A Framework for the Categorisation of General-Purpose AI Models under the EU AI Act

Lorenzo Pacchiardi[1], John Burden[1], Fernando Martínez-Plumed[2], José Hernández-Orallo[1,2], Emilia Gómez[3], David Fernández-Llorca[3]

[1]Leverhulme Centre for the Future of Intelligence, University of Cambridge, UK
[2]Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Spain
[3]European Commission, Joint Research Centre (JRC), Seville, Spain
lp666@cam.ac.uk, jjb205@cam.ac.uk, fmartinez@dsic.upv.es, jorallo@upv.es, emilia.gomez-gutierrez@ec.europa.eu, david.fernandez-llorca@ec.europa.eu

Abstract

We propose a framework for categorising AI models as General-Purpose AI (GPAI) models as defined in the European Union (EU) AI Act, based on their capabilities and generality. Our framework breaks down the core components of the GPAI model definition into measurable elements, focusing on four primary cognitive domains: Attention and Scan, Comprehension and Compositional Expression, Conceptualisation, Learning and Abstraction, and Quantitative and Logical Reasoning. We suggest using the Annotated Demand Levels (ADeLe) procedure to evaluate AI models' capabilities in these domains, and provide a methodology for combining domain-level scores into a single measure of generality. The framework is illustrated with empirical results from existing models, and policy recommendations are made for selecting thresholds and metrics for GPAI model categorisation.

1 Introduction

The EU AI Act [7] entered into force on 1 August 2024, with the obligations for the providers of general-purpose AI (GPAI) models applying from 2 August 2025. While Article 3(63) defines GPAI models as AI models *"that [display] significant generality and [are] capable of competently performing a wide range of distinct tasks [...] and that can be integrated into a variety of downstream systems or applications"*, this definition does not set out specific criteria and its operationalisation allows for multiple approaches based on different interpretations.

The recently published guidelines on the scope of obligations for GPAI models [6] provide an indicative criterion based on the amount of compute used to train an AI model, as substantial literature on scaling laws [11, 9] shows that this has been, until recently, a relatively good proxy for a model's performance in a wide range of domains[1]. Here, we propose an alternative framework that does not rely on the continued validity of proxy metrics, but instead directly attempts to measure the model's capabilities and its potential for generality, ensuring consistency with evolving AI capabilities[2].

[1]These are already losing validity with some new optimisations, and the introduction of "reasoning" models, which reach stronger performance in some domains with the use of additional compute at test-time [19, 8].

[2]Note that the guidelines acknowledge that training compute is an imperfect proxy for generality and capabilities and that in the future, if deemed appropriate, the European Commission may take into account benchmarks to determine whether a model is a general-purpose AI model.

Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

In particular, our proposal expands the framework in [10] by operationalising capabilities and generality in a way that balances scientific rigour with practical feasibility. Our operationalisation breaks down an AI model's core capabilities into key cognitive domains, grounded in a careful identification of relevant cognitive and knowledge capabilities, inspired by several human and artificial intelligence taxonomies, including the Cattell-Horn-Carroll (CHC) theory of intelligence [2, 12], along with more recent adaptations for the AI domain [18]. The result is a list of 14 core cognitive abilities spanning a broad range of domains, chosen to reflect the kinds of flexible, domain-general behaviours that GPAI models are expected to exhibit. To reduce the burden on developers, we reduce this to a core set of four domains we believe to be most pertinent: *Attention and Scan*, *Comprehension and Compositional Expression*, *Conceptualisation, learning and abstraction*, and *Quantitative and logical reasoning*. Then, model capability on these domains can be measured through instance-level analysis of existing benchmarks to derive meaningful profiles that are combined into a final generality metric.

While we illustrate our approach concretely and consider various decision rules, aggregation functions and thresholds for the binary determination of whether an AI model should be considered a GPAI model, we do *not* propose specific aggregation functions or numerical thresholds, as these should be determined to align our proposed methodology with legal, technical and regulatory developments.

2 Considerations while developing and applying the framework

2.1 Development considerations

Our framework (see [1] for an extended version) attempts to: (1) **adhere to the definition of GPAI model** as outlined in Article 3(63) (Appendix B) of the EU AI Act and relevant recitals (Appendix B.1); (2) **exempt small models** trained using relatively small datasets; (3) enable **easy determination of whether a given AI model qualifies as a GPAI model** or not through objective criteria that are straightforward to apply; (4) be **difficult to circumvent** or manipulate; and (5) remain **relevant for approximately the next two years**.

2.2 Application considerations: system-level components in testing an AI model

Hernández-Orallo et al. (2024) [10] note that, while GPAI categorisation applies to a *model*, evaluating capabilities and generality benefit from considering the model together with *system-level* components; this aligns with the definition of GPAI model (Appendix B) which requires models to competently perform a wide range of tasks and have the potential to be integrated into diverse downstream systems; thus, generality can be assessed in the context of the systems in which a model is embedded.

System-level components influence performance, and therefore capability and generality, primarily via: (1) **User interaction** (UI/UX): for instance, an LLM without a usable interface is effectively inert; conversely, well-designed interfaces and APIs enabling multi-turn interaction and larger context windows help users correct and steer the model, improving task performance. (2) **Tool availability**: teaching and prompting models to use external tools and scaffolds expands capabilities and supports data enhancement [3]; for example, web search provides information the model lacks, and calculators overcome arithmetic limits, enabling more complex mathematical reasoning.

Therefore, evaluations of whether a model qualifies as a GPAI model should specify and, where possible, standardise the set of system-level components available during testing. In Appendix C, we propose a set of standardised conditions; however, our framework described in Section 3 can be applied with any set of conditions, as long as these are standardised.

3 Operationalising the definition of a GPAI model

To effectively categorise AI models as GPAI models, we propose to break down the core components of the GPAI model definition and make them measurable. According to the definition in Article 3(63) (Appendix B), a GPAI model: (1) **shows significant generality**, (2) **is capable of competently performing a wide range of distinct tasks**, (3) can be integrated into a variety of downstream systems or applications, and (4) is not exclusively used for research, development, or prototyping activities before it is placed on the market. The third point combines accessibility aspects with the use of system-level components (as discussed in Section 2.2), while the fourth is purely of legal nature. Therefore, here we focus primarily on the first two elements.

We frame generality in terms of **abstract capability domains**, i.e., constructs required for some cognitive tasks but not others. Consistent, competent performance across a wide range of domains suffices for GPAI categorisation[3]. From this perspective, we pose four key questions:

1) Which cognitive domains should be investigated? (Section 3.1). **2)** What tests and methodology should we use to evaluate each of the considered domains? (Section 3.2). **3)** What does it mean to perform consistently and competently in a particular domain? (Section 3.3). **4)** How should capability across domains be combined to determine generality? In particular, how "wide" a range of domains is needed to demonstrate generality? (Section 3.4). We explore these questions in the following sections, and Appendix E includes a step-by-step protocol to categorise a new AI model as a GPAI model which relies on the answers to these questions.

