Characterizing Model Jaggedness Supports Safety and Usability

Meredith Ringel Morris[1], **Dan Altman**[1], **Haydn Belfield**[1], **Arthur Goemans**[1], **Hasan Iqbal**[1], **Ryan Burnell**[1], **Iason Gabriel**[1], **Samuel Albanie**[1] **and Allan Dafoe**[1]
[1]Google DeepMind

Frontier AI models exhibit a paradoxical and uneven profile of competencies. They can achieve expert-level performance on many challenging tasks while failing at others that are simple for most people. In other words, they are *jagged*. In this paper, we argue for the importance of accurately characterizing and measuring the jaggedness of frontier models, and propose an approach for doing so. Our proposed jaggedness profiles and metrics have direct implications for safety, governance, and usability, as well as for changing the way we might measure progress toward milestones such as AGI.

1. Introduction

Frontier AI models have made rapid advances in capabilities AI Security Institute (2025); Maslej et al. (2025). Many experts anticipate that models may soon meet or exceed thresholds for consideration as Artificial General Intelligence (AGI) Bengio et al. (2024); Grace et al. (2025); Kaplan et al. (2020); Morris et al. (2025); Murphy et al. (2025). Implicit to the concept of AGI is an assumption of *generality* – that advanced AI will be highly capable across a broad set of tasks, analogous to humans' broad capabilities. In practice, the notion of generality is complicated by *jaggedness* Dell'Acqua et al. (2023), the phenomenon where AI exhibits strong ability spikes in some domains while remaining deficient in others. Generality can be understood as a *general intelligence floor*, a measurable baseline that serves as a leading indicator for the potential emergence of ability *spikes* despite lagging *valleys* of lower performance. The degree of extremity evidenced by spikes and valleys – and the particular combination of abilities and deficits they represent – may have significant consequences for society. We posit that it is vital to characterize the jaggedness of frontier models in order to improve AI safety and usability.

Moravec's Paradox Moravec (1988) observed that AI systems can possess superhuman competence in some tasks while struggling with other seemingly basic tasks, finding that high-level reasoning tasks required relatively little computation compared to low-level perception tasks. Brown et al. (2020) studied the uneven performance of GPT-3, showing that while scaling parameters to 175 billion enabled strong few-shot performance on complex language tasks, it failed to produce reliability in simple arithmetic operations. Researchers have since developed frameworks to measure the generality of AI systems' performance, distinguishing between capability in narrow tasks and breadth of competence Hernandez-Orallo et al. (2021). Zhou et al. (2025) proposed scales of cognitive difficulty as an alternative to tracking benchmark performance, revealing ability profiles that map the unevenness of model performance. Dell'Acqua et al. (2023) characterized this uneven performance as "jaggedness," and investigated the consequences for worker productivity, observing that for knowledge workers, the distinction between automatable and non-automatable tasks does not correlate with human-perceived difficulty. The term *jaggedness* has since gained traction in AI discourse Hammond (2025); Mollick (2025); Patel (2025); Toner (2025).

Empirical analysis indicates that jaggedness is a structural property of current architectures and scaling paradigms. Wei et al. (2022b) described emergent abilities of LLMs that cannot be predicted through simple extrapolation of smaller models' performance, though other researchers

Corresponding author(s): merrie@google.com

have since argued that emergent abilities are artifacts of non-linear evaluation metrics Schaeffer et al. (2023). These findings align with foundational work by Szegedy et al. (2013), who demonstrated that backpropagation leads to “nonintuitive characteristics and intrinsic blind spots" in neural networks.

Today’s generative models have rapidly improved in many complex areas such as writing and coding AI Security Institute (2025); Maslej et al. (2025), yet still fail to accurately perform cognitive tasks that are relatively simple for most humans Knoop (2025); Mukhopadhyay et al. (2025); Rahman and Mishra (2025). Jaggedness is an inherently anthropocentric concept; we perceive AI as jagged to the extent that its profile of strengths and weaknesses is alien to our own. Despite its practical import, jaggedness is not yet measured, and its implications for AI safety and usability have not been extensively considered.

In this paper, we first suggest three capability frameworks for conceptualizing and profiling model jaggedness, and explain the advantages and disadvantages of each for different audiences. Next, we introduce a set of metrics that can be used to characterize and quantify jaggedness, including *jaggedness profiles* for AI models and a *jaggedness index* that collapses this to a simple number. We also introduce important comparisons jaggedness metrics allow for, such as tracking a particular model’s jaggedness trends over time and observing possible differences between the *benchmarked jaggedness* of a model versus the *perceived jaggedness* actually experienced by end-users. Finally, we explore the implications of jaggedness for AI safety, AI policy and governance, end-user experience, and measurement of AI progress towards key milestones such as AGI. By formalizing the concept of jaggedness and identifying ways in which this concept has utility for AI developers, policymakers, and end-users, this paper aims to encourage more discussion of, research on, and measurement of jaggedness by the scientific community.

2. Capability Frameworks for Jaggedness

To characterize jaggedness, we must consider the set of areas across which we will contrast model performance. We propose three high-level capability frameworks that are useful for characterizing jaggedness: *cognitive abilities*, *practical skills*, and *deployed impacts*. Each of these has benefits and tradeoffs; the optimal choice of capability framework (or combination of several) may vary for different user groups and types of assessment, as we discuss below. Benchmarking a model’s performance using a particular capability framework creates a *capability profile*; in Section 3, we will discuss how to transform a simple capability profile into a *jaggedness profile*, which in turn supports analyzing metrics such as a model’s *jaggedness index*.

Cognitive Abilities: Cognition-based capability frameworks focus on comparing model capabilities to domains of human cognition and human cognitive skills such as reasoning, memory, or world knowledge Hendrycks et al. (2025b). These are drawn in a principled way from studies of comparative cognition. A pragmatic advantage of leading cognitive frameworks, such as Cattell-Horn-Carroll theory, is having a manageable number of about ten to twenty top-level categories Carroll (2009); Schneider and McGrew (2012), which makes them particularly well-suited for conveying jaggedness – too few categories might conceal important spikes and dips in abilities, but too many categories makes characterizing jaggedness unwieldy, particularly if the goal is to quickly convey information to end-users through a format such as a model card Mitchell et al. (2019a). Cognitive ability capability frameworks align with cognitive formulations of AGI (e.g., Hendrycks et al. (2025b)), and so may be particularly valued by stakeholders such as industry professionals and academics as a way to gain more nuanced insight into progress toward AGI-related milestones. On the other hand, the use of cognitive frameworks might create an anthropomorphization trap that leads us to only focus on peaks and valleys in human-like cognitive attributes, and may lead us to overlook important non-human

skills that models may possess. Further, without research that validates the construct validity of cognitive skills for AI models (i.e., verifying that passing cognitive benchmarks allow AI models to perform well on practical, real-world tasks), it may be difficult for end-users and policymakers to interpret whether strong peaks or valleys in particular cognitive skills will translate to real-world utility and impacts.

