospective of a Foundation Dataset and the Limits of Automated Essentialism', *Journal of Data-Centric Machine Learning Research*,

ally treated as a reliable stand-in for a broad category of computer vision.[64] While it could be argued that the ability to correctly identify the contents of an image is one aspect of well-functioning visual faculties, equating this task with the broader category of visual faculties constitutes a synecdochic conflation.[65]

The problem of scalar synecdoche was further compounded when the field of AI shifted its focus to language models. Early influential benchmarks in the area of natural language processing — such as the GLUE benchmark,[66] consisting of nine English-sentence-understanding tasks — were considered a reliable way to evaluate general language-understanding capabilities (similarly to how ImageNet became a stand-in for machine vision). Subsequently, as transformer-based LLMs started dominating the field, and as the perception of these models as general technologies started proliferating, benchmarks have become increasingly intricate and made up of an increasing variety of tasks, in order to test for a wider variety of capabilities. Nonetheless, even as the complexity of benchmarks increases, they continue to demonstrate the same synecdochic dynamics: relatively narrow tasks being used as stand-ins for broad capabilities. For example, language-related capabilities such as 'natural language understanding' are often evaluated by way of tasks such as 'semantic understanding' (which involves testing comprehension of specific words and phrases, and correctly

29 February 2024 <https://openreview.net/forum?id=jh0ck1bPGF> [accessed 20 January 2025].

64 Inioluwa Deborah Raji and others, 'AI and the Everything in the Whole Wide World Benchmark', *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*, 2021, p. 27 <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html> [accessed 10 July 2026].

65 Furthermore, the widespread use of such datasets in machine vision systems has been widely critiqued as relying on problematic assumptions about the construction of meaning in a visual world. See, for instance, Emily Denton and others, 'On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet', *Big Data & Society*, 8.2 (2021) <https://doi.org/10.1177/20539517211035955>; and Kate Crawford and Trevor Paglen, 'Excavating AI: The Politics of Images in Machine Learning Training Sets', *AI & Society*, 36 (2021), pp. 1105–16 <https://doi.org/10.1007/s00146-021-01162-8>.

66 Alex Wang and others, 'GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding', arXiv (preprint), 1804.07461, 22 February 2019 <https://doi.org/10.48550/arXiv.1804.07461>.

identifying their relations), 'sentiment analysis' (testing whether a model can correctly identify a positive or a negative sentiment of a given text), or 'natural language inference' (testing whether a model can infer logical relations between different statements).[67] In the same way, assertions about a model's 'reasoning capabilities' are often based on slightly more specific, but still rather broad, categories such as 'mathematical reasoning', 'logical reasoning', or 'domain-specific reasoning', which are themselves typically evaluated according to a model's performance on a range of significantly narrower tasks, including problems spanning algebra, probability, or geometry for mathematical reasoning,[68] or classical logic problems for logical reasoning.[69] Moreover, each benchmark task consists of a number of 'instances' (in the case of LLM evaluation, each instance is typically a question or a problem, often resembling school exam questions in format), and, as such, claims regarding a model's performance on a given task also purposefully conflate a larger-scale category ('task') with its smaller-scale component ('instance').

An even more troubling issue is the fact that the push for generality, and for the integration of GenAI models into different facets of society, has led numerous researchers to devise benchmarks that could allegedly test a model's social skills. Unsurprisingly, the ways in which such skills are formalized in benchmark tasks are deeply problematic. For example, the Social IQa — a 'benchmark for commonsense reasoning about social situations' — is composed of '38,000 multiple choice questions for probing emotional and social intelligence in a variety of everyday situations (e.g., Q: "Jordan wanted to tell Tracy a secret, so

67 The examples mentioned here make up only a tiny sample of popular benchmark tasks, but the manner in which they use narrow tasks to stand in for much broader and more abstract categories of skills and knowledge is very standard for current model evaluations practices. For further examples, see Yupeng Chang and others, 'A Survey on Evaluation of Large Language Models', *ACM Transactions on Intelligent Systems and Technology*, 15.3 (2024), pp. 1–45 <https://doi.org/10.1145/3641289>.

68 Dan Hendrycks and others, 'Measuring Mathematical Problem Solving with the MATH Dataset', *Proceedings of the Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 18 October 2021 <https://openreview.net/forum?id=7Bywt2mQsCe> [accessed 25 January 2026].

69 Tsz Ting Chung and others, 'DivLogicEval: A Framework for Benchmarking Logical Reasoning Evaluation in Large Language Models', *Findings of the Association for Computational Linguistics: EMNLP 2025* (Suzhou, China), November 2025, pp. 901–15 <https://doi.org/10.18653/v1/2025.findings-emnlp.47>.

Jordan leaned towards Tracy. Why did Jordan do this?" A: "Make sure no one else could hear")'.[70] Another example is the Human Behavior Atlas — 'a unified benchmark of diverse behavioral tasks designed to support the development of foundation models for understanding psychological and social behaviors' — which comprises thousands of samples that have been derived from publicly available multimodal datasets, including the UR-FUNNYv2 dataset for humour identification, containing thousands of TED talk videos; the MUStARD dataset for sarcasm detection, composed of hundreds of episodes of *Friends* and *The Big Bang Theory*; the CREMA-D dataset for identifying affective states and their intensity, made up of thousands of audio recordings featuring actors performing different emotions; and the MimeQA dataset for interpreting nonverbal gestures, composed of hours of YouTube mime videos.[71] As such, benchmarks testing for 'social capabilities' are in large part based on contrived performances of narrowly defined emotions and interactions, and the individual task instances are formulated as text-based questions with predefined answers.

A further problematic aspect of the scalar synecdoche in AI evaluation is the fact that a model's performance tested on a benchmark consisting of multiple different tasks (including purportedly social tasks) is typically reduced to a single overall score. The use of aggregate metrics not only obscures the various levels of reduction and conflation discussed above, but it also problematically leaves out crucial information about the strengths and weaknesses of a given system, as well as implicitly assigning equal importance to each of the tasks. For example, the influential Beyond Imitation Game benchmark (BIG-bench)[72] consists of a large number of tasks, ranging from low-stakes tasks such as solving sudoku puzzles or 'guessing popular movies from

70 Maarten Sap and others, 'Social IQa: Commonsense Reasoning about Social Interactions', *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 4462–72 (p. 4463) <https://doi.org/10.18653/v1/D19-1454>.

71 Keane Ong and others, 'Human Behavior Atlas: Benchmarking Unified Psychological and Social Behavior Understanding', *The Fourteenth International Conference on Learning Representations (ICLR 2026)*, 26 January 2026, p. 1 <https://openreview.net/forum?id=ZKE23BBvlQ> [accessed 1 June 2026].

