ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240459

Caveats and Solutions for Characterising General-Purpose AI

Jose Hernandez-Orallo [a,b,c,d,*]

[a]VRAIN, Universitat Politecnica de Valencia
[b]VALGRAI
[c]Leverhulme Centre for the Future of Intelligence
[d]Centre for the Study of Existential Risk
ORCID (Jose Hernandez-Orallo): https://orcid.org/0000-0001-9746-7632

Abstract. The concept of General-Purpose AI (GPAI) has recently been permeating research papers, policy reports and legal regulations, as a way of referring to current and future models with high levels of capability and generality. Yet precisely characterising GPAI models remains elusive. Current definitions often describe GPAI models as those that 'competently perform a wide range of distinct tasks'. To properly characterise GPAI we need well-grounded definitions of capability and generality. In this paper, I will briefly introduce –or revisit– the concept of capability, going well beyond aggregate performance on benchmarks, and discuss practical procedures to evaluate the capability profile of AI systems, and derive generality metrics from them.

1 Introduction

The term General-Purpose AI (GPAI) has become increasingly popular as soon as large language models started to offer a wide-range of functionalities, becoming the first systems in the history of AI to be really general-purpose, doing many things for which they were not programmed. As technology will evolve and may even replace large language models or other foundation models as the dominant paradigm, the broader term GPAI has been used by researchers and regulators[1], but no widely accepted definition has been given. There is no consensus in what elements should be necessary for GPAI, such as competency for a wide range of tasks, preferably out of the box, and what elements should be excluded, such as metacognition (e.g., 'acknowledgment of its own limitations') [41]. Also, there is confusion between GPAI and human-level machine intelligence, AGI or any other interpretation of very advanced artificial intelligence.

In this paper, we explore the range of interpretations and take a pragmatic view of GPAI as combining the notions of capability and generality. While capability and generality have many different interpretations, there is some common ground to build on. Informally, capability will be defined as a property of a system that allows it to perform well on tasks that demand that capability. For instance, a model has addition capability level 7 if it can add up numbers correctly up to 7 digits. Generality will be defined as regularity in the capability levels for a range of domains. For instance, a model is general if it has similar capability levels in arithmetic, history, commonsense reasoning, social situations, etc. Our goal in this paper is to give more precise definitions of these concepts, and set some operational criteria so that we can locate current and future models in the space of generality and capability, as illustrated by Figure 1.

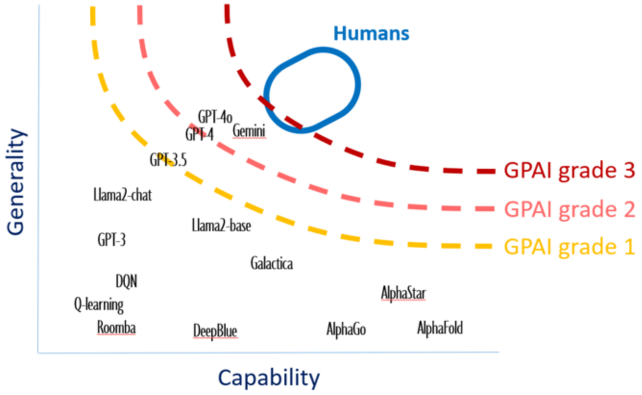

Figure 1. As per 2024, our approximate intuition about models (and systems) that are General-Purpose AI (GPAI) can be represented by having some degree of generality and some degree of total capability. For instance, many early AI systems were neither capable nor general, such as Q-learning (a RL algorithm) or Roomba (a robotic cleaner). In the past few years we have seen very capable systems for narrow tasks (e.g., DeepBlue for chess, and AlphaGo for Go, with higher capability than the best players in the world), and also some general systems but with limited capability (such as GPT3). More recently, we are starting to see models of significant degrees of capability and generality at the same time, such as GPT4, performing a wide range of tasks reasonably well. In this paper we propose a way to map these systems into different grades of GPAI.

The rest of the paper is organised as follows. The following section 2 discusses why general-purpose AI is so relevant, followed by different interpretations of capability and generality, historically and from different disciplines in section 3. Section 4 identifies several misconceptions and caveats. Finally, section 5 is ready to introduce the *operational mechanism to characterise GPAI*. We close the paper with a short discussion about they way the criteria outlined here can be revised and kept up-to-date.

[*] Corresponding Author. Email: jorallo@upv.es

[1] For isntance, the EU AI Act https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/745708/EPRS_ATA(2023)745708_EN.pdf.

2 The impact of generality

The impact of AI, both in transformative power and risks, is highly dependent on how general a system is. GPAI systems have a faster penetration rate (ChapGPT reached millions of users within days), higher adaptability and interoperability with many other tools (plugins, agents and RAG with language models), stronger effects on human cognition (dependency) and everyday activities (education and jobs), and pose much more difficult challenges for safety and control (e.g., deception, manipulation, etc.). For instance, while AlphaFold may have important applications, its effect on society and its risks are more limited than systems such as chatGPT or Gemini.

This relevance of GPAI is supported by the well-known effect of generality in natural intelligence, starting with human intelligence. The association of generality with intelligence appears as soon as classical and medieval philosophy dealt with human faculties, and especially reasoning. Ramon Llull sought the ideal of an 'ars generalis', the power that could 'solve all solvable tasks' [31]. Yet our scientific understanding of intelligence is more recent, developed within the fields of (human) psychology and psychometrics, comparative (animal) cognition and other social sciences [16]. All these fields emphasise the importance of generality.

For instance, Spearman developed the notion of the *g* factor, a latent factor that explained the variance of performance in a human population for a wide range of cognitive tasks. In fact, generality is incorporated in the very definition of (general) 'intelligence' as "the capacity of getting along well in all sorts of situations" [30]. General intelligence was assimilated with both *breadth* and *depth* of capabilities, and good scores in IQ tests were correlated with 'success in life' and even life expectancy. In animal cognition, generality is usually seen as opposed to specialisation, and in the context of brain 'modularity' (some functions can only be developed in some specialised parts of the brain) vs 'plasticity' (many parts of the brain can do different functions) [35]. A similar sociotechnical perspective is framed under the concept of 'general-purpose technology', and in economics, around specialisation and division of labour. Overall, the relation between generality and intelligence in all sciences dealing with it is as important as complex, and so is the concept of 'general intelligence' (natural or artificial) [16, 19].