3.1 Identifying cognitive domains

The Cattell-Horn-Carroll (CHC) theory [2, 12] is a widely adopted model that organises cognitive abilities hierarchically, with general intelligence (g) at the top, broad abilities in the middle, and narrow abilities below. CHC specifies 10 broad and 70+ narrow abilities across a range of cognitive abilities. While human taxonomies can serve as inspiration, other kinds of intelligence, non-human, but especially AI, may not share this hierarchy. Integrating several hierarchies and taxonomies in the literature of human intelligence, artificial intelligence and cognitive science, Tolan et al. (2021) [18] derive a list of 14 cognitive abilities (domains) (see Table 1). These domains span many capabilities, though not all are necessary for a model to be categorised as a GPAI model (e.g., sensorimotor interaction may be unnecessary for many cognitively demanding tasks). In Table 1, we highlight what we consider as the four domains most pertinent for GPAI categorisation: *Attention and Scan* (AS), *Comprehension and Compositional Expression* (CE), *Conceptualisation, Learning, and Abstraction* (CL), and *Quantitative and Logical Reasoning* (QL). In our framework, we rely on these four domains. Additionally, CE and QL are each split into two subdomains. Complete definitions for the domains and subdomain, adapted and rephrased from [18, 20] for use beyond workplace contexts, are provided in Appendix D.

Table 1: AI cognitive domains from [18]. In bold, those that we use in our framework.

Memory processes (MP)	Communication (CO)
Sensorimotor interaction (SI)	Emotion and self-control (EC)
Visual processing (VP)	Navigation (NV)
Auditory processing (AP)	**Conceptualisation, learning and abstraction (CL)**
Attention and Scan (AS)	**Quantitative and logical reasoning (QL)**
Planning and sequential decision-making and acting (PA)	Mind modelling and social interaction (MS)
Comprehension and compositional expression (CE)	Metacognition and confidence assessment (MC)

The four domains we use in our framework constitute a minimum set which we deem necessary for a model to be considered as a GPAI model, but they are not the only capabilities that GPAI models may possess. Moreover, the identified capabilities concern parsing inputs and generating outputs, but the definition of GPAI model also requires goal-directed action; we operationalise this via behavioural evaluations and benchmarks that test a model's ability and propensity to complete assigned tasks.

3.2 Testing each capability domain

Having identified the four primary domains, we could identify AI benchmarks relevant to each domain and simply considering a model's aggregate performance (such as its accuracy). However, this is problematic because, among other issues [4], even if a benchmark claims to test a capability such as Attention and Scan, its instances may require a model to be skilled at other capabilities (e.g., Quantitative and Logical Reasoning). Simple aggregations do not yield a capability estimate that is independent of the way in which the benchmark instances are defined.

One could try to tackle this by averaging across multiple benchmarks, hoping that this "cancels out" these confounders. However, averages across benchmarks do not take into account the different difficulty of the instances (which arises from the needed skills across multiple capabilities) or the different random guess accuracy rates (e.g., 50% vs. 33%).

[3] Interpretations of "tasks" and "wide range" co-vary; tasks may be taken as domains or specific problem types, provided "wide range" is specified.

To solve these issues and obtain a more accuracy estimate of models' capabilities, we rely on the approach in Zhou et al. (2025) [20], based on measurement scales: annotate individual task instances by the "demands" they pose on the various capability domains and then analyse the performance of AI models at the level of individual instances to identify how the various demands impact performance. This requires initial scale calibration and rubric construction, after which annotation is automatable and rubrics are reusable. We describe this method in more detail in Section 3.2.2.

3.2.1 Testing modalities

Nevertheless, when selecting instances to evaluate a particular domain, we must first consider what *modality* an AI model operates in. AI models differ in input/output modalities, with the same capabilities (e.g., reasoning) manifesting and being tested per modality[4]. For multimodal systems, competence in any one modality suffices to claim domain competence for the model. Clearly, some modalities are too constrained to permit competence broadly (e.g., a single-bit output limits expression). Common modalities are text (including code), images, audio (including speech), sensorimotor, and tabular data, which can be mixed (e.g., text+image input). Most frontier models currently support text, image, and audio I/O, but robotics and novel modalities (e.g., olfactory) may expand this. Importantly, all input modalities are ultimately digitised so that, in principle, any input can be processed in any modality. In practice, evaluation should match the modality the model was designed for: a text-only model could receive audio-derived binaries, but will likely fail to interpret them without audio-specific training.

3.2.2 Annotating demands and measuring capabilities

Annotating Demands Zhou et al. (2025) [20] introduce the Annnotated Demand Levels (ADeLe) procedure for annotating task instances, at the core of which is the idea that a single task instance can pose demand of different levels on different cognitive capabilities. For example, a question might require a low level of Attention and Scan (AS) but high level of Quantitative and Logical reasoning (QL). By identifying the demands of each instance, an AI model's response to many task instances can be informative about its ability to successfully respond to varying levels of demands.

To practically annotate a large dataset of task instances, rather than having humans assign the demands manually, Zhou et al. (2025) [20] takes advantage of existing LLM's capability to robustly apply rubrics. By carefully developing precise rubrics identifying what the demands of a question are, an LLM can assign those demands automatically, which greatly reduces the burden of applying such a method. Zhou et al. (2025) [20] independently validated the introduced rubrics and the resulting annotation by human reviewers through inter-rater analysis and the Delphi method [13].

The resulting ADeLe battery[5] includes 16,000 task instances from high-quality modern benchmarks annotated for a wider range of capabilities, but including the four ones we propose using for GPAI model categorisation. Thus, we employ this battery in our empirical validation in Section 4 and recommend regulators and other stakeholders do the same. Of course, however, the annotation could be extended (and should, after some time has passed, to avoid contamination) to new datasets, which could provide more refinement for the considered capabilities or to cover specific areas.

The scales used in Zhou et al. (2025) are "ratio scales" [16], which contain an *absolute zero*, where demands are not present at all. Ratio scales also require the differences between levels to be consistent across the scale and ensure comparability across different capabilities. When defining the rubrics, Zhou et al. (2025) uses the rule of thumb that doubling the demand should halve the log odds of success. A second level of calibration can further turn these levels into more meaningful scales (e.g., where an ability at level l is representative of one in 10^l humans being able to solve the task). We describe how this could be done in Section 3.3.

Measuring Capabilities Once demands are annotated on a test battery, an AI model can be tested and its responses "sliced" across individual demands, yielding "subject characteristic curves" that detail how an AI model's average score depends on the level of a particular demand. Zhou et al.