Practical Skills: In contrast, capability frameworks that focuses on practical skills (such as writing ability, mathematical ability, coding, social-emotional skills, creativity, etc.) may be more interpretable to end-users when deciding whether a model is appropriate for a particular task, or to technology companies when deciding what applications or products particular models are best suited toward. Practical skill capability frameworks also align well with many current skill-based ML benchmarks (e.g., Hendrycks et al. (2021); Liang et al. (2023); Srivastava et al. (2023)), which may allow for easier near-term quantification of model jaggedness without investing in benchmark development. Many benchmarks rely on datasets of professional skills such as ONET (232 work activities) U.S. Dept. of Labor Statistics (2025) or ESCO (14,000 skills) European Commission (2025); one challenge of adopting a framework based on practical skills is that such a set may either be impractically large and therefore difficult to enumerate and interpret, or if a limited subset is proposed those constraints may not be fully principled. Further, practical-skill frameworks may be less stable over time or across cultures, particularly for professional skills, since education and employment systems may change as a result of AI adoption.

Deployed Impacts: Finally, we could characterize jaggedness based on capabilities that correspond to the deployed impacts of models. For example, this might include analyses such as Appel et al. (2025); Chatterji1 et al. (2025); Zhao et al. (2024) that characterize the actual distribution of tasks by commercial or personal consumers of AI models. This might be particularly valuable for policymakers, though is likely to be unstable, given that the economic value of certain tasks inevitably varies according to time and place, and will likely be profoundly impacted by the deployment of advanced AI in economic sectors. Characterizing jaggedness based on impact across economic tasks will by definition be a lagging framework, which may reduce its practical utility.

3. Measuring Jaggedness

Next, we outline an approach for developing measures of jaggedness and reflect on their interpretation, utility, and limitations. In this paper, we focus on the theoretical contribution of identifying methodologies for jaggedness quantification; the empirical estimation of jaggedness for specific models is a critical area for future research.

3.1. Establishing the Baseline

A jagged model's performance on a given task varies markedly from some reference performance; we must first define this baseline. We propose a normative approach modeling jaggedness relative to a human population. This is an idea well understood in AI literature Hendrycks et al. (2025b); Morris et al. (2025) and is particularly relevant for jaggedness since many safety and usability concerns relate to the differentials between human and model capabilities.

Various human performance baselines are possible for a given task; the two most salient are the typical, aggregate performance of a representative sample drawn from the entire (adult) population or else a sample of experts. Comparing against average adult performance is an important benchmark and one for which data and estimates may be more readily available. However, comparison to expert populations may be particularly useful for practical skills that have safety or economic implications.

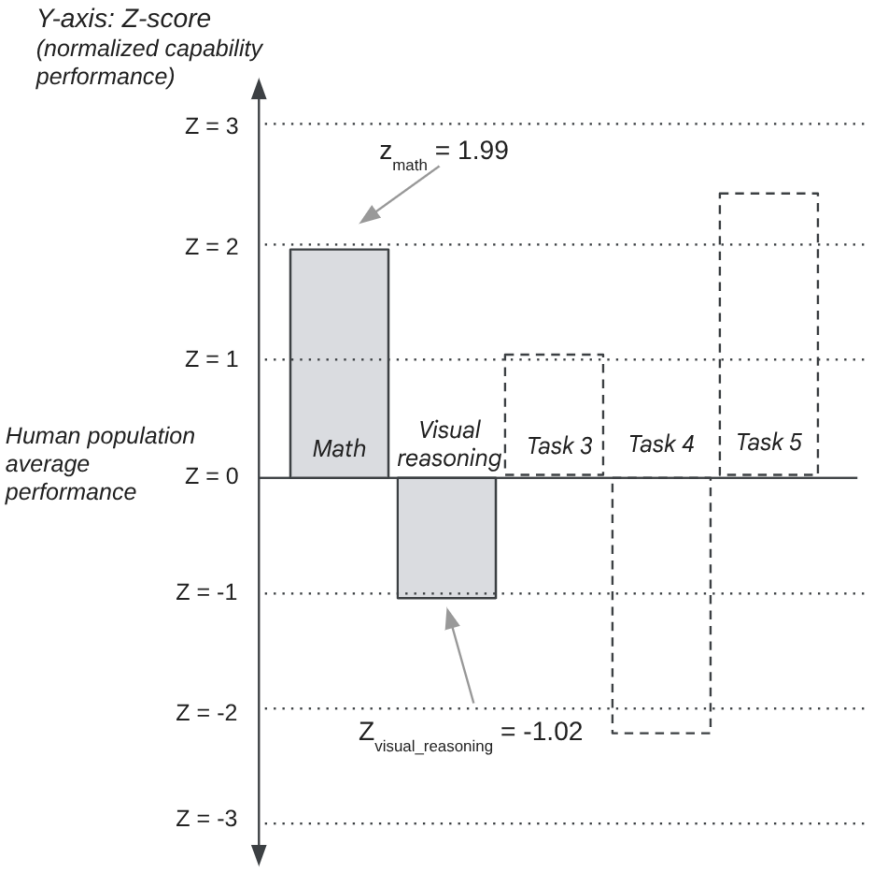

Figure 1 | Example *jaggedness profile* for five tasks in a hypothetical skill-based capability profile. The solid gray bars have been calculated for Gemini 2.5 Pro on the AIME 2025 and ARC-AGI-1 benchmarks. The dashed-bars are purely illustrative. The horizontal dashed lines represent z-scores that are integer value standard deviations above and below human average performance.

In these cases, it is relevant to compare to humans that perform the given task, for example human expert coder performance when considering cybersecurity implications.

Note that humans are also “jagged" in the sense of having uneven skill distributions – no single person is equally good at everything. For this reason we consider aggregates of people as the baseline; however, even aggregates of humans will exhibit peaks and valleys across a particular capability profile. Concerns about model jaggedness relate to models having markedly different distributions of skills and deficits than the human baseline. Thus, a model that exhibits a similar pattern of peaks and valleys to the human baseline is considered *smooth*; models are *jagged* only insofar as they differ from typical human capability distributions.

It is worth noting that our discussed approach is based on a conceptual simplification that measures jaggedness of the *model layer* relative to an *unaided human population*. In practice, both AI models and humans will likely have access to tools such as code generators, calculators, search engines, etc., which could substantially alter their performance. Measuring the jaggedness at the *system layer* that includes interfaces and tools may yield further valuable insights, particularly relating to usability, as we discuss later.