The history of artificial intelligence absorbs all these perspectives. The early ambition of AI was precisely to build a general-problem solver [28]. However, early failures shifted the focus to what was known as 'narrow' artificial intelligence, highly specialised systems, only solving one or few tasks. Under this perspective, generality was expected to be achieved as a bunch of tricks, to the extreme of systems becoming a 'big switch': determining which module to use for each particular problem. Such was the challenge in the early days of AI that McCarthy's Turing award lecture in 1971 had the name 'Generality in AI' [27]. For many decades the dream of having AI that could "adapt its behavior to meet goals in a range of environments" [9] looked unattainable. Concepts such as artificial general intelligence (AGI), human-level machine intelligence (HLMI) and strong AI were considered synonyms, even if some of them were, and still are, poorly defined. What was clear is that given the impact that natural general intelligence has had on our planet, represented by the homo sapiens, general AI could also have a transformative effect on every aspect of human life, from jobs to existential risks.

It is in this context when large language models (LLMs) took the world by surprise, and showed, for the first time in AI, effective generality. No-one can deny that some LLMs, better or worse, yet off-the-shelf, can do many tasks.

3 Perspectives of capability and generality

Most AI systems, influenced by the traditional narrow approach, are evaluated with benchmarks that measure performance on one or more tasks, such as object classification, task scheduling or language translation. Benchmarking is problematic because the results that are obtained for these benchmarks lead to overestimates (yet sometimes underestimates) of what models can really do [3]. For instance, looking at Figure 5, we may get the idea that most models today are superhuman, but then we realise they do not perform close to human in the real world or for variations of the tasks. Also, when models really become superhuman in some areas, it is hard to extrapolate beyond that, as most benchmarks saturate on a scale topped by 100% accuracy or human-level performance. Finally, in situations where tasks change, or we want to extrapolate for new tasks, the aggregate performance on a benchmark is mostly useless to extrapolate for new tasks, even in the same domain, since they can contain instances of different difficulty. While all these problems in the current evaluation paradigm are recognised, the main cause is an elephant in the room: performance measures some metric of success on a distribution of tasks, and the measurement will not apply when the distribution changes. This paradigm leads to poor test reliability and validity, and ultimately to the bad reputation of AI evaluation.

Instead of the dominant paradigm focused on 'performance', we need a different one based on the notion of 'capability', which characterises the system independently of the distribution, and may have predictability across different tasks. Because *capability is what we really need*, many AI labs, but also some academics, use the term 'capability' when they are still measuring aggregate performance [36]. This confusion between performance and capability is easy to make because performance can be a good proxy for capability in narrow benchmarks and applications. However, performance is usually a bad proxy for capability for general-purpose systems, and it can be especially problematic for estimating generality, as performance is usually incommensurate across different tasks and benchmarks.

To make the distinction clear, performance is a function of both the system and a distribution of tasks, whereas capability is a property of a system that informs about whether that system is able to succeed on a task that *demands* that capability. For instance, an AI system can be 85% successful in separating recyclable materials out of a tray, but if the tray distribution changes, that probability will likely change. On the other hand, if the capability of the system is estimated to be high for trays up to '4 materials', but unreliable when more than four materials are presented, then we can anticipate how the system will behave for a new batch of trays if we know how many elements each tray contains. The capability is still '4 materials', and does not change if we change the distribution. Given this intuition, we can give the following definition of capability:

Definition 1. *A capability is a property of a system such that once estimated at a certain level, success is likely for tasks that have demands below that level (i.e., easy tasks) and unlikely for tasks that have demands above that level (i.e., hard tasks). See Figure 2.*

Different formulations of capability have been introduced in many disciplines, especially in psychometrics and cognitive psychology. For instance, the previous definition is strongly inspired by the definition of ability in item response theory in psychometrics [8]. The notion of demand (or difficulty) of a task instance (or item) is crucial to *predict* whether a system is going to succeed for a particular problem. For instance, if a student has a capability for addition up to 10 digits (their capability), as in Figure 2, this is more predictive than

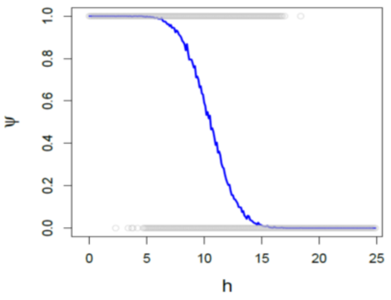

Figure 2. Figurative subject characteristic curve: expected success probability (y-axis) of a given subject (e.g., a human or an AI model) depending on the difficulty of items (x-axis). We see that for items of difficulty lower than 10 the probability of success is high. With items with difficulty greater than 10 the probability is low. Because of this we could say that the capability of the subject is approximately 10, around the steepest part of the curve where the probability of success plummets.

if we say that the student has achieved 85% success in a particular combination of sums (their performance). This high percentage may simply be caused by only having easy additions in the dataset.

Accurately estimating capability is not always an easy problem for non-human animals, humans and AI systems [16, 4]. But distinguishing capability from performance is crucial to address our following question: a concept of generality.