[4] E.g., common-sense reasoning can be tested in written or spoken form for text-to-text (LLMs) and audio-to-audio (e.g., Alexa) models.

[5] ADeLe v1.0: A battery for AI Evaluation with explanatory and predictive power. `https://kinds-of-intelligence-cfi.github.io/ADELE/`

(2025) employ a "non-dominant" strategy: for a considered domain and demand level, all instances for which other domains have demand levels above the considered one are discarded.

Following psychometric tradition, an AI model is assigned ability l for a considered domain if it can succeed at demand level (for that domain) l with 50% probability [17]. A single pass through the questions of a benchmark by an AI model can measure all capabilities simultaneously; Figure 1 reports the obtained abilities for some LLMs.

3.3 Competently performing in a domain

Now that we have identified a set of domains and a methodology for measuring capabilities, we need to answer the question: *"What do we mean in practice by competently performing in a particular domain?"*. Empirically, to answer this, we need to determine how to convert the capability scores obtained by the approach in Section 3.2.2 into a decision. In practice, we suggest to norm the scales from the ADeLe approach relatively to a human population and convert the model's capability to this human-normed scale. While alternative baselines could be used, such as a population of AI models or a single AI model sampled multiple times, these would be less representative of general-purpose intelligence (which is usually attributed to humans) and would be more contingent on the specific choice of AI models, thus yielding a less appropriate model categorisation.

The core idea is to calibrate each demand level (e.g. level l) to a corresponding human probability of success. In particular, we propose to use a logarithmic scaling, which corresponds to making sure that demand l corresponds to 1 in b^l humans (from a specified population) responding correctly (e.g. if base $b = 10$, then "level $3 \approx 1$ in 1,000 people can solve"). This does not require the creation of any new question, or altering them in any meaningful way; instead, humans can be tested on existing AI benchmarks (such as those in the ADeLe battery) using platforms such as Prolific[6]. For the lower demand levels, it would be sufficient to do this on $\sim$100s instances per domain[7], as dozens of human responses per item would be sufficient to accurately estimate them. On the other hand, higher levels of capability would require extremely large samples of human respondents to observe even a single success. For example, confirming that a task is at a "1 in 100,000" difficulty would naively require testing at least 10^5 people, clearly infeasible in terms of recruitment and cost. This can be addressed with several strategies.

It is important to realise that we only need to calibrate the scales once. We do not need to apply this procedure for any single annotation of a task instance, which will continue to be completely automated using a calibrated rubric. One strategy is selecting a sample of questions that are labelled as high level l and test a smaller number of people on them. For instance, if we have 1,000 people on 20 questions, we would have 20,000 evaluations of level 4, which we could then calibrate to see if the results are in the appropriate range. Another alternative is to recruit targeted human samples for difficult items rather than random individuals, thus increasing the likelihood of obtaining a few correct responses even for high-demand items. This biases the sample, but this would need to be recalibrated by considering the frequency of this population in the overall population. These strategies are considered, at least informally, in subsequent calibrations of [20]. At the moment, for the non-knowledge dimensions we are using here (the actual capabilities), we consider that for level l approximately 1 person gets it correctly in a sample of 2^l people). This human-normed calibration has margin of improvement, but will be used for the purpose of illustration in the rest of this paper.

3.4 Wide range and generality

We now have four primary domains (Section 3.1) that we propose for categorising a model as GPAI. The procedure discussed in Sections 3.2 and 3.3 can be used to obtain a score (i.e., the model's capability on the human-normed scale) for each domain. We now need to combine the results and categorise the model. The simplest approaches to do so include:

- Combining the capability levels across domains through an average and setting a threshold on the average (e.g., level 3, corresponding to a human-normed proportion for only 12.5% of the population being correct) above which the model is categorised as a GPAI model. Different mathematical averages can be used:

[6] `https://www.prolific.com/`

[7] Zhou et al. (2025) [21] include an "ADeLe light" subset of the larger ADeLe batter.

- The *Arithmetic mean* if we want to allow majority-performance to lift-up or drag-down areas of lower/higher performance.
- The *Geometric mean* or the *Harmonic mean* if we want to be more cautious about applying GPAI status.

• Identifying a suitable threshold value for each domain (e.g., level 2, corresponding to only 25% people being correct), marking each domain as Pass vs. Fail, and categorising a model as a GPAI model if there is a sufficient number of dimensions with a pass (e.g., 2+).

Both approaches allow to easily alter the threshold needed for GPAI categorisation, in order to keep the framework up-to-date with newly released models. However, small changes in performance near the threshold could potentially be gamed depending on the metric used.

In Section 4.1, we apply the methodology to a range of existing models and provide policy recommendations for how to appropriately select thresholds and metrics based on policy objectives.

4 Empirical Considerations of the proposed approach

4.1 Sensitivity analysis of classification thresholds and averages

We empirically assess how changing the thresholds and aggregation methods for domain-level capability scores impacts the classification of AI models as GPAI under the operational framework previously described. Given the relevant role of the choice of thresholds (e.g., the minimum domain ability score) and aggregation function (e.g., mean, harmonic mean) in determining GPAI status, it is important to understand the robustness and practical effects of these choices. We use a diverse cohort of publicly available LLMs, as analysed in [20]. For each model, we obtain domain-level ability scores (i.e., the demand level at which the model achieves at least 50% success) across the four primary domains described in Section 3.1, dervied from the ADeLe framework. This is important in our consideration of the levels. For instance, a model with level 4 does not mean it has 100% chance of succeeding at questions for which about 6% ($\sim 2^{-4}$) of the population is correct but 50% chance of doing so.

For the analysis, we explore a range of **Aggregation functions:** (arithmetic, harmonic, and geometric means), **Thresholds:** (values from which systems are considered capable; 3-4.5+), and **Pass/fail rules:** (requiring either all domains to exceed a threshold, or for at least N out of 5 domains).

4.1.1 Effect of aggregation function

The choice of aggregation function materially affects a model's aggregate score, and thus the set of models classed as GPAI models at any fixed threshold. The arithmetic mean is most forgiving: high scores in one domain can more easily compensate for lower scores in another. The geometric mean is slightly more stringent, as underperformance in any one domain will more severely dampen the average. The harmonic mean is most conservative and heavily penalises low scores in any domain.

However, as Table 2 demonstrates, the differences between these aggregation methods are relatively modest (generally less than 0.05–0.1) for the set of LLMs evaluated, especially for models that exhibit balanced performance across all domains. In cases where a model particularly weak in a single domain, the harmonic mean naturally penalises this more. However, for most of the models tested, abilities are sufficiently balanced that all aggregations yield similar GPAI categorisation outcomes. Consequently, the GPAI status categorisation for these models is typically robust to the specific averaging method chosen. Figure 1 shows radar plots of model ability profiles and further illustrates this point: the regular shapes of the high-performing models result in small differences across aggregation methods. Only models with pronounced bottlenecks in specific domains would be materially affected by aggregation function, which is our intention.