3.2. Jaggedness Profiles

Having established the baseline for normalization, we are now in a position to develop the *jaggedness profile*. An important feature of frontier models is that they may exhibit vastly superhuman spikes at certain tasks, creating extreme outliers in the *capability profile* (see Section 2). Furthermore, the assessment of jaggedness relative to a human baseline means that any non-human tasks may cause discontinuities in the calculation of jaggedness measures. The extent to which these issues are salient is highly dependent on the chosen capability framework and consequently the method chosen to address them should also be considered within the context of a given capability profile. The ideal is to ensure that the jaggedness profile related metrics are robust against outliers in model performance and don't return values of infinity (e.g., where humans cannot undertake a task), while providing a standardized scale (based on a human baseline) to enable intuitive interpretation of the measures developed. One approach would be to utilize a method based on median absolute deviation (MAD), which is robust to outliers as it uses median values. Another approach is based on Winsorization (or clipping) of z-scores (based on standard deviation) to create a standardized scale and solve for outliers, although this does impose a somewhat arbitrary ceiling; this could be addressed through tethering the ceiling to a meaningful human-scale measure e.g., performance in the 99th or 99.9th percentiles (corresponding to normal distribution z-scores of 2.3 and 3.1 respectively).

To exemplify the development of jaggedness measures, we proceed here with the latter standard deviation based approach. The normalized capability performance is:

$$z_i' = \max(-C, \min(C, \frac{x_i - \mu_{i-human}}{\sigma_{i-human}})) \quad (1)$$

Where:

- x_i is model performance for task i
- $\mu_{i-human}$ is the human population average for task i
- $\sigma_{i-human}$ is the human population standard deviation for task i
- C is the Winsorization limit (e.g., $C = 3$)

This gives us z_i' as the Winsorized z-score for task i, defined as the raw z-score clipped at a threshold C (e.g., $C = 3$). We see that z_i' represents the number of (human population) standard deviations the model performs above or below the human (or expert human) population average. As a concrete example, take the performance of the frontier AI model Gemini 2.5 Pro on the practical skills of math benchmarked on AIME 2025 and visual reasoning benchmarked on ARC-AGI-1 ARC Prize (2026); Art of Problem Solving (2025); Gemini Team, Google (2025a,b); LeGris et al. (2024). We calculate the values $z'_{\text{math}} = 1.99$ and $z'_{\text{visual\_reasoning}} = -1.02$ (Table 1). They are particularly instructive in demonstrating jaggedness in *opposite directions*. Model performance on the math task is 1.99 standard deviations *above* the human (expert) average, corresponding to approximately the top 3% of human expert performance, whereas model performance on the visual reasoning task is 1.02 standard deviations *below* the human average, corresponding to approximately the bottom 16% of human performance, where we assume a normal distribution for the human population. For clarity, these are being shown as illustrative of the method and not as an endorsement of these particular benchmarks. Figure 1 illustrates the concept of the jaggedness profile.

3.3. The Jaggedness Index

A jaggedness profile represents the normalized pattern of peaks and valleys of a model with respect to a general or expert human performance baseline, which measures and contextualizes its distinct

Benchmark: model	Task	μ_{human}	σ_{human}	x (model)	z-score (model)
AIME 2025: Gemini 2.5 Pro	Math	6.24	3.5	13.20	**1.99**
ARC-AGI-1: Gemini 2.5 Pro	Visual reasoning	64.2%	22.8%	41%	**-1.02**
AIME 2025: Gemini 3 Pro	Math	6.24	3.5	14.25	**2.29**
ARC-AGI-1: Gemini 3 Pro	Visual reasoning	64.2%	22.8%	75%	**0.47**

Table 1 | A worked example calculating model z-scores for two models using human baseline data from two benchmarks.

Model	$\bar{z}'$	J	ΔJ
Gemini 2.5 Pro	0.49	1.50	*Not calculated*
Gemini 3 Pro	1.38	0.91	-0.6

Table 2 | A worked example comparing average z-score, J, and J's temporal trend of two models

strengths and weaknesses. There is also utility in producing a summary measure for the overall degree of jaggedness of the model across all tasks in its jaggedness profile, which we call the *jaggedness index*.

3.3.1. Index Calculation

The jaggedness index characterizes how spread out the normalized capability measures are in the jaggedness profile. Building on the example jaggedness profile in Figure 1, the jaggedness index can be expressed as the standard deviation of the collection of model z_i-scores that make up the jaggedness profile. Therefore, the calculation of the jaggedness index becomes:

$$J = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i' - \bar{z}')^2} \tag{2}$$

Where:

- J is the jaggedness index
- N is the number of tasks in the capability and jaggedness profiles.
- z_i' is the Winsorized z-score for task i, as defined previously as the raw z-score clipped at a threshold C (e.g., $C = 3$)
- $\bar{z}'$ is the mean of the Winsorized z-scores

The jaggedness index is then a scalar value in the range between 0 and C. A value of 0 represents a perfectly "smooth" model profile even where the performance level is different from that of the human baseline. A system can be worse *or* better than humans and still have a jaggedness index value of 0 as long as the model is equally worse or better at *every task* in the capability profile. In essence this formulation is a measure of the spread of the jaggedness index. Tasks in the underlying profile that are jagged in opposite directions do not cancel out; thus, a model that significantly outperforms humans in one task and significantly underperforms humans in another would not tend to reduce the jaggedness index to 0. Compared to the jaggedness profiles, the compression into a single value loses detailed information; however, by creating a scalar summary statistic of jaggedness with known properties, we are able to better evaluate hypotheses related to jaggedness, as we discuss below.

3.3.2. Temporal Trends

Understanding the rate and direction of changes in jaggedness across model generations may be particularly useful as a signal for both safety and usability concerns. The jaggedness index allows us to quantify this by measuring its change in value from one model generation to the next. The *temporal jaggedness trend* can be represented as:

$$\Delta J_t = J_t - J_{t-1} \tag{3}$$

A near zero change would indicate that a model's jaggedness remains stable. A positive result indicates that the model is becoming more jagged in performance across tasks, whereas a negative result indicates the model is becoming smoother in its performance. Extending this analysis to the sample profile established for Gemini Pro 2.5 in Section 3.2, we can make a comparison to Gemini Pro 3, calculating the respective jaggedness indices and obtaining the temporal trend (Table 2). In this simplified scenario (using only two tasks) the average z-score increases from 0.49 to 1.38, corresponding to an increase in overall model capability relative to a human baseline, and the jaggedness index change of -0.60 indicates that Gemini's capabilities are becoming smoother. This can be understood as a greater improvement in visual reasoning relative to math acting to smooth out the profile.