First, it is important that all stakeholders are aware of the different interpretations of generality, including technical developers and evaluators, but also policy-makers and regulators. It is also important that all understand whether the definitions and assessments are capturing the different intuitive notions about generality. Let us enumerate some of the most common perceptions of generality, again from different disciplines:

1. **All-tasks**: Success in all situations. This is a very simple definition, but without any qualification on 'all', it leads to all sort problems such as incomputability, no-free-lunch results, etc., even if a universal distribution [24] is chosen. *Caveat*: no system can succeed in all situations.
2. **Many-tasks**: Success in a wide range of situations. This is a qualification of the above that makes the definition subjective on the definition of that range, or distribution, of situations. Also, are we expecting infallible success in all situations in that range or an aggregate of that success weighted by the distribution? *Caveat*: who chooses the task *usage* distribution that matters, and the aggregation or threshold function?
3. **Human-tasks**: Able to do everything a human can do. This is a refinement from the above by defining the distribution of situations in an anthropocentric way, but sets a very high threshold; we already have general systems that are far from doing everything a human can do. *Caveat*: who identifies what humans can do? Is this of an average human or the best human?
4. **Capabilities**: Having an elemental range of capabilities. This is more abstract than 'a wide range of situations'. As a capability can serve for many tasks, and many tasks involve many capabilities, this is another way to define the breadth that is expected in generality. Also, some capabilities, such as the ability of learning, are potential rather than actual. *Caveat*: how can we determine the number and structure of capabilities that define a general system? Are these capabilities actual or potential?
5. **Out-of-distribution**: Success out of distribution the system has been trained or habituated to. In principle, if the system does not require any adaptation during deployment then this is associated with the concept of 'generalisation power' in learning. If some adaptation is needed then we are in the following item, 'transfer'. *Caveat*: how can we define out-of-distribution and what is the level of decreased performance that is acceptable?
6. **Transfer**: Flexibility to adapt to new tasks. This introduces a dynamic aspect or potentiality; the system may not able to solve everything, or even a wide range of tasks, at the actual moment, but can adapt easily to new tasks in an *autonomous* way. *Caveat*: this requires notions of task similarity (very different tasks are expected to be harder), effort (time taken or number of examples needed) and external scaffolding (help from humans through hints or teaching).
7. **Compositionality**: Integration of different skills for complex tasks. A system can solve tasks involving specific skills (e.g., arithmetic and medicine) but may be incapable of combining these two skills for tasks that require both. There has been disagreement on whether language models, for instance, are simply stochastic parrots [2] or can really compose skills [46] to some extent. *Caveat*: how to identify skills and modes of composition?
8. **Multimodality**: Integration of different input and output modalities, such as text, audio, video, etc. This is a less common understanding of generality in its own, but it is true that the more perceptual capabilities and kinds of interaction the model supports, the more general we consider it to be. When actuators are complex this can also be applied to psychomotor skills as well. *Caveat*: We have examples of reduced modality in humans (e.g., deafness) and machines (text-only GPT) that are still considered very general.

Some definitions of generality or general-purpose AI focus on one or more of these perspectives. For instance, [12] combine item 2 (many-tasks) and an integration of items 5 and 6 into the following definition of GPAI: "an AI system that can accomplish or be adapted to accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained". In general, the perspectives overlap significantly, and many definitions tend to make two major clusters: items 2 (many-tasks) and 4 (capabilities) representing the dimension of breadth, and items 5 (out-of-distribution) and 6 (transfer) representing the dimension of adaptability or re-purposing. Figure 3 illustrates the breadth conception.

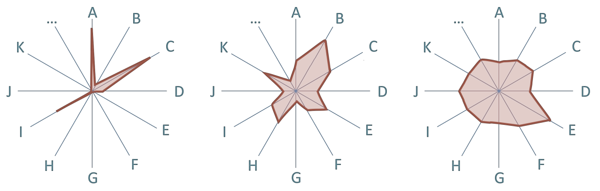

Figure 3. The *profiles* of three AI systems performing differently for a range of domains A, B, ..., . The system on the left has high capability levels for two domains (A and C) and intermediate capability level for I. The system in the middle does not excel in any domain but has some capability in all domains. The system on the right has good capability levels for all domains. Depending of the identification of domains (the usage distribution), the scales for each of them, and the metrics of capability and generality, we can determine which of these systems are general-purpose AI systems.

Many other definitions of general AI conflate generality with high capability. Of course, we would like to have both capability and generality, but they are not the same thing [19]. We know of humans that can master a small set of things compared to others that can do many different things but not very well. While the former seemed to correspond to narrow AI systems (such as AlphaGo), the latter had not been possible in AI until large language models, such as GPT-3, showed that they could do many things (even if not to an advanced degree yet). Figure 1 showed a schematic representation of models and systems in this space.

However, when these definitions and conceptions are turned into evaluation instruments, what we usually find is the *many-benchmarks* approach, which is half-way between the many-tasks and human-tasks perspectives. To approach the diversity angle of generality or to assimilate a group of benchmarks with a 'capability', benchmarks are grouped by domains, such as 'language understanding' [14], 'mathematics' [15], 'common-sense reasoning', etc. Identifying these domains, as used in Figure 3, will be fundamental for determining a system *profile*, as the set of domains has to be a good clustering of tasks, so that generality really captures diversity. On many other occasions, benchmarks are just created by agglutination, such as BIG-Bench [36] or with the intention, yet no proof, of being diverse, such as HELM [25]. Averages are calculated on the whole benchmark or on selected ablations, such as HELM-light, trying to be more representative of the tasks that matter. Even if the selection is performed with care, the aggregation of results can give misleadingly good aggregate results even if the system is performing very well on only some domains, and very poorly on the rest, as in Figure 3 (left). In other words, *performance aggregation hinders the analysis of generality*. But, more importantly, an apparently irregular set of results may be caused by benchmarks with *different difficulty distributions*, even if they are of the same domain. For instance, Figure 4 shows that even if MMLU includes questions whose domains overlap with those of GPQA and MATH, when we look at GPQA, it may give the impression that these models are weak on the "domain of GPQA". This is simply wrong, since models score worse on GPQA just because this benchmark contains a distribution of more difficult questions. This example shows that the notion of generality needs a calibration of the results for different domains so that they are commensurate, i.e., the magnitude estimated for one domain is on the same scale than the other domains, avoiding aggregating apples and pears.