For this reason, the arithmetic mean may be a reasonable choice given it is the simplest, most intuitive aggregation method, and produces almost identical results to other aggregation methods based on current model profiles

Table 2: Arithmetic mean, geometric mean, and harmonic mean of per-domain ability scores for evaluated LLMs. Scores are calculated across four key domains used for GPAI categorisation. The choice of aggregation function affects whether uneven domain performance is penalised (harmonic mean) or compensated (arithmetic mean).

	Model	Arith. Mean	Geom. Mean	Harm. Mean
DeepSeek	DK-R1-Dist-Qwen-1.5B	2.85	2.83	2.81
	DK-R1-Dist-Qwen-7B	3.68	3.66	3.64
	DK-R1-Dist-Qwen-14B	4.13	4.11	4.09
	DK-R1-Dist-Qwen-32B	4.40	4.37	4.34
GPT	Babbage-002	0.57	0.50	0.43
	Davinci-002	0.90	0.83	0.77
	GPT-3.5-Turbo	2.30	2.28	2.26
	GPT-4o	3.90	3.87	3.84
	OpenAI o1-mini	4.44	4.42	4.40
	OpenAI o1	5.33	5.26	5.19
LLaMA	LLaMA-3.2-1B-Instruct	1.61	1.59	1.56
	LLaMA-3.2-3B-Instruct	2.35	2.33	2.32
	LLaMA-3.2-11B-Instruct	2.63	2.61	2.60
	LLaMA-3.2-90B-Instruct	3.63	3.60	3.58
	LLaMA-3.1-405B-Instruct	3.73	3.71	3.70

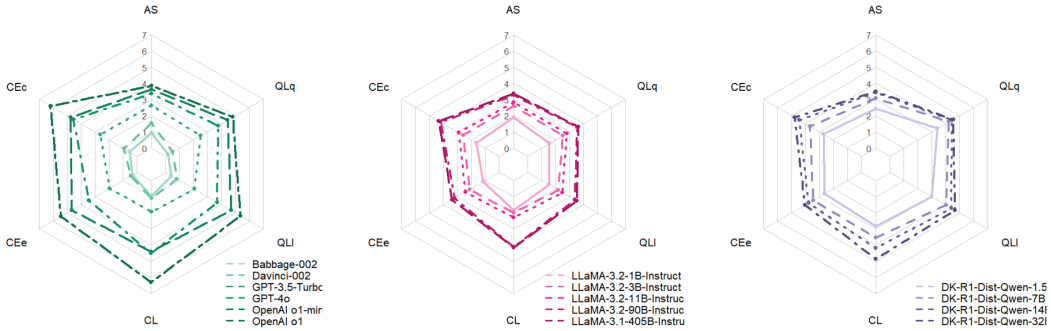


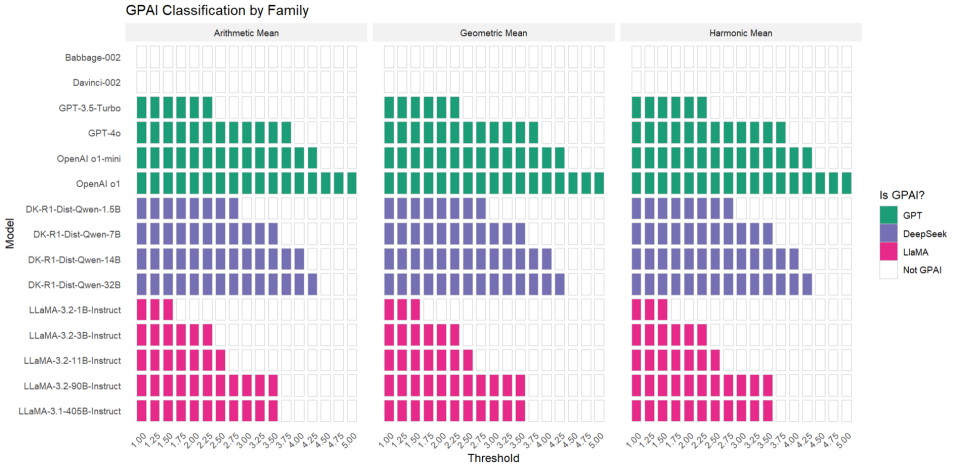
Figure 1: Radar plots showing per-domain ability scores (ADeLe scale; higher is better) for major LLM families. Left: OpenAI models; Middle: LLaMA models; Right: DeepSeek (DK-R1-Dist-Qwen) models). These profiles illustrate strengths and weaknesses across the evaluated abilities and underpin aggregate GPAI scoring. For the rest of the paper the two subdomains CEc-CEe and QLq-QLl are merged into single domains by arithmetic averaging.

4.1.2 Effect of threshold value

Selecting threshold values is also important in categorising GPAI models, as it directly defines what constitutes 'competent' performance in each cognitive domain or in aggregate. In our experiments, we systematically varied the minimum required ability threshold using the ADeLe scale, which typically ranges from 3.0 (intermediate) to 4.5 (well above average), to observe its impact on model classification. As visualised in Figure 2a, we observe the following trends:

- At **Lower Thresholds** (e.g., $\sim 3.0 - 3.5$): A broad range of models, including smaller and older models, are categorised as GPAI models. Many models included at this level may not exhibit robust, human-comparable abilities across all domains. This risks over-inclusivity, where the "GPAI" designation is granted to models that users or experts may not intuitively consider truly general-purpose.

- At **Intermediate Thresholds** (e.g., 4.0): Only models with consistently high abilities across domains, typically the most capable, modern LLMs, achieve GPAI status. This setting aligns well with regulatory expectations for "competent" performance and appears to capture the point at which models that users or experts intuitively consider to be truly general-purpose (e.g. GPT4-o in Figure 2a) are categorised as GPAI models.



(a)

(b)

Figure 2: Comparative visualisations of GPAI classification outcomes. for all models (rows) as a function of threshold (columns) and aggregation method (three panels: arithmetic mean, geometric mean, harmonic mean). Each cell indicates whether the model is classified as GPAI at the given threshold under the specified aggregation.(a) Heatmap of outcomes by threshold and aggregation method. (b) Classification results under varying N-domain rules.

- At **High Thresholds** (e.g., ≥ 4.5): Only the top-performing models (e.g., OpenAI o1) retain GPAI status. Slight deficits in a single domain can exclude models that are otherwise highly capable.