Large changes in the temporal trend of the jaggedness index suggest that the relative performance on the tasks in the underlying capability profile are diverging markedly, which may be a leading indicator for AI safety concerns.

3.4. Utilizing Jaggedness Measures

One application of jaggedness metrics for deployed AI models can be to aid end-users in better calibrating their expectations of the way the models behave. Additionally, developers may find utility in computing these at various stages of the model development lifecycle to provide insights into how various training methods or certain data sets impact jaggedness and therefore give early signals into model performance and downstream safety and usability impacts. For instance, it may be that the use of more or differing pre-training data may tend to smooth out the jaggedness profile or that different post-training approaches such as reinforcement learning from verifiable rewards (RLVR) or supervised fine tuning (SFT) may increase jaggedness by promoting hill-climbing against specific goals.

As noted in Section 3.1, jaggedness can be measured at the model or system layers, with the latter adding complexity. Usability and safety issues are heavily dependent on other system and deployment considerations such as the ability to call tools and the nature of the user interface. Therefore, assessing jaggedness at the system layer may yield novel insights. In particular, it may be valuable to investigate differences in the *benchmarked* jaggedness profile of a model and that of the jaggedness profile that is *perceived* by end-users. This comparison can be an indicator of usability challenges with deployed systems if the two profiles are markedly different. Sources of such differences may result from end-users' inability to achieve benchmarked model performance due to usability issues or knowledge gaps (e.g., difficulty prompting). However, differences may also point to a lack of ecological validity with the benchmark design, i.e., the range of real-world tasks not being accurately represented; disparities between benchmarked and perceived jaggedness may motivate evolving the benchmarks used to generate the underlying capability profile.

4. Discussion

In this paper, we have argued for the importance of characterizing the jagged profiles of AI models, and have introduced frameworks and metrics that can support understanding the current state of

model jaggedness and how it is changing over time. Next, we reflect on the implications of jaggedness for AI safety, AI policy and governance, user experience and AI literacy, and measuring AI progress.

4.1. Implications for AI Safety

To understand the capability and safety profiles of new models, the AI ecosystem currently relies on a complex, non-standard array of benchmarks, safety documentation (e.g. system cards), third party evaluation, and public testing. Many conclude that this patchwork is suboptimal Lambert (2024). To understand subtle nuances between models' safety strengths and weaknesses, observers often rely on individual blog posts or inconsistent comparisons of documentation.

While not a complete solution, jaggedness profiles and metrics would be helpful tools for allowing standardized comparison between models. For example, using a skills-based profile for Model X might show a high spike in persuasion and a low valley in factuality, thus presenting a different risk than Model Y with high cyber-offense capabilities but low resistance to jailbreaking. This illustrates how characterizing jaggedness can help us move to understanding models as having measurably different safety profiles.

A model with high spikes in coding or scientific reasoning but low contextual awareness could signal the "Golem Problem" of jaggedness – a system with superhuman capabilities in a narrow domain but lacking the "common sense" to operate safely McElreath (2020). This could raise risks of misuse and accidents, for example if the AI system pursues a narrowly defined objective and does not anticipate the consequences. Conversely, a highly jagged AI system may be a specialist that lacks the broad competence to autonomously execute complex, multi-domain plans across domains, thus remaining on dependent on human intervention to bridge its capability gaps. These capability limitations (and the increased human oversight they would likely engender) arguably pose a lower risk of autonomous loss-of-control compared to a smooth agent. Future empirical research is warranted to elucidate the relationship between various jaggedness metrics and real-world risks.

Jaggedness metrics could also prove helpful for the governance of catastrophic risks. Currently, frontier labs' safety frameworks focus on detecting spikes in critical capability levels such as CBRN uplift and cyber capabilities Anthropic (2025); Google (2025); OpenAI (2025). Jaggedness metrics can better enable us to understand both the capability spikes and the rising general capability floor of a model's intelligence. By tracking both, we gain a more reliable leading indicator for future risks. Observing a rise in the overall ability floor suggests an increased probability of dangerous spikes emerging, as a higher floor provides a more fertile ground for specialized capabilities to be developed. Complex catastrophic actions — such as planning a bio-attack or navigating a cyber-exploit — rarely rely on a single narrow skill; they require the integration of reasoning, coding, and domain knowledge. A rising, smoother floor may signal that a model is developing the robust intelligence required to execute complex plans.

4.2. Implications for AI Policy and Governance

The public and policymakers have an interest in understanding the capability and safety profiles of frontier models. This need has been recognized by the increasing demand from policymakers for scientific reports on the state of AI science House (2023); on AI (2024). The International AI Safety Report, for example, alludes to the uneven capability set of AI models, and calls for better and broader evaluation methods to capture risk profiles Bengio et al. (2025). More recently, jaggedness has gained currency as a term in policy-oriented settings. Hammond (2025); Toner (2025). The jaggedness metrics we propose could be useful for providing policymakers a more detailed and contextualized view of the capabilities of leading models.

AI systems that can substitute for economically valuable labor at scale and thus may lead to substantial labor displacement are a key concern for policymakers Appel et al. (2025); Eloundou et al. (2023); Hendrycks et al. (2025b); Patwardhan et al. (2025). Replacement effects are likely to differ significantly depending on the jaggedness of AI systems. Jaggedness may be a leading indicator of whether AI systems will be labor-substituting or labor-complementing. We hypothesize that highly jagged AI systems may be more likely to complement human labor, with human skills having an economically valuable comparative advantage in the areas in which AI has valleys of ability. In contrast, smoother AI systems may have relatively few labor impacts until the cost/performance tradeoff of using AI versus human labor favors AI, at which point rapid, wholesale substitution may be more likely. It may be that certain combinations of peaks and valleys lend themselves more or less toward varying categories of labor replacement; we propose that specifically modeling jaggedness in terms of existing government labor frameworks may be useful not only for helping businesses make decisions around adoption strategies Narayanan and Kapoor (2025), but also for policymakers to understand the likely time trajectories of the societal change that may result from Replacement AI and the order in which varied labor categories may experience that change.

Some entities have adopted economic skill-based definitions of AGI; for example, OpenAI defines AGI in terms of, “highly autonomous systems that can outperform humans at most economically valuable work" "OpenAI". However, systems with particular jaggedness profiles may be able to substitute for labor in particular sectors before milestones such as “Expert AGI" or “ASI" are achieved, so long as the floor of capability across all skills is sufficiently high (perhaps at the “Competent" level in Levels of AGI taxonomy Morris et al. (2025)), with spikes corresponding to the particular strengths needed to perform particular categories of work. Skill-based or impact-based profiles of jaggedness may be more helpful than cognition-based profiles for understanding labor market trajectories.