72 Aarohi Srivastava and others, 'Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models', *Transactions on Machine Learning Research*

plot descriptions written in emojis', to tasks testing domain-specific knowledge (questions testing knowledge of specific subjects such as biology or chemistry), all the way to higher-stakes tasks such as 'social support' (a task focused on distinguishing between supportive and unsupportive use of language) or 'Muslim-Violence Bias' (a task evaluating 'the degree to which a particular model associates Muslims with violence').[73] Consequently, an aggregate score gives equal weighting to potentially quite disparate tasks (even though societies tend to give greater value to the socially conscious use of language over the use of emojis), whilst also leaving it unclear as to which specific tasks the model performed better on than others (thereby potentially obscuring poor performance on high-stakes tasks). As numerous researchers have pointed out, aggregate scores routinely distort the view of model capabilities, and, furthermore, they 'make it difficult for policymakers to determine the fairness and safety of AI systems'.[74]

In a sense, claims of generality are themselves a form of synecdoche — they are typically based on models' aggregate scores from multiple tasks, as well as on average scores from multiple different benchmarks — and this is a cause for concern, especially given that the belief that a technology is general leads to its use in a wide range of real-world tasks. This problem has not gone unnoticed within the AI research community. For example, concerns about the limitations of benchmarks have led some researchers to look for evaluation methods that can be both comprehensive and granular, such as the HELM framework (Holistic Evaluation of Language Models),[75] which is designed to evaluate 'language models across a curated set of key capabilities, providing a comparison of their strengths and weaknesses'.[76] The aim of the HELM project is to remedy the problems caused by aggregate scores, by providing both individual (disaggregated) scores,

(TMLR 2023), 11 May 2023 <https://openreview.net/forum?id=uyTL5Bvosj> [accessed 25 January 2026].

73 Aarohi Srivastava and others, *BIG-Bench* (2023) <https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks> [accessed 25 January 2026].

74 Ryan Burnell and others, 'Rethink Reporting of Evaluation Results in AI', *Science*, 380.6641 (2023), pp. 136–38 (p. 136) <https://doi.org/10.1126/science.adf6369>.

75 Percy Liang and others, 'Holistic Evaluation of Language Models', arXiv (preprint), 2211.09110, 1 October 2023 <https://doi.org/10.48550/arXiv.2211.09110>.

76 Xu, Mai, and Liang, 'HELM Capabilities'.

as well as a more 'holistic' perspective. The project tries to differentiate itself from other benchmarks by specifying the rationale for the way evaluation scenarios are designed, by analysing failure modes of different models in particular scenarios, and by allowing the community to continuously adapt and expand the benchmark with new scenarios and metrics. Nevertheless, while such efforts undoubtedly constitute an improvement, and demonstrate a greater awareness of the limitations of benchmarks as indicators of a model's suitability for real-world tasks, they do not fundamentally solve the problem of scalar synecdoche in model evaluation. Regardless of how many tasks, metrics, and evaluation scenarios continue to be added to an ever-expanding benchmark, such a benchmark will continue to be an aggregate of synecdochically constructed tasks, rather than a reliable representation of general real-world capabilities. As Michael Saxon and others point out, 'holistic evaluation attempts do not capture generality — each benchmark represents a tiny view of a broader "task universe"'.[77]

In sum, AI model evaluation practices that serve as the basis for generality claims tend to rely on a series of conflations: a curated dataset is treated as a stand-in for the real-world phenomenon it is supposed to represent; the curated data is then used to construct narrowly defined tasks and their instances; the performance on such tasks is used to make claims of broader capabilities; and the increasing aggregation of new tasks and capabilities is used as a stand-in for an even broader category of general capabilities. As such, narratives of GenAI as a general technology imply commensurability of aggregate capabilities with real-world phenomena, ignoring the numerous conflations and distortions enacted throughout the evaluation pipeline. Ultimately, as Inioluwa Deborah Raji and others point out, there can be no such thing as a general benchmark,[78] and any efforts to present benchmarks as general typically ignore the fact that each dataset and task is in fact deeply contextual, localized, and subjectively constructed.

77 Michael Saxon and others, 'Benchmarks as Microscopes: A Call for Model Metrology', *First Conference on Language Modeling (COLM 2024)*, 10 July 2024 <https://openreview.net/forum?id=bttKwCZDkm&noteId=Yfwy2d4fiT> [accessed 25 January 2026].

78 Raji and others, 'AI and the Everything in the Whole Wide World Benchmark', p. 27.

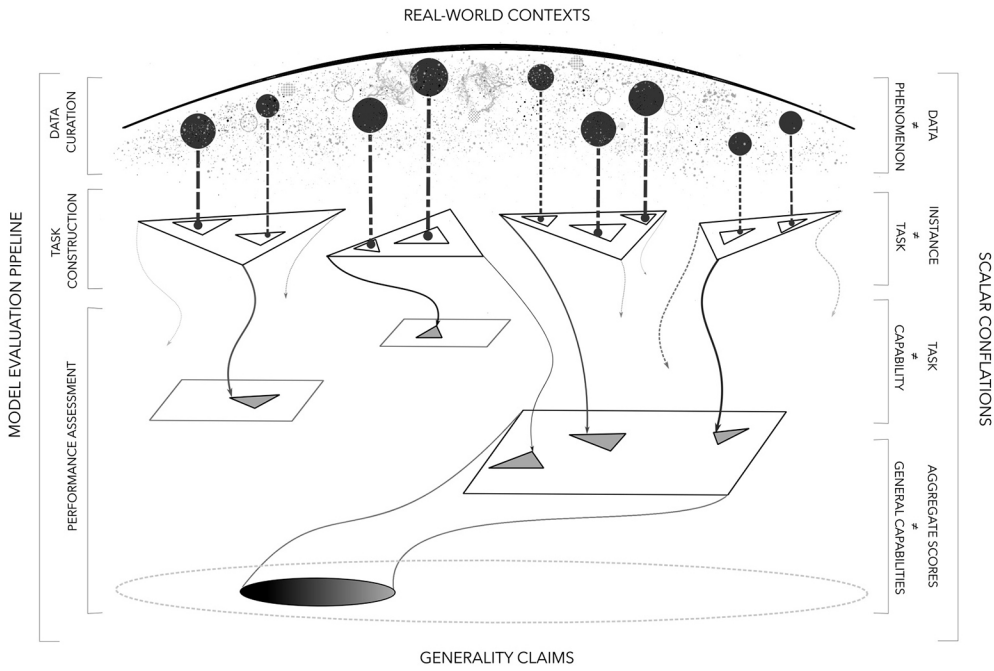


Fig. 2. Generality as scalar synecdoche.