The choice of threshold provides policymakers with an adjustable tool to tune the strictness of the GPAI model definition. A lower threshold involves more AI models within the scope of the obligations for GPAI models, but could diminish the meaningfulness of the "GPAI" standard. A higher threshold increases reliability but can quickly shrink the pool of qualifying models, possibly below what the law intends. Importantly, these trends hold across all three aggregations (arithmetic,

geometric, harmonic means) for our data, indicating that it is the threshold value, rather than the precise aggregation formula, that most directly drives changes in model categorisation.

When a threshold for a model being a GPAI model has been set at a clearly justified reference point (e.g., 4.0 on the ADeLe scale), it should be periodically recalibrated based on new model performance data and evolving policy objectives. Policymakers may also consider publishing the chosen threshold and rationale to support transparency and regulatory certainty.

4.1.3 Domain pass/fail policy

Figure 2b illustrates the effects of domain-level policies. If the rule requires all domains to meet the threshold then only models with uniform capability are classified as GPAI, and this set shrinks rapidly as the threshold rises. In contrast, relaxed policies (e.g., GPAI if ≥ 3 of 4 domains pass) result in broader inclusion, accommodating models that are strong in most, but not all, domains, a realistic consideration given that even the best models have the occasional 'blind spot' leading to particular deficits. Such policies may disqualify models that, by most practical standards, would still perform at a level consistent with the intent of general-purpose systems.

Our results suggest that the threshold value itself is not the only important factor in determining GPAI status; the point at which the domain-level policy is set (i.e., how many domains must reach the bar) is also important, particularly for models close to the threshold. For many current LLMs, a small change in the number of domains required to pass can result in several models changing their GPAI status, particularly when their aggregate abilities are tightly clustered.

One possible approach in this regard would be to require models to exceed the competency threshold in at least three out of four (or a similar proportion) of the assessed domains to qualify as a GPAI model. Any chosen approach needs to balance the need for broad, reliable capabilities with realistic expectations regarding minor weaknesses, and should be reviewed periodically as domain-specific requirements evolve.

5 Conclusions

There are increasing concerns that proxies based on number of parameters, training or inference FLOP, or even project budget may be insufficient to capture the actual capabilities and generality of an increasingly more diverse landscape of LLMs (CoT, RL, multimodal, routed, ensemble, etc.) and their integration into systems and agents with changing affordances for real-world applications. There is no reliable shortcut for determining the capabilities of an AI model other than *measuring* them.

We introduced a framework that could be used to determine whether an AI model should be categorised as GPAI model, drawing on measurement scales, cognitive psychology, psychometrics and traditional concepts of generality in AI, with the goal of grounding GPAI categorisation in scientifically robust criteria. Central to this framework is a set of four core cognitive abilities, selected to reflect a diverse and representative range of domains that jointly characterise general-purpose intelligence. To apply this framework, we assign, to each task instance, a demand profile, enabling us to treat capabilities as latent traits expressed to varying degrees across tasks. We adopt the ADeLe methodology [20] for this purpose, leveraging LLM to annotate task demands using carefully constructed and validated rubrics. Given an already annotated battery such as ADeLe, a reliable estimation of capability levels of any new models can be done with a few hundred examples, which is extremely efficient compared to evaluating on a range of (usually large) benchmarks.

We proposed two complementary strategies for assigning GPAI status based on the final capability profile. The first aggregates across dimensions using averages. The second uses a thresholding approach: a model qualifies as a GPAI model if it meets or exceeds a fixed performance bar in a sufficient number of cognitive domains. We deliberately refrain from prescribing fixed thresholds. The appropriate standard for GPAI categorisation should be determined in accordance with legal interpretations of the EU AI Act and updated as the field evolves. However, we empirically study the effect of varying metrics and thresholds on the categorisation of existing AI models. Together, our work establishes an empirically grounded and practically useful operationalisation of the definition of GPAI models: one that can evolve alongside the models it is designed to evaluate.

Disclaimer

The views expressed in this document are purely those of the authors and may not, under any circumstances, be regarded as an official position of the European Commission.

References

[1] J. Burden, L. Pacchiardi, F. Martínez Plumed, and J. Hernández Orallo. A Framework for General-Purpose AI Model Categorisation. In D. Fernández Llorca and E. Gómez, editors, Collection of External Scientific Studies on General-Purpose AI Models under the EU AI Act. Publications Office of the European Union, Luxembourg, JRC143256, 2025.

[2] John Bissell Carroll. Human cognitive abilities: A survey of factor-analytic studies. Number 1. Cambridge university press, 1993.

[3] Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos, and Guillem Bas. Ai capabilities can be significantly improved without expensive retraining. arXiv preprint arXiv:2312.07413, 2023.

[4] Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. arXiv preprint arXiv:2502.06559, 2025.

[5] European Commission. Approval of the content of the draft Communication from the Commission - Commission Guidelines on the definition of an artificial intelligence system established by Regulation (EU) 2024/1689 (AI Act). `https://ec.europa.eu/newsroom/dae/redirection/document/112455`, 2024.

[6] European Commission. Guidelines on the scope of the obligations for general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act). `https://ec.europa.eu/newsroom/dae/redirection/document/118340`, 2025.

[7] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. `https://eur-lex.europa.eu/eli/reg/2024/1689/oj`, 2024.

[8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

[9] Lennart Heim and Leonie Koessler. Training compute thresholds: Features and functions in ai regulation. arXiv preprint arXiv:2405.10799, 2024.

[10] J. Hernández-Orallo, J. Sevilla, E. Gómez, and D. Fernández-Llorca. General-Purpose AI Models in the AI Act: Capabilities, Generality, Systemic Risks and Compute. European Commission, Joint Research Centre, JRC139341, 2024.

[11] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

[12] Timothy Z Keith and Matthew R Reynolds. Cattell–horn–carroll abilities and cognitive tests: What we've learned from 20 years of research. Psychology in the Schools, 47(7):635–650, 2010.

[13] HA Linstone. The delphi method: Techniques and applications in linstone ha, turoff m. URL: http://www. is. njit. edu/pubs/delphibook, 1975.

[14] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering, September 2018. arXiv:1809.02789 [cs].

[15] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark, November 2023. arXiv:2311.12022 [cs].

[16] Stanley Smith Stevens. On the theory of scales of measurement. Science, 103(2684):677–680, 1946.

[17] Louis Leon Thurstone. Ability, motivation, and speed. Psychometrika, 2(4):249–254, 1937.

[18] Songül Tolan, Annarosa Pesole, Fernando Martínez-Plumed, Enrique Fernández-Macías, José Hernández-Orallo, and Emilia Gómez. Measuring the occupational impact of ai: tasks, cognitive abilities and ai benchmarks. Journal of Artificial Intelligence Research, 71:191–236, 2021.

[19] Pablo Villalobos and David Atkinson. Trading off compute in training and inference, 2023. Accessed: 2025-07-10.