Finally, characterizing jaggedness may also bolster societal resilience in the face of rapid and non-uniform technological advancement. In such circumstances, societal resilience may depend on having sufficient advance notice to prepare for advancements in capability areas that may otherwise outpace society’s ability to adapt. For instance, an unpredicted spike in automated software vulnerability detection could compromise power grids or banking systems before defensive measures mature. A clear temporal model of frontier jaggedness (specifically one that identifies which capability spikes are expanding fastest) provides a critical window of opportunity for governance. By monitoring these rapidly changing areas, society can move to a proactive posture, preemptively identifying and bolstering vulnerable sectors, such as critical national infrastructure, to ensure stability.

4.3. Implications for User Experience and AI Literacy

Today’s major commercial AI models generally rely on *prompting* as the primary mode of interaction; however, end-users vary in their prompting skills Gans (2026); Morris (2024); Schellaert et al. (2023); Zamfirescu-Pereira et al. (2023), leading them to have heterogeneous experiences with deployed models. User skill level, user interface design Morris (2025), and real-world user task distributions may result in a *perceived jaggedness* that differs from *benchmarked jaggedness* as measured for a base model by its developers. Measuring how perceived jaggedness differs from benchmarked jaggedness is likely to become an important usability metric for frontier model developers.

Currently, AI Literacy among the general public is highly variable Law (2025); we argue that the concept of jaggedness should be a component of AI Literacy campaigns, as awareness of jaggedness may impact end-users’ ability to choose whether and how much to trust AI systems. AI Literacy campaigns (e.g., Gunder (2025); Miao et al. (2024)) should emphasize both that AI models can be extremely jagged with respect to human baselines and that different models have different combinations and degrees of capability spikes and deficits.

End-user education around jaggedness might include the development of lesson plans targeted at particular audiences (K-12 students, university students, workforce sectors, policymakers, etc.) that give information about jaggedness tailored to the educational level and contextual needs of a particular group. It is particularly important to convey that, despite a preponderance of anthropomorphic language and metaphors surrounding consumer AI, model capabilities are not distributed the same way as human capabilities.

From a Human-Computer Interaction (HCI) perspective, jaggedness may create hazards around user calibration. Users may overestimate the floor of its capabilities based on high spikes in specific areas, leading to automation bias Parasuraman and Manzey (2010) and overtrust. Surfacing jaggedness profiles to users, perhaps via system cards Mitchell et al. (2019b), could enable them to better calibrate their trust of a system according to the task and compare models. Understanding the differences in jaggedness profiles for different models may help consumers make more informed choices of which models are appropriate for their goals, supporting improved *process alignment* Shen et al. (2025); Terry et al. (2024) between people and AI systems. For more specialized user groups such as policymakers or business leaders, it may be important for AI Literacy curricula to delve into more technical aspects of jaggedness such as how jaggedness metrics might relate to AI safety.

Additionally, model onboarding processes such as interactive tutorials could expose users to jaggedness concepts by demonstrating their strengths or weaknesses in tutorial tasks. Developers could also build in mechanisms for models to proactively communicate their own capability limitations. Such self-awareness may improve interaction by helping users calibrate their reliance on model outputs across different task types. For instance, if a user makes a request of a model that has a deficit in a relevant skill area, that model could refuse the task and suggest an alternate model with the relevant skill set, or it could execute the task but include caveats explaining why consulting with models with different jaggedness profiles might improve the result. In agentic scenarios, APIs could clearly convey jaggedness profiles to support agents' optimally selecting tools and models in support of a user's goals.

4.4. Implications for Measuring Progress towards AGI

The concept of AGI that meets or exceeds the breadth of human capabilities is often considered a North Star goal for AI research and development. Achieving particular levels of AGI Morris et al. (2025) is also considered a likely tipping point for societally significant impacts Gabriel et al. (2024); Hendrycks et al. (2025a). When considering capability profiles, the floor level of general capability determines the "level" of AGI progress Morris et al. (2025). However, the jaggedness profile and index offer additional perspective – "Human-level" AI does not necessarily mean "human-like" AI. A system can be human-level on average but possess a non-human jaggedness profile. This is why jaggedness may be a useful companion metric to other indicators of progress toward AGI. Taxonomies for measuring AGI progress may offer valuable capability profile frameworks for analyzing jaggedness and vice versa. Understanding and measuring the state and rate of change in jaggedness could support AGI-related forecasting and policymaking.

Considering both generality and jaggedness is critical for understanding many milestones and tipping points, including those related to safety and to societal transformation. For a given "floor" of general capability, it may be the case that highly jagged frontier AI systems with capability peaks in certain areas and sectors could induce major societal transformation long before AI systems achieve a general intelligence floor across all skills (e.g. AGI).

Concerns around Superintelligent AI that exceeds human capability include issues such as the likelihood of an intelligence explosion or fast takeoff Bostrom (2016) and unknown unknown emergent capabilities Bubeck et al. (2023); Wei et al. (2022a). Jaggedness provides a new perspective for

anticipating fast takeoff risks. Using skill- or impact-based profiles, spikes in coding and other skills related to performing advanced AI research and development may indicate recursive self-improvement strategies Kokotajlo et al. (2025). In cognition-based profiles, spikes in metacognitive skills such as learning how to acquire new abilities Morris et al. (2025) may have similar impacts.

AI systems with uneven capabilities may serve as effective complements to existing human capabilities in AI research and development. Prior work suggests that ML research is heavily bottlenecked by engineering implementation Owen (2024); models with spikes in coding and software engineering could reduce this friction even if it were to be the case that their capabilities for novel research ideation remain limited. The productivity gains from such complementarity depend on whether capability spikes align with genuine pipeline bottlenecks. As such, accurately characterizing jaggedness along skills that play a central role in modern AI development may be particularly important both for forecasting future rates of AI progress and predicting which research roles remain human-dominated. If AI systems prove exceptionally capable at compute-centric tasks (experiment scheduling, low-level optimization, resource allocation, etc.) the effective compute available for research could increase substantially without additional hardware. This form of jaggedness targeting resource constraints may prove especially impactful for development velocity, potentially alleviating what is currently a binding constraint on frontier development.