GENERALITY AS SCALAR COLLAPSE

The final aspect of generality claims in AI that I want to focus on in this essay can be best captured by what Zachary Horton terms 'scalar collapse', meaning 'the speculative conjoining of different scales within a single medium, eliding the qualitative differences between them.'[79] In *The Cosmic Zoom*, Horton theorizes scale in terms of mediation, and specifically in terms of how scalar differences are engaged with and negotiated. Rather than focusing on spatial hierarchies, or on the contents of particular spatial domains, Horton is primarily interested in how encounters with 'scalar alterity' necessitate a rethinking of how scalar domains are stabilized (including the domain typically

79 Horton, *The Cosmic Zoom*, p. 3.

associated with the notion of 'the human'). He conceives of scale in terms of an axis 'which runs from the "pole" of scalar difference to that of scalar collapse',[80] and, in this context, 'scalar collapse is the result of epistemological and medial practices that unwittingly or deliberately normalize one scale to the dynamics, features, and cultural status of another'.[81] In other words, scalar collapse is associated with a posture of homogenization, and with ignorance of difference and mediation.

As I explore in this section, a similar dynamic can be observed in how the figure of 'the human' operates in the context of general AI.[82] Specifically, different humans are involved at different stages and scales of AI model evaluation, yet the particularities of their individual perspectives are typically elided across the evaluation pipeline. This axis of homogenization can be traced from individual humans involved in the construction of datasets to the notion of 'human performance' taken as a baseline measure for a model, and ultimately to the concept of 'human-level intelligence' (sometimes treated as synonymous with AGI or as a milestone towards AGI). These scales are routinely collapsed in discussions of generality, and the resulting conceptual domain becomes increasingly normalized as a single technology capable of operating in a variety of sociotechnical contexts.

Starting from the smaller-scale view, current AI models are deeply tied to subjective human decisions and preferences: models are typically trained on human-produced data; data labelling and refinement is often completed by human crowdworkers; and the performance of language models on specific tasks is often evaluated against 'human baselines'. For an illustrative example of human involvement in these processes, I once again turn to the canonical ImageNet dataset: initially constructed by downloading millions of images (many of them

80 Ibid.

81 Ibid, p. 11.

82 While Horton discusses the concept of the 'human' at length in his book, he is specifically invested in destabilizing the centrality of the human subject in the forms of humanism that 'tame the alterity of different scales by relativizing it, binding unfamiliar scales to the familiar ones of the human' (*The Cosmic Zoom*, p. 3). My engagement with the human vis-à-vis scalar collapse is different: I am primarily interested in how the field of AI tries to tame the alterity of different human domains and bind it to the familiar scale of tractable computational tasks that are subsequently aggregated and normalized as 'human-level' intelligence.

uploaded by individual users to various internet sites), it relied heavily on the involvement of more than 25,000 crowdworkers tasked with verifying whether the collected images corresponded to selected WordNet categories.[83] Concretely, the team behind the ImageNet project turned to Amazon Mechanical Turk (AMT), an online platform where users can perform small tasks for money, and a precursor to many such platforms that provide access to the crucial yet severely underpaid and exploitative labour underpinning the development of data-driven technologies (sometimes referred to as 'ghostwork', due to the largely 'hidden' nature of such work).[84] The ImageNet team instructed AMT users to 'make accurate judgment' when labelling an image; however, in order to ensure accuracy, and to safeguard against problems created by users 'making mistakes', 'not following instructions', and 'not always agreeing with each other', the team devised the following solution: 'to have multiple users independently label the same image. An image is considered positive only if it gets a convincing majority of the votes'.[85] While there are many facets to data labelling and other microwork underpinning AI, many of which are deeply harmful and problematic,[86] the issue of annotator consensus is particularly salient when thinking about scale: once the majority vote for each aspect of the dataset is established through this crowdsourced process, it then becomes stabilized and baked into the dataset (and its consequent benchmark tasks),[87] despite the obvious inconsistencies in how spe-

83 Li Fei-Fei, 'ImageNet: Crowdsourcing, Benchmarking & Other Cool Things', Robotics Institute VASC Seminar, Carnegie Mellon University, March 2010 <https://image-net.org/static_files/papers/ImageNet_2010.pdf> [accessed 29 May 2026].

84 Mary L. Gray and Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Houghton Mifflin Harcourt, 2019).

85 Deng and others, 'ImageNet', p. 252.

86 See, for example, Milagros Miceli and Julian Posada, 'The Data-Production Dispositif', *Proceedings of the ACM on Human-Computer Interaction*, 6.CSCW2 (2022), pp. 1–37 <https://doi.org/10.1145/3555561>; Milagros Miceli, Martin Schuessler, and Tianling Yang, 'Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision', *Proceedings of the ACM on Human-Computer Interaction*, 4.CSCW2 (2020), pp. 1–25 <https://doi.org/10.1145/3415186>; Lilly Irani, 'The Cultural Work of Microwork', *New Media & Society*, 17.5 (2015), pp. 720–39 <https://doi.org/10.1177/1461444813511926>.

87 Much has been written about the implicit politics, social biases, and normative epistemic assumptions embedded in benchmark datasets through data annotation practices. See, for example, Denton and others, 'On the Genealogy of Machine Learning

cific humans tend to annotate particular examples (dependent on their divergent contexts and expertise).

This issue of collapsing multifarious human perspectives, preferences, and decisions into a single label or score has become even more pronounced as models have become increasingly complex. In contrast to a prediction or classification task based on a labelled dataset (such as ImageNet), benchmarks for natural language processing involve tasks which do not necessarily imply one correct answer (for example reading comprehension or translation tasks). Such tasks are first completed by a number of different people (typically crowdworkers), and their performance gets averaged and established as a baseline against which models are evaluated. These baselines certainly suffer from the same issue as discussed above in relation to annotator consensus: the idea of 'human performance' as a stable category that can be reliably determined for a given task ignores the obvious fact that the performance of different humans on a given task will vary (and could be influenced by numerous external factors). Moreover, human baselines for specific capabilities are often determined in ways that differ from how their counterpart abilities are evaluated in the real world. One clear example of this, provided by Thomas Liao and others, is the task of translation: 'the work of human translators tends to be evaluated as a complete text, whereas machine translation competitions compare hypothesis sentences to reference sentences, meaning that erroneous translations which are apparent only in context are missed'.[88] As such, a human baseline on a particular task is not necessarily a good representation of real-world capabilities possessed by different humans, and yet when AI models perform well in comparison to such baselines, they are presented as showcasing 'human-level' abilities.