[20] Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M Collins, Yael Moros-Daval, Seraphina Zhang, Qinlin Zhao, Yitian Huang, Luning Sun, Jonathan E Prunty, et al. General scales unlock ai evaluation with explanatory and predictive power. arXiv preprint arXiv:2503.06378, 2025.

[21] Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. Nature, 634(8032):61–68, 2024.

A List of abbreviations and definitions

AI Artificial Intelligence

API Application Programming Interface

AUROC Area Under the Receiver Operating Characteristic (curve)

CoP Code of Practice

FLOP Floating Point Operations

GPAI General-Purpose Artificial Intelligence

GPAI models General-Purpose Artificial Intelligence models

GPAISRs/GPAISR/GPAI-SR General-Purpose Artificial Intelligence models with Systemic Risks

LLM Large Language Model

ROC Receiver Operating Characteristic (curve)

UI User Interface

UX User Experience

B Background: definitions and recitals

In this section, we provide some definitions and recitals that provide useful context regarding the categorisation of AI models as GPAI models.

First, in the EU AI Act, the definition of GPAI model is given in Article 3(63):

- *'General-purpose AI model' means an AI model, including* ***where such an AI model is trained with a large amount of data using self-supervision at scale****, that* ***displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market***

B.1 Recitals

A number of recitals of the EU AI Act clarify how the concept of GPAI model should be interpreted. For instance, Recital 97 states:

- *'[...] The definition should be based on the key functional characteristics of a general-purpose AI model, in particular the generality and the capability to competently perform a wide range of distinct tasks. [...]'*

Recital 98:

- *" Whereas the generality of a model could, inter alia, also be determined by a number of parameters,* ***models with at least a billion of parameters and trained with a large amount of data using self-supervision at scale*** *should be considered to display significant generality and to competently perform a wide range of distinctive tasks.'*

And recital 99:

- *'**Large generative AI models are a typical example for a general-purpose AI model, given that they allow for flexible generation of content**, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks.'*

In the above paragraphs, bolding is ours.

B.2 AI model and system

The definition of AI model is not given in the AI Act. The glossary in Hernández-Orallo et al. (2024) [10] gives the following definition of AI model:

- *An operative abstraction of a parcel of the world, parametrised or not, which is usually trained from data. The better the model represents the world and captures its patterns, the more it can be used to make predictions, give explanations or perform simulations about the world.*

While Sec 2.1 of Hernández-Orallo et al. (2024) [10] says:

- *'AI model' is a physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, process or data, that is used to make inferences from inputs in order to produce outputs. An AI system is typically built by combining one or more AI models.*

In contrast, the EU AI Act (Article 3(1)) defines an AI system as:

- *'a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.'*

We notice that the European Commission has recently released guidance for the interpretation of the definition of AI systems [5].

C Testing conditions

To ensure that performance metrics accurately reflect a model's true capabilities across different domains, we need to define a set of standardised test conditions to (1) promote reproducibility; and (2) allow fair comparison between models. The following guidelines outline recommended testing conditions:

- **Autonomous evaluation:** Models must be evaluated in a fully autonomous, hands-off manner, without real-time human intervention during test. The goal is to minimise the variability that may be introduced by human assistance.

- **Standardise the environment and protocols:** All evaluations should be conducted in a pre-specified computational environment with fixed hyperparameters (e.g., sampling strategy, temperature, prompt format) to ensure consistency across tests. Evaluation scripts, datasets and benchmarking code must be made publicly available.
- **Handling of system-level components:** The set of additional system-level components (e.g., external APIs or tool integrations) the AI models has access to during evaluation must be specified in advance and be homogeneous across models. The evaluation protocol should specify and explain the role of these additional components and how the contributions of these components are measured.
- **Data contamination and sandbagging:** Test datasets must be monitored for contamination; repeated exposure or "leakage" of test items into training can artificially inflate model performance. Evaluators should routinely update datasets or make statistical adjustments to mitigate these risks. Protocols should be designed to prevent deliberate "sandbagging" of benchmarks (e.g., by ensuring that test sets are administered in controlled environments with restricted access)
- **Robustness to perturbations:** The test environment should include controlled perturbations (e.g., input paraphrasing, controlled noise injection, or minor formatting variations) in addition to canonical test conditions.

D Domains

We report here the definitions of the domains we use in our framework. These definitions have been adapted from [18] and [20] for simplicity and suitability to consideration for GPAI models outside of the workplace:

- **AS: Attention and Scan:** The ability to focus on relevant information in a stream of data and to find items that meet certain criteria.
- **CE: Comprehension and compositional expression:** The ability to understand and extract meaning from natural language or other semantic representations, and to generate and express ideas. Subdomains:
  - **CEc: Verbal Comprehension:** Understand text, stories or the semantic content of other representations of ideas in different formats or modalities.
  - **CEe: Verbal Expression:** Generate and articulate ideas, stories, or semantic content in different formats or modalities.
- **CL: Conceptualisation, learning and abstraction:** The ability to generalise from examples, to learn from instructions or demonstrations, or to accumulate knowledge at different levels of abstraction.
- **QL: Quantitative and logical reasoning:** The ability to represent quantitative and logical information and infer new information to solve problems, including probabilities and counterfactuals. Subdomains:
  - **QLl: Logical Reasoning:** Match and apply rules, procedures, algorithms or systematic steps to premises to solve problems, derive conclusions and make decisions.
  - **QLq: Quantitative Reasoning:** Work with and reason about quantities, numbers, and numerical relationships.

E GPAI model categorisation framework: step-by-step protocol

When a new AI model is considered, the following protocol can be used to categorise it as a GPAI model:

1. **Clearly identify and specify the subject to be evaluated and the system-level components to be considered.** This means we need to delineate what is and is not considered as part of the evaluation. This involves identifying the specific model to be evaluated, but also the requirements to appropriately test the model (e.g., modalities, Section 3.2.1). Further, we need to consider the system-level components (Section 2.2) that may affect generality and capability (and ultimately categorisation).

2. **Exclude the model if it is unable to receive instructions for flexible tasks.** These models are not capable of being used across a wide variety of tasks capably enough to be considered GPAI models. Note that providing instructions does not need to be via textual prompts but could also be provided in other modalities (e.g., audio) or by providing few-shot demonstrations.
3. **For each of the four identified cognitive domains (Section 3.1), apply the set of relevant tests to the AI model.** Currently, the ADeLe battery [21] can be used for models receiving text as input, but this can be expanded on and updated over time. These tests need to be presented in the appropriate modality for the model (Section 3.2.1). If multiple modalities are available, consider the modality under which the model demonstrated strongest competence, measured as discussed in the point below.
4. **For each domain, analyse the results and obtain a domain-level capability score.** We propose to follow the ADeLe methodology outlined in Section 3.3: for every relevant domain, plot the subject's accuracy according to the human-normed demand level of the instances. Then identify the demand level where success probability is 0.5 and convert this to the human-normed scale (obtained as discussed in Section 3.3). An example of the application of this method is provided in Section F.
5. **Combine the domain-level scores in a single measure of "generality", using one of the approaches discussed in Section 3.4.** This measure will operationalise competence over a "wide-range" of tasks. An example of how this can be done with different choices of aggregation metrics and thresholds is given in Section 4.1.