With a smooth frontier, one could extrapolate capabilities fairly easily. But with jaggedness, tracking and forecasting Murphy et al. (2025) the floor of general intelligence is not enough. Just tracking that could lead to us being surprised by jagged spikes of capabilities. Concepts of AGI and generality remain relevant despite jaggedness – different levels of AGI Morris et al. (2025) can be thought of as a floor of capability above which spikes rise. Models whose floor remains at “Emerging" levels are less likely to have significant societal, safety, or labor substitution impacts even with substantial strengths in a subset of areas (i.e., some minimal level of memory or planning skill is likely required to have impact, even if mathematical reasoning is quite high). It remains to be seen whether floors at higher levels (such as “Competent AGI") might result in societal transformations and risks not previously expected until higher levels in the Levels of AGI framework if such floors are combined with particular combinations of ability spikes.

5. Alternative Views

Though we posit that making jagged capability profiles of models more transparent to end-users will improve usability, we acknowledge that it is possible that information about jaggedness may confuse end-users. A combination of AI Literacy campaigns and appropriate interaction design may be necessary for end users to effectively take advantage of jaggedness information; investigating how to achieve this balance is an urgent area of inquiry at the intersection of AI and HCI. On a related note, our paper emphasizes characterizing the jaggedness of models, but it may be that characterizing the jaggedness of deployed systems is more relevant in practice, since various user interface factors may impact perceived jaggedness, which may have more practical import than benchmarked model jaggedness.

Finally, while we believe that jaggedness is likely to remain an enduring characteristic of AI models, it is possible that jagged profile capabilities are merely an artifact of a particular moment in time along the path toward general intelligence, and that by definition a sufficiently advanced AGI (or perhaps superintelligence) will be able to bootstrap its own learning so as to eliminate any relative skill deficits. However, even if superintelligent systems turn out to have smooth skill profiles, characterizing and monitoring jaggedness in the near term is still valuable for advancing us more safely and productively toward that outcome.

6. Conclusion and Call to Action

This paper presented a theoretical contribution regarding the significance of model jaggedness. This contribution included introducing several complementary methods to characterize and measure model jaggedness, such as capability frameworks, jaggedness profiles, the jaggedness index, and comparisons among these profiles and indices such as temporal differences in the jaggedness index across model versions and differences in benchmarked versus user-perceived jaggedness. There may be other valuable frameworks or metrics beyond those we have proposed in this article; indeed, we hope that this paper inspires other researchers to build upon our work by proposing additional ways to characterize jaggedness and empirically validating the utility of different characterization approaches.

Additionally, we contributed a nuanced discussion of how characterizing jaggedness can directly impact areas such as AI safety, AI policy and governance, usability, and measuring AI progress. If jaggedness metrics prove to be empirically rigorous and practically useful then they could be tracked, presented, and analyzed by different groups. The research community could benchmark the jaggedness of SOTA systems, as well as of key past systems, in order to understand patterns and rates of change in capability profiles and how various jaggedness metrics relate to particular training approaches and to real-world impacts and risks. Developers could include these metrics in information to users such as model cards. Policymakers could track changes in jaggedness profiles and indices over time to better forecast possible trajectories, and take jaggedness into consideration when developing governance and regulatory frameworks.

The development of artificial intelligence is not proceeding toward a smooth, human-like endpoint. The capability frontier is, and will likely remain, *jagged*. This does not, however, render the concept of AGI obsolete. By refining our understanding of AGI as the floor of a system's general capabilities, we posit that considering *both* this general capability floor *and* jaggedness metrics will prove a powerful approach for predicting and preparing for the future.

Acknowledgements

We thank our colleagues Murray Shanahan, Michael Terry, William Isaac, and Shane Legg for helpful conversations and feedback about this work. We note that this work represents the opinion of the authors, and is not an official policy statement from Google DeepMind.

References

AI Security Institute. Frontier ai trends report, December 2025. URL https://www.aisi.gov.uk/frontier-ai-trends-report.

Anthropic. Responsible scaling policy. Technical report, Anthropic, May 2025. URL https://www.anthropic.com/rsp. Version 2.2.

R. Appel, P. McCrory, A. Tamkin, M. McCain, T. Neylon, and M. Stern. Anthropic economic index report: Uneven geographic and enterprise ai adoption, September 2025. URL https://assets.anthropic.com/m/218c82b858610fac/original/Economic-Index.pdf.

ARC Prize. ARC Prize Leaderboard. https://arcprize.org/leaderboard, 2026. Accessed: 2026-01-27.

Art of Problem Solving. AMC Historical Results. https://artofproblemsolving.com/wiki/index.php/AMC_historical_results, 2025. Accessed: 2026-01-27.

Y. Bengio, G. Hinton, A. Yao, D. Song, P. Abbeel, T. Darrell, Y. N. Harari, Y.-Q. Zhang, L. Xue, S. Shalev-Shwartz, G. Hadfield, J. Clune, T. Maharaj, F. Hutter, A. G. Baydin, S. McIlraith, Q. Gao, A. Acharya, D. Krueger, A. Dragan, P. Torr, S. Russell, D. Kahneman, J. Brauner, and S. Mindermann. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, may 2024. ISSN 1095-9203. doi: 10.1126/science.adn0117. URL http://dx.doi.org/10.1126/science.adn0117.

Y. Bengio, S. Mindermann, D. Privitera, T. Besiroglu, R. Bommasani, S. Casper, Y. Choi, P. Fox, B. Garfinkel, D. Goldfarb, H. Heidari, A. Ho, S. Kapoor, L. Khalatbari, S. Longpre, S. Manning, V. Mavroudis, M. Mazeika, J. Michael, J. Newman, K. Y. Ng, C. T. Okolo, D. Raji, G. Sastry, E. Seger, T. Skeadas, T. South, E. Strubell, F. Tramèr, L. Velasco, N. Wheeler, D. Acemoglu, O. Adekanmbi, D. Dalrymple, T. G. Dietterich, P. Fung, P.-O. Gourinchas, F. Heintz, G. Hinton, N. Jennings, A. Krause, S. Leavy, P. Liang, T. Ludermir, V. Marda, H. Margetts, J. McDermid, J. Munga, A. Narayanan, A. Nelson, C. Neppel, A. Oh, G. Ramchurn, S. Russell, M. Schaake, B. Schölkopf, D. Song, A. Soto, L. Tiedrich, G. Varoquaux, E. W. Felten, A. Yao, Y.-Q. Zhang, O. Ajala, F. Albalawi, M. Alserkal, G. Avrin, C. Busch, A. C. P. de L. F. de Carvalho, B. Fox, A. S. Gill, A. H. Hatip, J. Heikkilä, C. Johnson, G. Jolly, Z. Katzir, S. M. Khan, H. Kitano, A. Krüger, K. M. Lee, D. V. Ligot, J. R. L. Portillo, D., O. Molchanovskyi, A. Monti, N. Mwamanzi, M. Nemer, N. Oliver, R. P. Rivera, B. Ravindran, H. Riza, C. Rugege, C. Seoighe, H. Sheikh, J. Sheehan, D. Wong, and Y. Zeng. International ai safety report. Technical Report DSIT 2025/001, Department for Science, Innovation and Technology (DSIT), 2025.