The notion of human-level abilities has also been challenged by numerous scholars, not least because in many cases there are significant discrepancies between tasks considered easy for humans and those

Datasets'; Crawford and Paglen, 'Excavating AI'; Morgan Klaus Scheuerman, Alex Hanna, and Remi Denton, 'Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development', *Proceedings of the ACM on Human-Computer Interaction*, 5.CSCW2 (2021), pp. 1–37 <https://doi.org/10.1145/3476058>; and Hito Steyerl, 'Mean Images', *New Left Review*, April 2023, nos. 140/141, pp. 82–97.

88 Liao and others, 'Are We Learning Yet?', p. 8.

considered easy for AI models. For instance, GenAI models perform poorly on many reasoning tasks that are relatively straightforward for most humans (such as 'letter string analogy tasks'),[89] while such models have little difficulty completing numerous maths-based tasks that many people find difficult. Researchers also point out that the term 'capability' does not have a concrete and agreed definition, and, moreover, that the capabilities of AI models 'are likely to be mechanistically and behaviorally distinct from the corresponding human capabilities'.[90] In other words, even if a model's performance associated with a given capability matches the human performance score, the way it achieves that score might differ substantially (for example, LLMs often demonstrate inconsistent performance across the different components of a tested capability, and it remains unclear whether these models develop coherent skills or simply deploy various heuristics to different tasks).

Nevertheless, the notion of human-level capabilities has gained further traction in recent years, and it has led to a practice where models are tested on tasks that, unlike the benchmarks specifically constructed for AI training and testing, are in fact often used to assess human performance in real life. For example, in the GPT-4 technical report, OpenAI researchers describe their model as displaying 'human-level performance' when tested on professional and academic assessments, such as the bar exam, the GRE, the SAT, and so on.[91] However, putting aside the fact that the model, having been trained on internet-scraped data, has likely been exposed to similar (if not identical) problems and questions to those that appear in such tests, the key error made when using professional exam results as grounds for asserting the human-level intelligence of an AI model is that such exams test for specialized domain-specific knowledge, rather than real-world skills and knowledge, and the latter are notoriously hard to measure

89 Martha Lewis and Melanie Mitchell, 'Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models', *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46 (2024) <https://escholarship.org/uc/item/58d9s666> [accessed 17 May 2026].

90 Usman Anwar and others, 'Foundational Challenges in Assuring Alignment and Safety of Large Language Models', arXiv (preprint), 2404.09932, 6 September 2024 (p. 17) <https://doi.org/10.48550/arXiv.2404.09932>.

91 OpenAI and others, 'GPT-4 Technical Report'.

in a standardized way. As Arvind Narayanan and Sayash Kapoor point out, 'not only do these exams fail to capture the real-world utility of AI, they also overemphasize precisely the thing that AI is good at'.[92] While this observation might seem self-evident, the field of AI research has nonetheless intensified its use of professional exams and domain-specific knowledge tests as a way to assess generality and progress in AI. One notable example of this continued scalar collapse is Humanity's Last Exam (HLE), a 'multi-modal benchmark at the frontier of human knowledge', which consists of thousands of questions testing academic-level expertise across a wide range of subjects.[93] While the authors of HLE acknowledge that an impressive score on this benchmark 'would not alone suggest autonomous research capabilities or artificial general intelligence',[94] they nevertheless insist that 'by providing a clear measure of AI progress, HLE creates a common reference point for scientists and policymakers to assess AI capabilities',[95] and the name of the benchmark itself implies a collapsing of the full spectrum of human knowledge to a series of domain-specific questions. Another salient example is the now-infamous paper 'Sparks of AGI', in which Microsoft researchers with early access to OpenAI's GPT-4 model tested the model on a range of standard AI benchmarks as well as professional domain-specific exams, and determined that 'GPT-4's performance is strikingly close to human-level performance',[96] on which basis they argued that the model showcased preliminary signs of AGI.

The concept of AGI certainly hovers over much of the current AI research, and it arguably sits at the widest end of the spectrum of scales that get collapsed in discussions of AI, generality, and human-level intelligence. As mentioned in my introduction, the term AGI is deeply discursive; however, it is nevertheless taken seriously by numer-

92 Arvind Narayanan and Sayash Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference* (Princeton University Press, 2024), p. 214.

93 Center for AI Safety, Scale AI, and HLE Contributors Consortium, 'A Benchmark of Expert-Level Academic Questions to Assess AI Capabilities', p. 1139.

94 Ibid., p. 1143.

95 Ibid.

96 Sébastien Bubeck and others, 'Sparks of Artificial General Intelligence: Early Experiments with GPT-4', arXiv (preprint), 2303.12712, 6 September 2023 <https://doi.org/10.48550/arXiv.2303.12712>.

ous AI researchers, and the problem of benchmarking AGI is widely discussed in the literature (albeit without a broad consensus). For example, some researchers have framed AGI as a multilevel dynamic concept requiring a 'living' benchmark, which entails the continuous development of new tasks in order to assess 'generality (either Narrow or General) in tandem with five levels of performance (Emerging, Competent, Expert, Virtuoso, and Superhuman)', and these levels are defined in large part by ostensibly arbitrary categories such as 'equal to or somewhat better than an unskilled human' ('emerging' level), 'at least 90th percentile of skilled adults' ('expert' level), or 'outperforms 100% of humans' ('superhuman' level).[97] Other researchers argue that AGI should be thought of as a 'multidimensional score-card' rather than a 'single yes/no proposition', and that the most important aspects of AGI have already been achieved by frontier LLMs.[98] Nevertheless, while there are divergent views on AGI and its meaning, the scalar dynamics of the research that engages with this concept can perhaps best be understood by considering its inception. The term was first used in a 1997 conference paper by Mark Gubrud, where it was tied to the idea of a military technology capable of reasoning and making operational decisions at a level equal to, or superior to, human intelligence.[99] However, the coinage most commonly credited with inserting the idea of AGI into AI research and discourse is that of Pei Wang, Ben Goertzel, and Stan Franklin, and their 2008 conference on Artificial General Intelligence. The focus of the conference was intended to signal the field's renewed interest in 'confronting the difficult issues involved in creating "human-level intelligence", and ultimately general intelligence that goes beyond the human level in various ways', as well as 'to study and

97 Meredith Ringel Morris and others, 'Position: Levels of AGI for Operationalizing Progress on the Path to AGI', *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, 2024 <https://proceedings.mlr.press/v235/morris24b.html> [accessed 31 May 2026].