F Analysing LLM performance with respect to human difficulty scores

In this section we provide an illustration of human norming as described in Section 3.3[8]. Figure 3 shows the results where the x-axis orders examples for each task in terms of human success rate (from 100 to 0, in four bins), and for each bin the performance of different models. With this we can compare the results across very different tasks which otherwise would be incomparable.

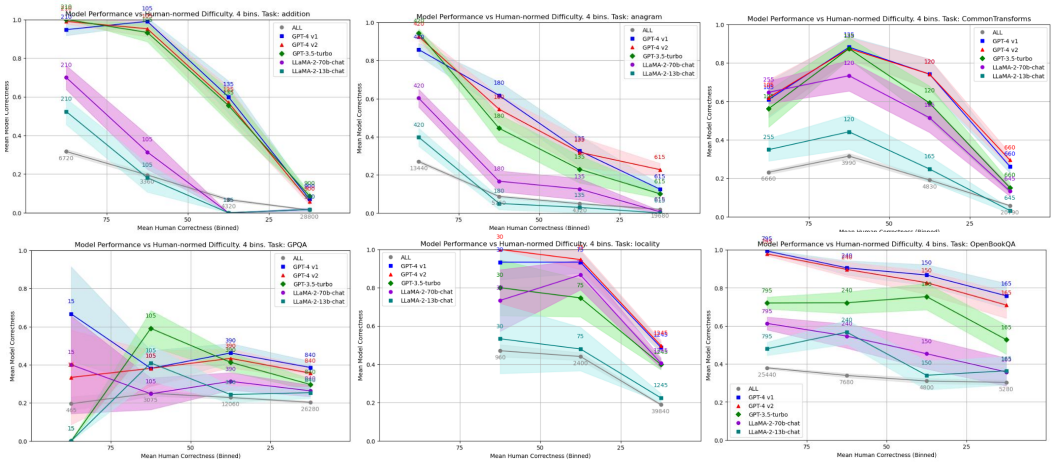


Figure 3: Results for Addition, anagram, CommonTransforms, GPQA [15], locality, OpenBookQA [14] for several models. The x-axis locates each example in bins depending on the percentage of human success.

G Correlation of capability levels with model size/compute

Here we analyse how a model's measured capabilities (its domain-wise capability levels from the ADeLe battery) correlate with the resources used to train that model, in particular, the number of parameters and total training FLOP.

[8]Data taken from `https://github.com/wschella/llm-reliability` and associated with [21].

Intuitively, larger models (and those trained on more data/computation) are expected to achieve higher capability levels. Indeed, previous research using ADeLe has observed clear scaling trends, with newer, larger models achieving higher capabilities in almost all dimensions than older, smaller models (see Figures 4 and 5).

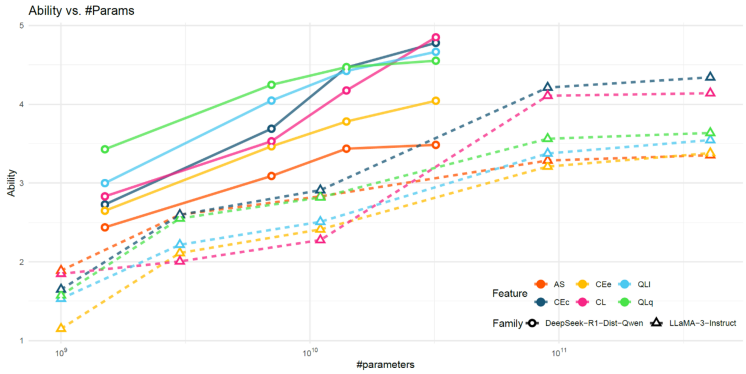


Figure 4: The scaling curves (number of parameters) of actual abilities for LLaMA and DK-R1-Distilled-Qwen families across four broad demands: Attention & Scan (AS); Comprenhension and Expression (with Verbal Comprehension (CEc) & Verbal Expression (CEe) as specific dimensions); Conceptualisation, Learning & Abstraction (CL); and Quantitative & Logical Reasoning (with Logical Reasoning (QLl) and Quantitative Reasoning (QLq) as specific dimensions). Data from [20].

Traditional performance scaling analyses (such as the one shown in Figures 6 and 7, which aggregates results across 20 benchmarks from ADeLe) are prone to saturation effects. These arise not only because the y-axis is constrained (e.g., accuracy is capped at 100%), but also because there may be some abstruse or even wrongly labelled questions that make the percentages never reach 100%. For the most powerful models, the composite performance scores flatten across many benchmarks, making it difficult to interpret incremental improvements as model size increases. This saturation can mask subtle but important gains in specific cognitive abilities.

In contrast, capability scaling curves (Figures 4 and 5), based on ratio scale measurements derived from ADeLe, remain sensitive across the full range of model sizes. They are not affected by benchmark saturation (as benchmarks can be swapped while the scale remains applicable) and show clear trends even for state-of-the-art systems, e.g., while larger models still yield better performance, the rate of ability growth slows significantly beyond a certain size.

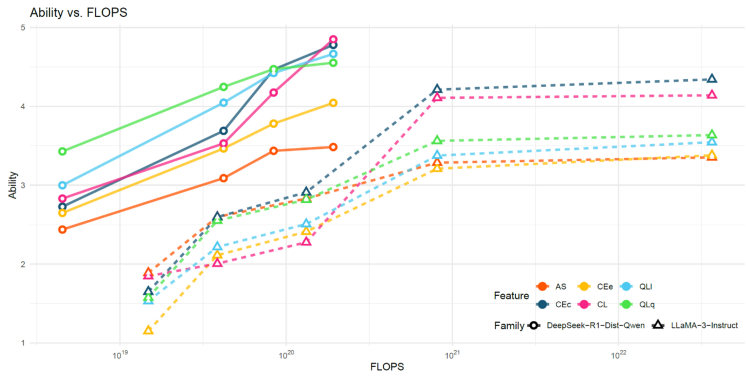


Figure 5: The scaling curves (FLOP) of actual abilities for LLaMA and DK-R1-Distilled-Qwen families across four broad demands (as in Figure 4). Data from [20]. Training compute estimates based on available data and scaling laws (FLOP $\approx 6 \times N \times D$, where N = parameters, D = tokens trained on). For models lacking explicit details, estimates are derived from comparable architectures or official disclosures[9].