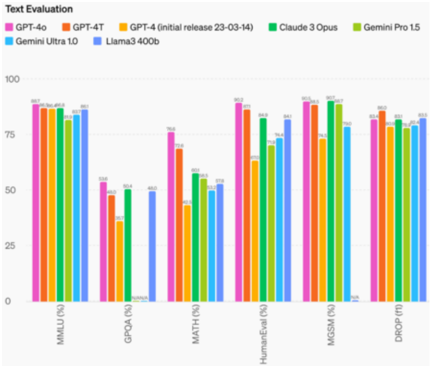

Figure 4. Results for several foundation models on common benchmarks as reported by OpenAI (May 2024). We see that two benchmarks that are about a similar domain (MMLU and GPQA) show different levels of performance, because one (GPQA) has instances of higher difficulty. Does this look more like the middle or the right plot in Figure 3? (Image from https://openai.com/index/hello-gpt-4o/).

The question of the magnitude of the scale is also very important for another reason: *benchmarks saturate*, approaching the maximum of the performance metric or the human standard that is used for ground truth [18]. For instance, Figure 5 (red curve) shows how MMLU is saturating, as well as some other benchmarks. This suggests that more difficult instances in the test set are needed to track the progress of these systems. Even if the training set has limited difficulty, the test results could well go beyond the difficulty levels in the training set; it has been shown that big models extrapolate well from easy to difficult examples [13, 38]. For all this to make sense, we need to calibrate old and new results. However, these extended or harder examples typically deviate from the distribution, so the results are affected by more factors than just an increase of difficulty. As a result, there is *no calibration of results* when benchmarks are evaluated with performance rather than capability.

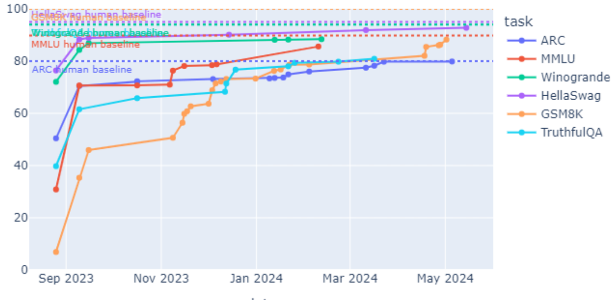

Figure 5. Saturation behaviours for some other benchmarks, also compared with human average. Plot from: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

.

Seen all the problems of the aggregation of benchmark performance, we need an alternative paradigm for defining general-purpose AI, working on some of the dimensions seen in this section.

4 Caveats in measuring generality

Any metric of generality has to make some assumptions about what we call the *task usage distribution* $\mathcal{T}$, i.e., a probability distribution over tasks depending on what we consider relevant or meaningful in this world. We can use $\mathbb{P}(t|u)$ to exemplify that this distribution of tasks t is different for each user u. This may also be different in different times and contexts. A fine-grained view of this distribution does not preclude us to make coarser aggregations, such as considering the population of users, or even all humans, to get a common task usage distribution $\mathbb{P}_{\mathcal{T}}(t) = \mathbb{P}(t|u)\mathbb{P}(u)$, assuming independence between users.

Given a 'subject' system s to be evaluated and a user u, we can randomly sample from this distribution $\mathcal{T}$ and calculate for each t the utility or validity $v(s, t, u)$. With this we can estimate the expected value $\mathbb{E}_{t \sim \mathcal{T}}[v(s, t, u)]$, an aggregate performance metric for task distribution $\mathcal{T}$. However, imagine we are only interested in the domains of *arithmetic* and *diplomacy*, and assume that arithmetic has 50% of the mass of the distribution in $\mathcal{T}$ and diplomacy has the remaining 50%. A perfect system on arithmetic, failing on all diplomacy tasks would score the same as another system succeeding on half of each domain. We tend to think the second system is more general. This illustrates that what matters is the distribution of success across *domains*, not aggregation over all individual tasks. In order to accommodate for this, it is usual to group tasks into more abstract domains $D_1, D_2, ..., D_k$, and apply this distributional analysis on the domains and not on the possibly infinite number of tasks. For clarity, since the term 'task' can represent a specific task example (e.g., "What's the sum or 593 + 256?") or a broad set of examples (e.g., the addition task), from now on we will use the term 'instance' for the former and 'domain' for the latter.

If we now have two levels, instances and domains, what do we do with the results of the instances of each domain? In the previous example, what does it mean to succeed in 50% of examples in the arithmetic domain? The number of arithmetic operations we might conceive is infinite, so we again need a distribution or sample of common arithmetic operations. In other words, we need $\mathbb{P}(i|D)$, the probability of instance i happening in domain D. This is what a usual benchmark is, a collection of likely or common examples. For instance, a simple addition involving numbers of 10e500 digits cannot be solved in this universe, and we do not expect this one to appear in the sample. However, while this commonplace sampling is the traditional approach in AI benchmarks, this has the problem that this $\mathbb{P}(i|D)$ is usually hard to estimate and it may change drastically from one benchmark to another of the same domain, intentionally or unintentionally. For instance, two image classification benchmarks can contain similar images, but one can contain more *difficult* ones. The same system can score 20% on one benchmark and 90% on the other. Which of these percentages should we use as the *capability* for that domain? The answer is clear: none of these, or any aggregate of performance, should be used as an estimate for capability[2].

As introduced in definition 1, performance is the aggregate observed quality of a pair system and distribution, while capability is a property of the system, independent from the distribution. Only with this interpretation can we have a capability that does not change when we change the distribution in the benchmark (but should naturally lead to changes in performance out of the distribution). Consider again the very simple example of addition. If we define the difficulty of addition to be the average number of digits of the two addends, and we have the data represented in Figure 2, then we could say that the capability of a system is estimated at 10, because this is the inflection point where errors become more systematic.

Under this view, for each domain D, we need to introduce a difficulty function $h(i)$ for each task instance in the domain ($i \in D$). These functions can be derived from intrinsic characteristics of the task or can be derived empirically from a human population, such as done in item response theory (IRT) [8, 43]. Given this function we can then sample instances in a sufficiently wide range of difficulties, so that we can estimate a *subject characteristic curve*, showing the evolution of the response of the system on the y-axis for increasing difficulty on the x-axis. See Figures 2 and 6. From this (ideally decreasing) curve, there are different ways to estimate the capability of the system, such as the point of maximum slope, the point where response goes below a threshold, or, if there is proper scaling of the difficulties on the x-axis, the area under the curve. The latter interpretation is actually assuming a uniform distribution of difficulties, so $\mathbb{P}(i|D) = \mathbb{P}(i|h, D)\mathbb{P}(h|D)$ where $\mathbb{P}(h|D)$ is uniform and $\mathbb{P}(i|h, D)$ should not affect the result in expectation, as a good difficulty function should be defined such that $\forall i, j \in D, h(i) = h(j) : \mathbb{E}[v(s, i)] = \mathbb{E}[v(s, j)]$, i.e., there is the same expectation of success for instances of the same difficulty.