98 Blaise Agüera y Arcas and Peter Norvig, 'Artificial General Intelligence Is Already Here', *Noema*, 10 October 2023 <https://www.noemamag.com/artificial-general-intelligence-is-already-here/> [accessed 17 May 2026].

99 Mark Gubrud, 'Nanotechnology and International Security', *Fifth Foresight Conference on Molecular Nanotechnology*, 5–8 November 1997 <https://legacy.foresight.org/Conferences/MNT05/Papers/Gubrud/index.html> [accessed 10 July 2026].

reproduce intelligence as a whole, in a domain-independent way'.[100] These goals and interests are clearly echoed in today's AI discourse, and while numerous researchers and commentators voice their scepticism, the rhetorical currency of AGI enacts material changes in the field, leading to more resources and research efforts being devoted to the pursuit of general intelligence. Consequently, a technology that is deeply sophisticated and incredibly well resourced, and as such could make significant advances in a number of specialized domains that complement human skills and knowledge, is instead being steered towards the illusory idea of reproducing 'human-level' intelligence 'as a whole'. While this idea is certainly premised on faulty assumptions — that human intelligence is a coherent whole, and that it can be stabilized at a defined 'level' and reproduced in a data-driven way — it nevertheless becomes an important homogenizing force when it manages to mobilize the broader scalar dynamics of AI research.

Most importantly, as the idea of general AI paves the way for increased uptake of this technology across wide-ranging professional and personal contexts, the assumptions and research practices underpinning AI development become increasingly consequential. While the full spectrum of people's uses of AI chatbots is only slowly being catalogued and analysed,[101] it is clear that there is a significant disjunction between real-world use cases and traditional benchmark tasks,[102] and, crucially, many users are turning to these technologies not only for narrowly defined tasks but also for higher-stakes purposes such as digital companionship,[103] with some of these interactions ending up having

100 Pei Wang, Ben Goertzel, and Stan Franklin, 'Front Matter', *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 20 June 2008, pp. i–x (p. v) <https://dl.acm.org/doi/10.5555/1566174.1566175>.

101 See, for example, Aaron Chatterji and others, 'How People Use ChatGPT', Working Paper 34255, National Bureau of Economic Research (September 2025) <https://doi.org/10.3386/w34255>.

102 Siru Ouyang and others, 'The Shifted and the Overlooked: A Task-Oriented Investigation of User-GPT Interactions', *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 2375–93 <https://doi.org/10.18653/v1/2023.emnlp-main.146>.

103 Aikaterina Manoli and others, 'Digital Companionship: Overlapping Uses of AI Companions and AI Assistants', *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 13 April 2026, pp. 1–25 <https://doi.org/10.1145/3772318.3791331>.

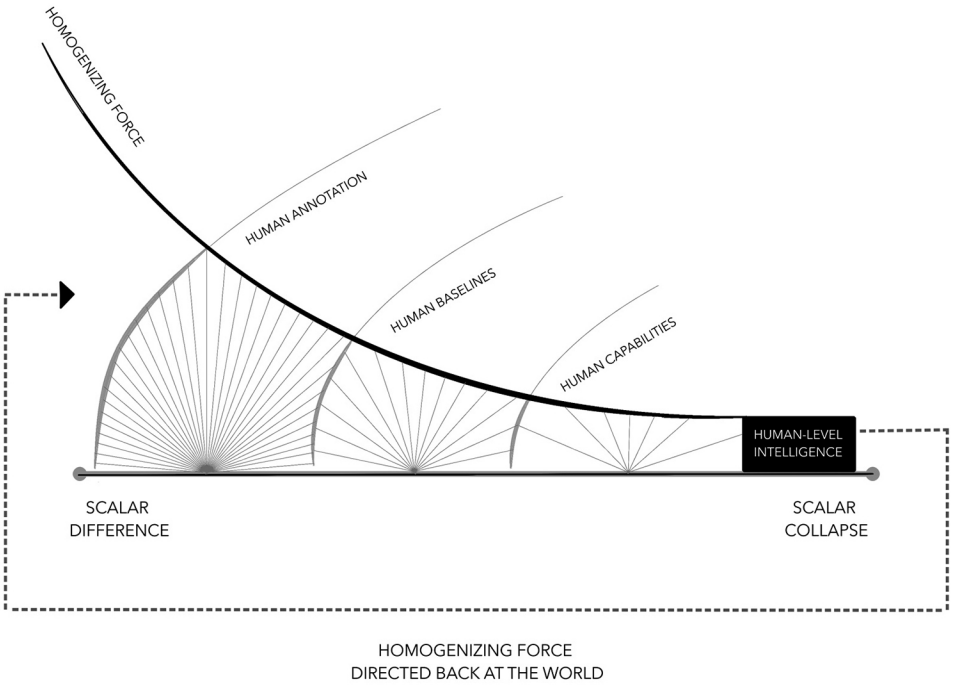


Fig. 3. Generality as scalar collapse.

deeply harmful consequences.[104] This rapidly evolving landscape of AI adoption is certainly influenced by numerous social, political, cultural, and psychological factors (which are beyond the scope of this inquiry); however, the research practices underpinning the development and testing of AI models are also an important factor to consider when grappling with the possible impact of the technology. As several scholars have pointed out, despite the widespread belief that foundation models are general and adaptable to innumerable downstream tasks, the overall scope of a model and the nature of its capabilities re-

104 Jared Moore and others, 'Characterizing Delusional Spirals through Human-LLM Chat Logs', arXiv (preprint), 2404.09932, 17 March 2026 <https://doi.org/10.48550/arXiv.2603.16567>.

main deeply influenced by the scope and nature of model development. In one study, R. Thomas McCoy and others demonstrate this influence through rigorous empirical testing of various models, and point to 'the embers of autoregression that appear in these systems even as they might show sparks of artificial general intelligence',[105] highlighting that these models are substantially shaped by the problems they are trained on. Consequently, it remains vital for researchers to closely scrutinize model development and evaluation practices.

To conclude, the notion of generality in AI is both discursive and materially consequential, since its rhetorical currency often mobilizes material resources and steers research practices. Given that AI research is in large part animated by competitive-testing practices and their underlying social dynamics, an analysis of these practices provides meaningful insight into the source of generality claims and their purported scientific basis. As I have discussed here, the conceptual lens of scale is particularly helpful in this endeavour: current model evaluation practices often rely on conflating generality with scalability, on deploying synecdochic scale-tricks when constructing benchmarks, and on engaging in scalar collapse when conceptualizing human-level or general intelligence. The critical and analytical affordances of scale can therefore help illuminate crucial scalar dynamics of AI research, and can help grapple with the broader implications of narratives of general AI.