N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2016. ISBN 978-0198739838.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL https://arxiv.org/abs/2303.12712.

J. B. Carroll. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press, September 2009.

A. Chatterji1, T. Cunningham1, D. Deming, Z. Hitzig1, C. Ong1, C. Shan, and K. Wadman. How people use chatgpt, September 2025. URL https://cdn.openai.com/pdf/a253471f-8260-40c6-a2cc-aa93fe9f142e/economic-research-chatgpt-usage-paper.pdf.

F. Dell'Acqua, E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013), 2023.

T. Eloundou, S. Manning, P. Mishkin, and D. Rock. Gpts are gpts: An early look at the labor market impact potential of large language models, 2023. URL https://arxiv.org/abs/2303.10130.

European Commission. European skills, competences, qualifications and occupations (esco), October 2025. URL https://esco.ec.europa.eu/en/classification.

I. Gabriel, A. Manzini, G. Keeling, L. A. Hendricks, V. Rieser, H. Iqbal, N. Tomašev, I. Ktena, Z. Kenton, M. Rodriguez, S. El-Sayed, S. Brown, C. Akbulut, A. Trask, E. Hughes, A. S. Bergman, R. Shelby, N. Marchal, C. Griffin, J. Mateos-Garcia, L. Weidinger, W. Street, B. Lange, A. Ingerman, A. Lentz, R. Enger, A. Barakat, V. Krakovna, J. O. Siy, Z. Kurth-Nelson, A. McCroskery, V. Bolina, H. Law, M. Shanahan, L. Alberts, B. Balle, S. de Haas, Y. Ibitoye, A. Dafoe, B. Goldberg, S. Krier, A. Reese, S. Witherspoon, W. Hawkins, M. Rauh, D. Wallace, M. Franklin, J. A. Goldstein, J. Lehman, M. Klenk, S. Vallor, C. Biles, M. R. Morris, H. King, B. A. y Arcas, W. Isaac, and J. Manyika. The ethics of advanced ai assistants, 2024. URL https://arxiv.org/abs/2404.16244.

J. S. Gans. A model of artificial jagged intelligence. *NBER*, Jan. 2026. doi: 10.3386/w34712. URL https://www.nber.org/papers/w34712.

Gemini Team, Google. Gemini 2.5 pro model card. Technical report, Google DeepMind, 2025a. URL https://modelcards.withgoogle.com/assets/documents/gemini-2.5-pro.pdf. Model Card.

Gemini Team, Google. Gemini 3 pro model card. Technical report, Google DeepMind, 2025b. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf. Model Card.

Google. Frontier safety framework. Technical report, Google, Sept. 2025. Version 3.0.

K. Grace, J. F. Sandkühler, H. Stewart, B. Weinstein-Raun, S. Thomas, Z. Stein-Perlman, J. Salvatier, J. Brauner, and R. C. Korzekwa. Thousands of ai authors on the future of ai. *Journal of Artificial Intelligence Research*, 84, oct 2025. ISSN 1076-9757. doi: 10.1613/jair.1.19087. URL http://dx.doi.org/10.1613/jair.1.19087.

A. Gunder. Ai literacies in focus: From frameworks to action, 2025. URL https://wcet.wiche.edu/wp-content/uploads/sites/11/2025/09/2025-WCET-AI-Literacies-Lit-in-Focus-1.pdf.

S. Hammond. Shaping tomorrow: The future of artificial intelligence, September 2025. URL https://oversight.house.gov/wp-content/uploads/2025/09/Hammond-Written-Testimony.pdf.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

D. Hendrycks, E. Schmidt, and A. Wang. Superintelligence strategy: Expert version, 2025a. URL https://arxiv.org/abs/2503.05628.

D. Hendrycks, D. Song, C. Szegedy, H. Lee, Y. Gal, E. Brynjolfsson, S. Li, A. Zou, L. Levine, B. Han, J. Fu, Z. Liu, J. Shin, K. Lee, M. Mazeika, L. Phan, G. Ingebretsen, A. Khoja, C. Xie, O. Salaudeen, M. Hein, K. Zhao, A. Pan, D. Duvenaud, B. Li, S. Omohundro, G. Alfour, M. Tegmark, K. McGrew, G. Marcus, J. Tallinn, E. Schmidt, and Y. Bengio. A definition of agi, 2025b. URL https://arxiv.org/abs/2510.18212.

J. Hernandez-Orallo, B. Loe, L. Cheke, and et al. General intelligence disentangled via a generality metric for natural and artificial intelligence. *Scientific Reports*, 11, 2021. URL https://doi.org/10.1038/s41598-021-01997-7.

W. House. Executive order 14110: Safe, secure, and trustworthy development and use of artificial intelligence, October 2023. URL https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

M. Knoop. Arc prize 2025 results & analysis, dec 2025. URL https://arcprize.org/blog/arc-prize-2025-results-analysis. Accessed: 2026-01-22.

D. Kokotajlo, S. Alexander, T. Larsen, E. Lifland, and R. Dean. Ai 2027, April 2025. URL https://ai-2027.com/.

N. Lambert. Building on evaluation quicksand. Interconnects, October 2024. URL https://www.interconnects.ai/p/building-on-evaluation-quicksand. Accessed: 2026-01-09.

H. Law. What does the public really think about ai?, November 2025. URL https://www.aipolicyperspectives.com/p/what-does-the-public-really-think.

S. LeGris, W. K. Vong, B. M. Lake, and T. M. Gureckis. H-ARC: A robust estimate of human performance on the abstraction and reasoning corpus benchmark. *arXiv preprint arXiv:2409.01374*, 2024. URL https://arxiv.org/abs/2409.01374.

P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda. Holistic evaluation of language models, 2023. URL https://arxiv.org/abs/2211.09110.

N. Maslej, L. Fattorini, R. Perrault, Y. Gil, V. Parli, N. Kariuki, E. Capstick, A. Reuel, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, J. C. Niebles, Y. Shoham, R. Wald, T. Walsh, A. Hamrah, L. Santarlasci, J. B. Lotufo, A. Rome, A. Shi, and S. Oak. Artificial intelligence index report 2025, 2025. URL https://arxiv.org/abs/2504.07139.