Reached this point, the *estimation of generality* would rely on:

1. **Domain partition**: A partition of task domains $D_1, D_2, ..., D_k$, with a domain distribution $\mathcal{D}$. If the partition is balanced (all domains are equally important in the task usage distribution), then $\mathcal{D}$ can be assumed uniform: $\mathbb{P}_{\mathcal{D}}(D_i) = 1/k$.
2. **Capabilities for each domain**: A metric of capability for each domain, ideally based on subject characteristic curves using difficulty. For instance, the capability of system s for domain D could be derived by sampling instances from a range of difficulties for that domain to estimate $\Psi(s, D) = \sum_{i \in D} \mathbb{P}(i|D) \cdot v(s, i) = \sum_{i \in D, h \in 1..k} \mathbb{P}(i|h, D) \cdot \mathbb{P}(h|D) \cdot v(s, i) = 1/k \sum_{i \in D, h \in 1..k} \mathbb{P}(i|h, D) \cdot v(s, i)$ by assuming $\mathbb{P}(h|D)$ uniform (capability as area under the curve).
3. **Regularity across domains**: A metric of generality based on the regularity of capabilities. If the difficulties for different domains are commensurate (capabilities in the same scale) and $\mathcal{D}$ is uniform, this could be based on the standard deviation of $\Psi(s, D)$ for all domains (for a justification of this choice, see proposition 1 in [19]), or a unit-less coefficient of variation for higher profile monotonicity. Alternatively, we could use the number of domains above a capability threshold, such as 25% of humans.

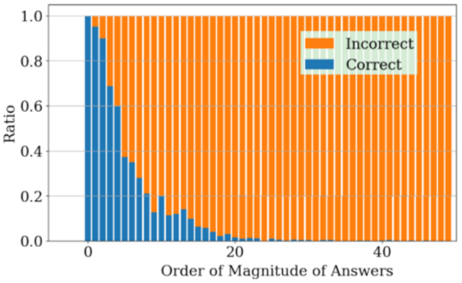

Figure 6. Actual subject characteristic curve (blue). Results of GPT-NeoX for 100,000 instances i of the addition of two decimal numbers, all of them with the same prompt "What is a + b?", where a and b refer to numbers up to 100 digits (plot only showing until 50 digits). Results are shown by the number of digits of the sum (up to 100 digits), which works as a proxy for difficulty $h(i)$, and the scale for demand and capability. The blue 'curve' shows the ratio of correct answers, while the rest (in orange) is errors. If we were aiming for a probability of success of 0.5, we would say that the capability is 5 digits. If we were aiming to secure a probability of success of 0.9 we would say that the capability is 3 digits. The area of the blue curve is approximately 6.4, which given the sigmoid shape of the curve we could also take as a proxy for capability. Note that the accuracy on the whole dataset is so much distribution-dependent that for this distribution in particular accuracy is only 0.0065 (a totally uninformative number). Plot from [10].

The items above are not straightforward and require scientific expertise on evaluation and significant consensus. For instance, the domain distribution might contain a few domains, or might contain hundreds. It must be broad, as the list of capabilities in psychometrics, the set of disciplines used in education or the collection of work tasks identified from human occupations. For instance, these could be elicited from sources such as the second level of the Cattell-Horn-Carroll hierarchy [34], the list of abilities, knowledge areas or skills in the occupational database O*Net (https://www.onetonline.org/), the frequency of computer tasks (https://os-world.github.io/), etc. The choice of domains is as important as tricky, but at least these examples from different disciplines are based on some criteria of balance and representativeness, and supported by some intuition of what task usage distribution represents. In contrast, many other compendiums of tasks used in AI are created by the availability of benchmarks, such as BIG-Bench [36], rather than a clear selection criteria. Also, the identification of a good validity metric, $v(s, i)$, even if generalised for all users, requires a thorough analysis of all the elements discussed in the previous subsection. Finally, the difficulties or demands for each domain require expert effort and careful analysis of key tasks in the domain, using proxies of observable features (e.g., number of digits) or a rubric that is used to annotate each example. Deriving scales that

[2] A large majority of references in AI (with notable exceptions, e.g., [32]) use the term 'capability' in a very vague way or as a synonym of performance. This is simply inaccurate and at odds with the traditional use of capability in psychology as a construct, a latent trait rather than an observable variable.

are commensurate (e.g., comparing level 3 in arithmetic with level 2 in diplomacy) must currently rely on human standardisation, until more theoretical or computational underpinnings exist [17].

Even if these questions are resolved, there are some extra caveats. First, we have the problem of *evaluation contamination*, where items in the test dataset, or very similar questions, have been used in the training dataset for these systems. This can happen in more subtle ways than one might expect, as simply querying a large language model through an API could lead to the developers of the language model to use that query to improve the model for subsequent builds [1]. The problem of contamination usually leads to *overestimation of capabilities*, and indirectly to *overestimation or underestimation of generality*: models tend to be better on those common domains for which there is a lot of data and tests (e.g., educational tests for mathematics, geography, physics, etc.), and worse on those domains for which fewer tests are available (e.g., spatial reasoning).