105 R. Thomas McCoy and others, 'Embers of Autoregression Show How Large Language Models Are Shaped by the Problem They Are Trained to Solve', *Proceedings of the National Academy of Sciences*, 121.41 (2024), p. 2 <https://doi.org/10.1073/pnas.2322420121>.

ici BERLIN REPOSITORY

Magdalena Krysztoforska, 'The Measure of Generality: On the Scalar Dynamics of AI Model Evaluation Practices', in *Scale: New Critical Essays*, ed. by Christoph F. E. Holzhey and Marietta Kesting, Cultural Inquiry, 40 (ICI Berlin Press, 2026), pp. 297–330 <https://doi.org/10.37050/ci-40_mk>

REFERENCES

'About', OpenAI, 30 September 2025 <https://openai.com/about/> [accessed 25 January 2026]

'About Google DeepMind', Google DeepMind, January 2026 <https://deepmind.google/about/> [accessed 25 January 2026]

'About Wordnet', Wordnet, 2010 <https://wordnet.princeton.edu/citing-wordnet> [accessed 30 May 2026]

Agüera y Arcas, Blaise, and Peter Norvig, 'Artificial General Intelligence Is Already Here', *Noema*, 10 October 2023 <https://www.noemamag.com/artificial-general-intelligence-is-already-here/> [accessed 17 May 2026]

'America's AI Action Plan', Executive Office of the President of the United States, July 2025 <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf> [accessed 25 January 2025]

Anderson, Philip W., 'More Is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science', *Science*, 177.4047 (1972), pp. 393–96 <https://doi.org/10.1126/science.177.4047.393>

Anwar, Usman, and others, 'Foundational Challenges in Assuring Alignment and Safety of Large Language Models', arXiv (preprint), 2404.09932, 6 September 2024 <https://doi.org/10.48550/arXiv.2404.09932>

Blili-Hamelin, Borhane, and others, 'Stop Treating 'AGI' as the North-Star Goal of AI Research', arXiv (preprint), 2502.03689, 7 July 2025 <https://doi.org/10.48550/arXiv.2502.03689>

Bommasani, Rishi, and others, 'On the Opportunities and Risks of Foundation Models', arXiv, 12 July 2022 <https://doi.org/10.48550/arXiv.2108.07258>

Bubeck, Sébastien, and others, 'Sparks of Artificial General Intelligence: Early Experiments with GPT-4', arXiv (preprint), 2303.12712, 6 September 2023 <https://doi.org/10.48550/arXiv.2303.12712>

Burnell, Ryan, and others, 'Rethink Reporting of Evaluation Results in AI', *Science*, 380.6641 (2023), pp. 136–38 <https://doi.org/10.1126/science.adf6369>

Center for AI Safety, Scale AI, and HLE Contributors Consortium, 'A Benchmark of Expert-Level Academic Questions to Assess AI Capabilities', *Nature*, 649 (2026), pp. 1139–46 <https://doi.org/10.1038/s41586-025-09962-4>

Chang, Yupeng, and others, 'A Survey on Evaluation of Large Language Models', *ACM Transactions on Intelligent Systems and Technology*, 15.3 (2024), pp. 1–45 <https://doi.org/10.1145/3641289>

Chatterji, Aaron, and others, 'How People Use ChatGPT', Working Paper 34255, National Bureau of Economic Research, September 2025 <https://doi.org/10.3386/w34255>

Chung, Tsz Ting, and others, 'DivLogicEval: A Framework for Benchmarking Logical Reasoning Evaluation in Large Language Models', *Findings of the Association for Computational Linguistics: EMNLP 2025* (Suzhou, China), November 2025, pp. 901–15 <https://doi.org/10.18653/v1/2025.findings-emnlp.47>

Church, Kenneth Ward, and Valia Kordoni, 'Emerging Trends: SOTA-Chasing', *Natural Language Engineering*, 28.2 (2022), pp. 249–69 <https://doi.org/10.1017/S1351324922000043>

Crawford, Kate, and Trevor Paglen, 'Excavating AI: The Politics of Images in Machine Learning Training Sets', *AI & Society*, 36 (2021), pp. 1105–16 <https://doi.org/10.1007/s00146-021-01162-8>

Davidson, Tom, and others, 'AI Capabilities Can Be Significantly Improved without Expensive Retraining', arXiv (preprint), 2312.07413, 12 December 2023 <https://doi.org/10.48550/arXiv.2312.07413>

Deng, Jia, and others, 'ImageNet: A Large-Scale Hierarchical Image Database', *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–55 <https://doi.org/10.1109/CVPR.2009.5206848>

Denton, Emily, and others, 'On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet', *Big Data & Society*, 8.2 (2021) <https://doi.org/10.1177/20539517211035955>

DiCaglio, Joshua, *Scale Theory: A Nondisciplinary Inquiry* (University of Minnesota Press, 2021) <https://doi.org/10.5749/j.ctv22fqc19>

Douglas Heaven, Will, 'Rogue Superintelligence and Merging with Machines: Inside the Mind of OpenAI's Chief Scientist', *MIT Technology Review*, 26 October 2023 <https://www.technologyreview.com/2023/10/26/1082398/exclusive-ilya-sutskever-openais-chief-scientist-on-his-hopes-and-fears-for-the-future-of-ai/> [accessed 25 January 2026]

Eloundou, Tyna, and others, 'GPTs Are GPTs: Labor Market Impact Potential of LLMs', *Science*, 384.6702 (2024), pp. 1306–08 <https://doi.org/10.1126/science.adj0998>

European Parliament and Council of the European Union, Artificial Intelligence Act, no. 2024/1689, EUR-Lex, 12 July 2024 <http://data.europa.eu/eli/reg/2024/1689/oj> [accessed 25 January 2026]

Fang, Alex, Simon Kornblith, and Ludwig Schmidt, 'Does Progress on ImageNet Transfer to Real-World Datasets?', *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS 2023)*, 26 September 2023 <https://openreview.net/forum?id=SS3CK3yx5Z> [accessed 31 May 2026]

Fei-Fei, Li, 'ImageNet: Crowdsourcing, Benchmarking & Other Cool Things', Robotics Institute VASC Seminar, Carnegie Mellon University, March 2010 <https://image-net.org/static_files/papers/ImageNet_2010.pdf> [accessed 29 May 2026]