[9]LLaMA-3.1 and 3.2 models: `https://build.nvidia.com/meta/llama-3.2-3b-instruct/modelcard`; `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`;

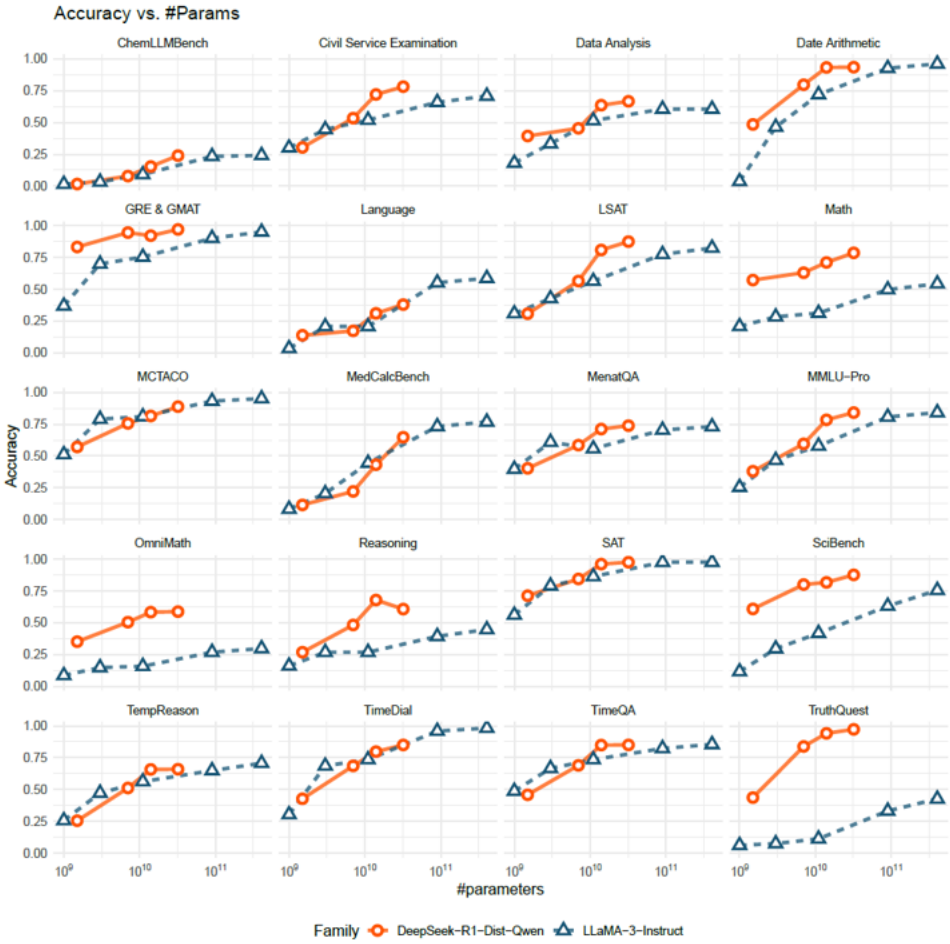


Figure 6: The scaling curves (number of parameters) of performance for LLaMA and DK-R1-Distilled-Qwen families across 20 different AI benchmarks. From [21].

https://ai.meta.com/blog/meta-llama-3-1/; DK-R1-Distill models: https://huggingface.co/AXERA-TECH/DeepSeek-R1-Distill-Qwen-1.5B; https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B; https://huggingface.co/RedHatAI/DeepSeek-R1-Distill-Qwen-7B-quantized.w8a8

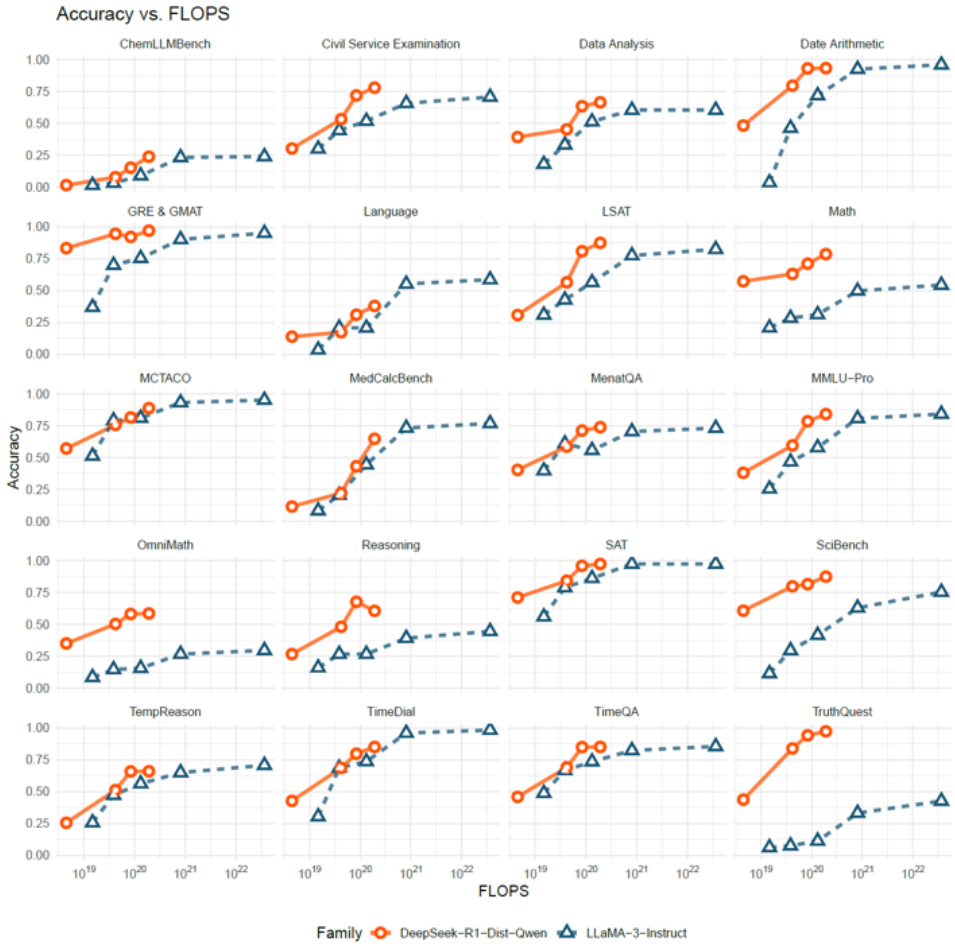


Figure 7: The scaling curves (FLOP) of performance for LLaMA and DK-R1-Distilled-Qwen families across 20 different AI benchmarks. From [20].