A second problem that is yet to be properly recognised is *dissimulation* [23, 16, 11], now recently referred to as 'sandbagging' [42], by which *capabilities and generality can be underestimated*. A system, its developers or providers can pretend the system is less intelligent than it really is during testing, in order to be below some thresholds of capability or generality. Making their systems look worse than they really are seems to be against the logic of most organisations behind big models. However, they can still underperform during (private) testing while publicising good results publicly. For instance, they can fail on purpose in an particular subset of tasks to game the generality metric, thus getting the conformance to operate. Detecting dissimulation in a purely behavioural way is extremely complicated and probably impossible in the long term, despite the current approach of discovering capabilities that appear for many tasks [39] but suspiciously not for others. White-box approaches for mechanistic interpretability may be unfeasible for this, especially if dissimulation is conceived on purpose by the developers, and hence hidden in intricate ways [11]. The most effective option is to take punitive actions (fees) in case intentional dissimulation happens. But the fines must be considerable taking into account how much is at stake to get a product release and how difficult it is to detect dissimulation.

Table 1 summarises major caveats and solutions.

5 Characterising GPAI

We now outline a basic protocol for evaluating whether an AI component, model, system or product should be considered GPAI.

1. **Subject contour:** Determine the AI 'subject' s to be assessed, as a component, model, system or product, its boundaries, dependencies and interactions, especially on human computation or other intelligent systems. If s is accessed through API, determine that the system is not updated or modified to game the evaluation or use the evaluation test for updating its weights or filters [1]. *Example: The subject system is set to be the March 2024 'build' of GPT4-Turbo, with no Internet access or plug-ins, known system prompt and no compute limitations during deployment. For certification, there is previous agreement with Microsoft/OpenAI to evaluate a version on an inspectable or owned server not to be updated during the testing.*
2. **Task incentives and elicitation:** Identify incentives or assurance that the system will not dissimulate, i.e., pretending being less capable it really is. This may be intentional (companies seeking not to be catalogued as GPAI) or unintentional (alignment for toxicity, bias, risks, etc., ends up with a less capable or less general system). *Example: A specific red-teaming effort to elicit dormant or cancelled capabilities shows that there is an underestimate of capability or generality in our assessment. For instance, we cannot discard alignment through RLHF may have introduced some possibility of dissimulation in GPT4 that can be elicited back by some clever contextual information.*
3. **Usage distribution, domains and difficulties:** Elicit a partition and distribution of task domains $\mathbb{P}(D)$ given the context or the population of potential users, a difficulty function $h(i)$ for each instance in each task domain $i \in D$. Identify a difficulty probability $\mathbb{P}(h|D)$ and conditional instance probability $\mathbb{P}(i|h,D)$ per task domain. Sampling can be done by $\mathbb{P}(i) = \mathbb{P}(i|h,D)\mathbb{P}(h|D)\mathbb{P}(D)$. *Example: for a context of generalist mathematics assistants we determine that the addition of natural numbers appears in 5% of the queries of primary-school students according to some real-world usage database of language models (e.g., some specialised versions of [21, 45]). So* $\mathbb{P}(Addition) = 0.05$. *The difficulty function* $h(i)$ *is the sum of the number of digits, the difficulty probability* $\mathbb{P}(h|D)$ *is uniform from* $h = 0$ *to* $h = 20$ *digits and zero elsewhere, and* $\mathbb{P}(i|h,D)$ *is also uniform. Basically, this means that we will calculate the capability for the task by looking at additions of up to 20 digits in total. The same procedure is performed for each of the other considered mathematics domains.*
4. **Battery construction:** Construct a battery of tests using a sample of instances using $\mathbb{P}(i)$ (or stratified according to $\mathbb{P}(D)$ and the other conditional probabilities). Determine the degree of contamination by preliminary testing of the system, introducing variations (new instances) on some of the tasks controlling per (human) difficulty. Explore instructional variations (e.g., prompts) to determine factors or biases that may affect the performance for each task. *Example: For each task explore prompt sensitivity and possible dissimulation, and other variations for each instance, to see what changes affect performance. For addition, this could be ways in which the query and the numbers can be represented, or included in other problems (context).*
5. **Main Testing:** Actually test the system. The whole sample of instances and variations can be used. Alternatively, the evaluation can be adaptive to maximise information about tasks and difficulty thresholds. Capture as much information as possible, including time as a function of difficulty. Collect data and keep it at the instance level for reanalysis [5]. *Example: For addition we could sample instances of a sufficient range of difficulties, or in an adaptive way, we could start measuring additions of medium difficulty and adapt according to the observed performance, so that we would end up calibrating the right level of capability for that task. Testing should log some other information, such as the times for each instance, and controlling whether the length of the queries (because of larger numbers) has some other effects. For internal consistency, the test (or part of it) should be repeated to ensure results are the same for the same queries after some time, checking for memorisation, contamination or cache effects.*
6. **Metrics and Results**. Interpret and plot results (subject characteristic curves). Calculate capability metrics for each D by aggregating $v(s,i)$ per difficulty. Then aggregate by area or derive capability by a threshold. Once done this for all domains, we will try to put each capability in a commensurate scale (e.g., human referencing), to get a *capability profile*. From here, we can calculate a generality metric from the dispersion of this profile. *Example: Results show GPT-4 is very good for addition up to 20 digits (95% accuracy), less good at spatial navigation in a map represented by*

Caveat	Proposed Pragmatic Solution
Benchmark saturation	We use capability levels instead of benchmark performance (saturating near 100%), based on a sufficiently wide range of difficulties, shown as a subject characteristic curve, with no capability limit.
Poor reliability and validity when using benchmark performance	We use capability instead of performance, to extrapolate for different distributions and problems. This frames evaluation as a prediction problem, which can, in turn, be evaluated.
Performance aggregates for different benchmarks are incommensurate	We use capability levels that are anchored on human percentiles and normalised into z-scores, so notions of variation or regularity on a range of domains become meaningful and comparable with humans.
Many interpretations of generality	We choose the most common and actionable interpretation, as the regularity in the capability levels for a range of domains.
Repurposing effort and 'capability elicitation' is hard to quantify	We first consider off-the-shelf models (with state-of-the-art generic elicitation techniques), instructable by non-expert human users. But we can then incorporate cost, effort, safety, etc., in the score function v.
Identifying domains and usage distribution	We initially consider the 14 capabilities in [40] as domains, with a uniform distribution. Subsequently we can use an expert committee and/or perform human studies (daily tasks) to refine this set.
Establishing GPAI models thresholds	We propose several grades (GPAI grade I, II, III, etc.) that depend on the metrics of generality and overall capability, as shown in Figure 1.
Contamination	We can create variations of instances (controlling by difficulty) and keep them private. They can be renewed periodically to counteract testing leaks.
Sandbagging	We should monitor systems in real environments after conformance is given. Increase fines in case companies try to do sandbagging when tested by regulators (higher than in the Volkswagen emission case).
Higher complexity over compute estimation or benchmark performance	Estimating generality and capability gives us profiles and metrics that do not suffer a moving target phenomenon, and can be compared with human profiles to estimate impacts on jobs, education, safety, etc.