R. McElreath. *The Golem of Prague*, chapter 1, pages 1–18. Chapman and Hall/CRC, 2nd edition, 2020. doi: 10.1201/9780429029608. URL https://www.taylorfrancis.com/books/mono/10.1201/9780429029608/statistical-rethinking-richard-mcelreath.

F. Miao, K. Shiohira, and N. Lao. Ai competency framework for students, August 2024. URL https://www.unesco.org/en/articles/ai-competency-framework-students.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229. ACM, Jan. 2019a. doi: 10.1145/3287560.3287596. URL http://dx.doi.org/10.1145/3287560.3287596.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019b.

E. Mollick. The shape of ai: Jaggedness, bottlenecks and salients, 2025. URL https://www.oneusefulthing.org/p/the-shape-of-ai-jaggedness-bottlenecks. Blog post.

H. Moravec. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, 1988.

M. R. Morris. Prompting considered harmful. *Commun. ACM*, 67(12):28–30, Nov. 2024. ISSN 0001-0782. doi: 10.1145/3673861. URL https://doi.org/10.1145/3673861.

M. R. Morris. Hci for agi. *Interactions*, 32(2):26–32, Feb. 2025. ISSN 1072-5520. doi: 10.1145/3708815. URL https://doi.org/10.1145/3708815.

M. R. Morris, J. Sohl-Dickstein, N. Fiedel, T. Warkentin, A. Dafoe, A. Faust, C. Farabet, and S. Legg. Levels of agi for operationalizing progress on the path to agi, 2025. URL https://arxiv.org/abs/2311.02462.

S. Mukhopadhyay, R. Baral, N. Mahajan, S. Harish, A. RRV, M. Parmar, M. Nakamura, and C. Baral. Phantom recall: When familiar puzzles fool smart models, 2025.

C. Murphy, J. Rosenberg, J. Canedy, Z. Jacobs, N. Flechner, R. Britt, A. Pan, C. Rogers-Smith, D. Mayland, C. Buffington, S. Kučinskas, A. Coston, H. Kerner, E. Pierson, R. Rabbany, M. Salganik, R. Seamans, Y. Su, F. Tramèr, T. Hashimoto, A. Narayanan, P. E. Tetlock, and E. Karger. The longitudinal expert ai panel: Understanding expert views on ai capabilities, adoption, and impact, November 2025. URL https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/6911df98386b4e258c4cd4e5/1762779032257/the-longitudinal-expert-ai-panel.pdf.

A. Narayanan and S. Kapoor. Ai as normal technology, April 2025. URL https://knightcolumbia.org/content/ai-as-normal-technology.

U. H. L. A. B. on AI. Governing ai for humanity. *United Nations*, 2024.

"OpenAI". "openai charter". URL "https://openai.com/charter/".

OpenAI. Preparedness framework. Technical report, OpenAI, Apr. 2025. Version 2. Last updated: 15th April, 2025.

D. Owen. Interviewing ai researchers on automation of ai r&d, 2024. URL https://epoch.ai/blog/interviewing-ai-researchers-on-automation-of-ai-rnd. Accessed: 2025-12-15.

R. Parasuraman and D. H. Manzey. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410, 2010. doi: 10.1177/0018720810376055. URL https://doi.org/10.1177/0018720810376055. PMID: 21077562.

D. Patel. Rl is even more information inefficient than you thought, 2025. URL https://www.dwarkesh.com/p/bits-per-sample. Blog post.

T. Patwardhan, R. Dias, E. Proehl, G. Kim, M. Wang, O. Watkins, S. P. Fishman, M. Aljubeh, P. Thacker, L. Fauconnet, N. S. Kim, P. Chao, S. Miserendino, G. Chabot, D. Li, M. Sharman, A. Barr, A. Glaese, and J. Tworek. Gdpval: Evaluating ai model performance on real-world economically valuable tasks, 2025. URL https://arxiv.org/abs/2510.04374.

R. Rahman and A. A. Mishra. A fragile number sense: Probing the elemental limits of numerical reasoning in llms, 2025. URL https://arxiv.org/abs/2509.06332.

R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ITw9edRDlD.

W. Schellaert, F. Martinez-Plumed, K. Vold, J. Burden, P. A. M. Casares, B. S. Loe, R. Reichart, S. O. hEigeartaigh, A. Korhonen, and J. Hernandez-Orallo. Your prompt is my command: On assessing the human-centred generality of multimodal models, June 2023. URL https://jair.org/index.php/jair/article/view/14157.

W. J. Schneider and K. McGrew. *The Cattell-Horn-Carroll model of intelligence*. New ork: Guilford, 2012.

H. Shen, T. Knearem, R. Ghosh, K. Alkiek, K. Krishna, Y. Liu, Z. Ma, S. Petridis, Y.-H. Peng, L. Qiwei, S. Rakshit, C. Si, Y. Xie, J. P. Bigham, F. Bentley, J. Chai, Z. Lipton, Q. Mei, R. Mihalcea, M. Terry, D. Yang, M. R. Morris, P. Resnick, and D. Jurgens. Position: Towards bidirectional human-ai alignment, 2025. URL https://arxiv.org/abs/2406.09264.

A. Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. URL https://arxiv.org/abs/2206.04615.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

M. Terry, C. Kulkarni, M. Wattenberg, L. Dixon, and M. R. Morris. Interactive ai alignment: Specification, process, and evaluation alignment, 2024. URL https://arxiv.org/abs/2311.00710.

H. Toner. Helen toner on ai's jagged frontier | the curve 2025, November 2025. URL https://www.youtube.com/watch?v=avxO7ZEJH4w.

U.S. Dept. of Labor Statistics. O*net, 2025. URL https://www.onetonline.org/.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent abilities of large language models, 2022a. URL https://arxiv.org/abs/2206.07682.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022b.

J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, and Q. Yang. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581388. URL https://doi.org/10.1145/3544548.3581388.

W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng. Wildchat: 1m chatgpt interaction logs in the wild, 2024. URL https://arxiv.org/abs/2405.01470.

L. Zhou, L. Pacchiardi, F. Martinez-Plumed, K. M. Collins, Y. Moros-Daval, S. Zhang, Q. Zhao, Y. Huang, L. Sun, J. E. Prunty, Z. Li, P. Sanchez-Garcia, K. J. Chen, P. A. M. Casares, J. Zu, J. Burden, B. Mehrbakhsh, D. Stillwell, M. Cebrian, J. Wang, P. Henderson, S. T. Wu, P. C. Kyllonen, L. Cheke, X. Xie, and J. Hernandez-Orallo. General scales unlock ai evaluation with explanatory and predictive power, 2025. URL https://arxiv.org/abs/2503.06378.