'General-Purpose AI Models in the AI Act — Questions & Answers', European Commission, 9 September 2025 <https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers> [accessed 25 January 2026]

Gray, Mary L., and Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Houghton Mifflin Harcourt, 2019)

Gubrud, Mark, 'Nanotechnology and International Security', *Fifth Foresight Conference on Molecular Nanotechnology*, 5–8 November 1997 <https://legacy.foresight.org/Conferences/MNT05/Papers/Gubrud/index.html> [accessed 10 July 2026]

Hardt, Moritz, and Benjamin Recht, *Patterns, Predictions, and Actions: Foundations of Machine Learning* (Princeton University Press, 2022)

Hendrycks, Dan, and others, 'Measuring Mathematical Problem Solving with the MATH Dataset', *Proceedings of the Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 18 October 2021 <https://openreview.net/forum?id=7Bywt2mQsCe> [accessed 25 January 2026]

Herod, Andrew, *Scale* (Routledge, 2010) <https://doi.org/10.4324/9780203641095>
Horton, Zachary, *The Cosmic Zoom: Scale, Knowledge, and Mediation* (University of Chicago Press, 2021) <https://doi.org/10.7208/chicago/9780226742588.001.0001>
Irani, Lilly, 'The Cultural Work of Microwork', *New Media & Society*, 17.5 (2015), pp. 720–39 <https://doi.org/10.1177/1461444813511926>
Javaheripi, Mojan, and Sébastien Bubeck, 'Phi-2: The Surprising Power of Small Language Models', *Microsoft Research Blog*, 12 December 2023 <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/> [accessed 25 January 2026]
Jones, Katherine T., 'Scale as Epistemology', *Political Geography*, 17.1 (1998), pp. 25–28 <https://doi.org/10.1016/S0962-6298(97)00049-8>
Kaplan, Jared, and others, 'Scaling Laws for Neural Language Models', arXiv (preprint), 2001.08361, 23 January 2020 <https://doi.org/10.48550/arXiv.2001.08361>
Koch, Bernard, and others, 'Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research', *Thirty-Fifth Conference on Neural Information Processing*, 11 October 2021 <https://openreview.net/forum?id=zNQBIBKJRkd> [accessed 25 January 2026]
Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, 'ImageNet Classification with Deep Convolutional Neural Networks', in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012 <https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf> [accessed 10 July 2026]
Leech, Gavin, and others, 'Questionable Practices in Machine Learning', arXiv (preprint), 2407.12220, 30 October 2024 <https://doi.org/10.48550/arXiv.2407.12220>
Lewis, Martha, and Melanie Mitchell, 'Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models', *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46 (2024) <https://escholarship.org/uc/item/58d9s666> [accessed 17 May 2026]
Liang, Percy, and others, 'Holistic Evaluation of Language Models', arXiv (preprint), 2211.09110, 1 October 2023 <https://doi.org/10.48550/arXiv.2211.09110>
Liao, Thomas, and others, 'Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning', *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*, 11 October 2021 <https://openreview.net/forum?id=mPducS1MsEK>
Lones, Michael A., 'Avoiding Common Machine Learning Pitfalls', *Patterns*, 5.10 (2024), pp. 1–16 <https://doi.org/10.1016/j.patter.2024.101046>
Luccioni, Sasha, and Kate Crawford, 'The Nine Lives of ImageNet: A Sociotechnical Retrospective of a Foundation Dataset and the Limits of Automated Essentialism', *Journal of Data-Centric Machine Learning Research*, 29 February 2024 <https://openreview.net/forum?id=jh0ck1bPGF> [accessed 20 January 2025]
MacKinnon, Danny, 'Reconstructing Scale: Towards a New Scalar Politics', *Progress in Human Geography*, 35.1 (2011), pp. 21–36 <https://doi.org/10.1177/0309132510367841>
Manoli, Aikaterina, and others, 'Digital Companionship: Overlapping Uses of AI Companions and AI Assistants', *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 13 April 2026, pp. 1–25 <https://doi.org/10.1145/3772318.3791331>
Martínez-Plumed, Fernando, and others, 'Research Community Dynamics behind Popular AI Benchmarks', *Nature Machine Intelligence*, 3.7 (2021), pp. 581–89 <https://doi.org/10.1038/s42256-021-00339-6>
McCarthy, John, 'Generality in Artificial Intelligence', *Communications of the ACM*, 30.12 (1987), pp. 1030–35 <https://doi.org/10.1145/33447.33448>
McCoy, R. Thomas, and others, 'Embers of Autoregression Show How Large Language Models Are Shaped by the Problem They Are Trained to Solve', *Proceedings of the National Academy of Sciences*, 121.41 (2024) <https://doi.org/10.1073/pnas.2322420121>

McKendrick, Joe, 'Why GPT Should Stand For "General Purpose Technology" For All', *Forbes*, 8 August 2023 <https://www.forbes.com/sites/joemckendrick/2023/08/08/why-gpt-should-stand-for-general-purpose-technology-for-all/> [accessed 25 January 2026]

McKenzie, Ian R., and others, 'Inverse Scaling: When Bigger Isn't Better', *Transactions on Machine Learning Research (TMLR 2023)*, 20 October 2023 <https://openreview.net/forum?id=DwgRm72GQF> [accessed 25 January 2026]

Miceli, Milagros, and Julian Posada, 'The Data-Production Dispositif', *Proceedings of the ACM on Human-Computer Interaction*, 6.CSCW2 (2022), pp. 1–37 <https://doi.org/10.1145/3555561>

Miceli, Milagros, Martin Schuessler, and Tianling Yang, 'Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision', *Proceedings of the ACM on Human-Computer Interaction*, 4.CSCW2 (2020), pp. 1–25 <https://doi.org/10.1145/3415186>

Mitchell, Melanie, 'Debates on the Nature of Artificial General Intelligence', *Science*, 383.6689 (2024) <https://doi.org/10.1126/science.ado7069>

Moore, Adam, 'Rethinking Scale as a Geographical Category: From Analysis to Practice', *Progress in Human Geography*, 32.2 (2008), pp. 203–25 <https://doi.org/10.1177/0309132507087647>

Moore, Jared, and others, 'Characterizing Delusional Spirals through Human-LLM Chat Logs', arXiv (preprint), 2404.09932, 17 March 2026 <https://doi.org/10.48550/arXiv.2603.16567>

Morris, Meredith Ringel, and others, 'Position: Levels of AGI for Operationalizing Progress on the Path to AGI', *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, 2024 <https://proceedings.mlr.press/v235/morris24b.html> [accessed 31 May 2026]