Table 1. Caveats for evaluating GPAI and the pragmatic choices and solutions we propose.

a graph, with only more than 95% accuracy up to 3 nodes and excellent causal reasoning of level 7 on a rubric between 0 and 10. If these three tasks were the only ones in $\mathbb{P}(D)$, and had the same weight (probability in $\mathcal{D}$), then after putting them in a commensurate scale with humans, we could have 96.2 percentile for addition, 30.3 percentile for spatial navigation and 65.1 percentile for causal reasoning. Given this capability profile (96.2, 30.3, 65.1), the generality is relatively low, since there is high standard deviation in the profile (33.97). The aggregate capability is 63.87.

7. **Final assessment**. Taking into account all the steps above, apply the established criteria to determine whether a system is GPAI. These could be set as a minimum of aggregate capability *and* a minimum of generality, or quadratic curves resembling the grades on Figure 1. Another option to define the grades or threshold for GPAI could be to set a minimum for a percentage of a domain. For instance, an integrated metric of 40 would mean that for 95% of the domains we expect at least 40 capability (which in our scale means better than 40% of humans). *Example: Following the previous case of addition, spatial navigation and causal reasoning, while the aggregate capability is quite high, 63.87 (higher than average humans), the generality is not enough to consider this system GPAI with the actual thresholds (30.3 is below 40%).*

The above stages are preliminary. We require a detailed protocol and a technical committee to apply it to any new system to be characterised as GPAI. Ideally, we should be able to determine grades of GPAI as shown in Figure 1. The contours of these grades would need further discussion, but a quadratic combination of generality and total capability makes more sense than a linear combination (e.g., a much more capable AlphaGo will never be considered GPAI).

Finally, each evaluation must come with validation. If the evaluation is based on capabilities, as they have predictive power, we can evaluate them as we do with any predictive model at the meta-level, using hold-out data not used in the training or testing of the base models [47]. Also, comparison with (informal) human experts is necessary, using more systematic procedures than leader boards of human preferences[3].

6 Conclusions

In the same way the measurements for a subject (system, model or component) need to be updated regularly, the protocol may need revision from time to time. In particular, since the application of the protocol incorporates many choices, these should be monitored such that the mechanism says something is a GPAI when it really agrees with our intuition. In general, it is better to think of metrics of generality and capability with which we can regulate several 'grades of GPAI', rather than a binary condition of what a GPAI is. Our proposal is to derive the metrics that can place AI systems in a plot like Figure 1, on either side of several grades or frontiers of GPAI. Things to revise periodically include the task usage distribution leading to the domain partition, the commensuration of capabilities (using human reference, by capitalising on tests administered for both humans and AI [37, 44], using equating, scaling or linking [22] or some other approaches) and the thresholds used for the several GPAI grades.

Finally, we have not explored the relation between generality and compute. There can be natural laws for the maximum level of capability per compute that can be achieved. It is no surprise that brain size (at least relative to body mass) correlates with capabilities in animals [26], and even in human populations. The theoretical relation between compute and intelligence is beyond the scope of this document (see, e.g., [17]), but there is empirical evidence in the scaling laws of deep learning [20, 33, 32, 7], showing improvement in all domains, increasing total capability but maintaining or even increasing the generality of the models. There are also theoretical results (compactness property in [19]): given the same compute, if we want to maximise the total capacity of a system, then the best way of doing this is to distribute this capacity by maximising generality, instead of maximising specific niches, because different domains are related in representations and reusability between them. In other words, it is better to have the effort distributed as in Figure 3 (c) than having several spikes of ultracapability. Or, under the perspective of Figure 1, the top-left quadrant of the figure is more efficient than the bottom-right quadrant. Generality in AI is here to stay.

[3] Such as ChatBot Arena (https://chat.lmsys.org/?leaderboard) [6], which is a very informative source, but should not be taken as gold standard [29].

Acknowledgements

This work was funded by GVA CIPROM/2022/6 (FASSLOW) and IDIFEDER/2021/05 (CLUSTERIA), the EC H2020-EU grant agreement No. 952215 (TAILOR), US DARPA HR00112120007 (RECoG-AI) and Spanish grant PID2021-122830OB-C42 (SFERA) under MCIN/AEI/10.13039/501100011033. I would like to thank the comments from Joe Castellano, David Fernández-Llorca, Dario Garigliotti, Emilia Gómez, Lorenzo Pacchiardi, Gabriel Recchia, Wout Schellaert, Jaime Sevilla and Lexin Zhou.

References

[1] S. Balloccu, P. Schmidtová, M. Lango, and O. Dušek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. *arXiv preprint arXiv:2402.03927*, 2024.

[2] E. M. Bender, T. Gebru, A. McMillan-Major, and M. Mitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[3] J. Burden. Evaluating ai evaluation: Perils and prospects. *arXiv preprint arXiv:2407.09221*, 2024.

[4] J. Burden, K. Voudouris, R. Burnell, D. Rutar, L. Cheke, and J. Hernández-Orallo. Inferring capabilities from task performance with bayesian triangulation. *arXiv preprint arXiv:2309.11975*, 2023.

[5] R. Burnell, W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, et al. Rethink reporting of evaluation results in ai. *Science*, 380(6641): 136–138, 2023.

[6] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

[7] R. Dominguez-Olmedo, F. E. Dorner, and M. Hardt. Training on the test task confounds evaluation and emergence. *arXiv preprint arXiv:2407.07890*, 2024.

[8] S. E. Embretson and S. P. Reise. *Item response theory*. Psychology Press, 2013.

[9] D. B. Fogel. *Evolutionary computation: toward a new philosophy of machine intelligence*. John Wiley & Sons, 2006.

[10] I. Fujisawa and R. Kanai. Logical tasks for measuring extrapolation and rule comprehension. *arXiv preprint arXiv:2211.07727*, 2022.

[11] R. Greenblatt, F. Roger, D. Krasheninnikov, and D. Krueger. Stress-testing capability elicitation with password-locked models. *arXiv preprint arXiv:2405.19550*, 2024.

[12] C. I. Gutierrez, A. Aguirre, R. Uuk, C. C. Boine, and M. Franklin. A proposal for a definition of general purpose artificial intelligence systems. *Digital Society*, 2(3):36, 2023.

[13] P. Hase, M. Bansal, P. Clark, and S. Wiegreffe. The unreasonable effectiveness of easy training data for hard tasks. *arXiv preprint arXiv:2401.06751*, 2024.

[14] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[15] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[16] J. Hernández-Orallo. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press, 2017.

[17] J. Hernández-Orallo. Unbridled mental power. *Nature Physics*, 15(1): 106–106, 2019.

[18] J. Hernandez-Orallo. Ai evaluation: On broken yardsticks and measurement scales. In *Workshop on evaluating evaluation of AI systems at AAAI*, 2020.

[19] J. Hernández-Orallo, B. S. Loe, L. Cheke, F. Martínez-Plumed, and S. Ó hÉigeartaigh. General intelligence disentangled via a generality metric for natural and artificial intelligence. *Scientific reports*, 11(1): 22822, 2021.

[20] 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.

[21] H. R. Kirk, A. Whitefield, P. Röttger, A. Bean, K. Margatina, J. Ciro, R. Mosquera, M. Bartolo, A. Williams, H. He, et al. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.

[22] M. J. Kolen, R. L. Brennan, and M. J. Kolen. Test equating, scaling, and linking: Methods and practices. 2004.

[23] S. Lem. *The Futurological Congress (from the Memoirs of Ijon Tichy)*. English version by Michael Kandel. Seabury Press, 1971.

[24] M. Li, P. Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer, 2008.

[25] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[26] C. J. Logan, S. Avin, N. Boogert, A. Buskell, F. R. Cross, A. Currie, S. Jelbert, D. Lukas, R. Mares, A. F. Navarrete, et al. Beyond brain size: uncovering the neural correlates of behavioral and cognitive specialization. 2018.

[27] J. McCarthy. Generality in artificial intelligence. *Communications of the ACM*, 30(12):1030–1035, 1987.

[28] A. Newell, J. C. Shaw, and H. A. Simon. Report on a general problem solving program. In *IFIP congress*, volume 256, page 64. Pittsburgh, PA, 1959.

[29] J. Ni, F. Xue, X. Yue, Y. Deng, M. Shah, K. Jain, G. Neubig, and Y. You. Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. *arXiv preprint arXiv:2406.06565*, 2024.

[30] R. Pintner. Intelligence and its measurement: A symposium–v. *Journal of Educational Psychology*, 12(3):139, 1921.

[31] E. Priani. Ramon Llull. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.

[32] Y. Ruan, C. J. Maddison, and T. Hashimoto. Observational scaling laws and the predictability of language model performance, 2024.

[33] W. Schellaert, R. Hamon, F. Martínez-Plumed, and J. Hernandez-Orallo. A proposal for scaling the scaling laws. In *SCALE-LLM*, pages 1–8. ACL, 2024.

[34] W. J. Schneider and K. S. McGrew. The cattell-horn-carroll model of intelligence. 2012.

[35] S. J. Shettleworth. *Cognition, evolution, and behavior*. Oxford university press, 2009.

[36] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[37] J. W. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo, S. Panzeri, G. Manzi, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11, 2024.

[38] Z. Sun, L. Yu, Y. Shen, W. Liu, Y. Yang, S. Welleck, and C. Gan. Easy-to-hard generalization: Scalable alignment beyond human supervision. *arXiv preprint arXiv:2403.09472*, 2024.

[39] E. Todd, M. L. Li, A. S. Sharma, A. Mueller, B. C. Wallace, and D. Bau. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*, 2023.

[40] S. Tolan, A. Pesole, F. Martínez-Plumed, E. Fernández-Macías, J. Hernández-Orallo, and E. Gómez. Measuring the occupational impact of ai: tasks, cognitive abilities and ai benchmarks. *Journal of Artificial Intelligence Research*, 71:191–236, 2021.

[41] I. Triguero, D. Molina, J. Poyatos, J. Del Ser, and F. Herrera. General purpose artificial intelligence systems (gpais): Properties, definition, taxonomy, open challenges and implications. *arXiv preprint arXiv:2307.14283*, 2023.

[42] T. van der Weij, F. Hofstätter, O. Jaffe, S. F. Brown, and F. R. Ward. Ai sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*, 2024.

[43] X. Wang, L. Jiang, J. Hernandez-Orallo, L. Sun, D. Stillwell, F. Luo, and X. Xie. Evaluating general-purpose ai with psychometrics. *arXiv preprint arXiv:2310.16379*, 2023.

[44] T. Webb, K. J. Holyoak, and H. Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

[45] T. Xie, D. Zhang, J. Chen, X. Li, S. Zhao, R. Cao, T. J. Hua, Z. Cheng, D. Shin, F. Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*, 2024.

[46] D. Yu, S. Kaur, A. Gupta, J. Brown-Cohen, A. Goyal, and S. Arora. Skill-mix: A flexible and expandable family of evaluations for ai models. *arXiv preprint arXiv:2310.17567*, 2023.

[47] L. Zhou, P. A. Moreno-Casares, F. Martínez-Plumed, J. Burden, R. Burnell, L. Cheke, C. Ferri, A. Marcoci, B. Mehrbakhsh, Y. Moros-Daval, et al. Predictable artificial intelligence. *arXiv preprint arXiv:2310.06167*, 2023.