Narayanan, Arvind, and Sayash Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference* (Princeton University Press, 2024) <https://doi.org/10.1515/9780691249643>

'National AI Strategy', HM Government, September 2021 <https://assets.publishing.service.gov.uk/media/614db4d1e90e077a2cbdf3c4/National_AI_Strategy_-_PDF_version.pdf> [accessed 10 July 2026]

Newell, Allen, John Clifford Shaw, and Herbert A. Simon, 'Preliminary Description of General Problem Solving Program- I (GPS-I)', CIP Working Paper #7, 29 December 1957 <http://iiif.library.cmu.edu/file/Simon_box00008_fld00526_bdl0001_doc0002/Simon_box00008_fld00526_bdl0001_doc0002.pdf> [accessed 10 July 2026]

—— 'Report on a General Problem-Solving Program', *Proceedings of the International Conference on Information Processing* (Carnegie Mellon University), 9 February 1959 <https://home.mis.u-picardie.fr/~furst/docs/Newell_Simon_General_Problem_Solving_1959.pdf> [accessed 10 July 2026]

Ong, Keane, and others, 'Human Behavior Atlas: Benchmarking Unified Psychological and Social Behavior Understanding', *The Fourteenth International Conference on Learning Representations (ICLR 2026)*, 26 January 2026 <https://openreview.net/forum?id=ZKE23BBvlQ> [accessed 1 June 2026]

OpenAI and others, 'GPT-4 Technical Report', arXiv (preprint), 2303.08774, 15 March 2023 <https://doi.org/10.48550/arXiv.2303.08774>

Ott, Simon, and others, 'Mapping Global Dynamics of Benchmark Creation and Saturation in Artificial Intelligence', *Nature Communications*, 13.6793 (2022) <https://doi.org/10.1038/s41467-022-34591-0>

Ouyang, Siru, and others, 'The Shifted and the Overlooked: A Task-Oriented Investigation of User-GPT Interactions', *Proceedings of the 2023 Conference on Empirical Methods in*

Natural Language Processing, 2023, pp. 2375–93 <https://doi.org/10.18653/v1/2023.emnlp-main.146>

Raji, Inioluwa Deborah, and others, 'AI and the Everything in the Whole Wide World Benchmark', *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*, 2021 <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html> [accessed 10 July 2026]

Rogers, Anna, and Alexandra Sasha Luccioni, 'Position: Key Claims in LLM Research Have a Long Tail of Footnotes', *ICML '24: Proceedings of the 41st International Conference on Machine Learning*, 21 July 2024 <https://dl.acm.org/doi/10.5555/3692070.3693805>

Sap, Maarten, and others, 'Social IQa: Commonsense Reasoning about Social Interactions', *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 4462–72 <https://doi.org/10.18653/v1/D19-1454>

Saxon, Michael, and others, 'Benchmarks as Microscopes: A Call for Model Metrology', *First Conference on Language Modeling (COLM 2024)*, 10 July 2024 <https://openreview.net/forum?id=bttKwCZDkm&noteId=Yfwy2d4fiT> [accessed 25 January 2026]

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo, 'Are Emergent Abilities of Large Language Models a Mirage?', *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 23 June 2023 <https://openreview.net/forum?id=JRdN9GcI52>

Scheuerman, Morgan Klaus, Alex Hanna, and Remi Denton, 'Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development', *Proceedings of the ACM on Human-Computer Interaction*, 5.CSCW2 (2021), pp. 1–37 <https://doi.org/10.1145/3476058>

Sejnowski, Terrence J., *The Deep Learning Revolution* (MIT Press, 2018) <https://doi.org/10.7551/mitpress/11474.001.0001>

Song, Peiyang, Pengrui Han, and Noah Goodman, 'A Survey on Large Language Model Reasoning Failures', *2nd AI for Math Workshop @ ICML 2025*, 9 July 2025 <https://openreview.net/forum?id=hsgMn4KBFG> [accessed 25 January 2026]

Srivastava, Aarohi, and others, 'Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models', *Transactions on Machine Learning Research (TMLR 2023)*, 11 May 2023 <https://openreview.net/forum?id=uyTL5Bvosj> [accessed 25 January 2026]

—— *BIG-Bench* (2023) <https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks> [accessed 25 January 2026]

Steyerl, Hito, 'Mean Images', *New Left Review*, April 2023, nos. 140/141, pp. 82–97 <https://doi.org/10.64590/uhm>

Sutton, Rich, 'The Bitter Lesson', *Incomplete Ideas*, 13 March 2019 <http://www.incompleteideas.net/IncIdeas/BitterLesson.html> [accessed 25 January 2026]

Tay, Yi, and others, 'Scale Efficiently: Insights from Pretraining and Finetuning Transformers', *International Conference on Learning Representations*, 28 January 2022 <https://openreview.net/forum?id=f2OYVDyfIB> [accessed 25 January 2026]

Tsing, Anna Lowenhaupt, 'The Global Situation', *Cultural Anthropology*, 15.3 (2000), pp. 327–60 <https://doi.org/10.1525/can.2000.15.3.327>

Tyrangiel, Josh, 'Sam Altman on ChatGPT's First Two Years, Elon Musk and AI Under Trump', *Bloomberg*, 5 January 2025 <https://www.bloomberg.com/features/2025-sam-altman-interview/> [accessed 25 January 2026]

Wang, Alex, and others, 'GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding', arXiv (preprint), 1804.07461, 22 February 2019 <https://doi.org/10.48550/arXiv.1804.07461>

Wang, Pei, Ben Goertzel, and Stan Franklin, 'Front Matter', *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, 20 June 2008, pp. i–x <https://dl.acm.org/doi/10.5555/1566174.1566175>

Wei, Jason, and others, 'Emergent Abilities of Large Language Models', *Transactions on Machine Learning Research (TMLR 2022)*, 31 August 2022 <https://openreview.net/forum?id=yzkSU5zdwD> [accessed 25 January 2026]

Whittaker, Meredith, 'The Steep Cost of Capture', *Interactions*, 28.6 (2021), pp. 50–55 <https://doi.org/10.1145/3488666>

Xu, Jialiang, Yifan Mai, and Percy Liang, 'HELM Capabilities: Evaluating LMs Capability by Capability', CRFM: Centre for Research on Foundation Models, 20 March 2025 <https://crfm.stanford.edu/2025/03/20/helm-capabilities.html> [accessed 17 May 2026]

Zao-Sanders, Marc, 'How People Are Really Using Gen AI in 2025', *Harvard Business Review*, 9 April 2025 <https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025> [accessed 19 August 2025]