

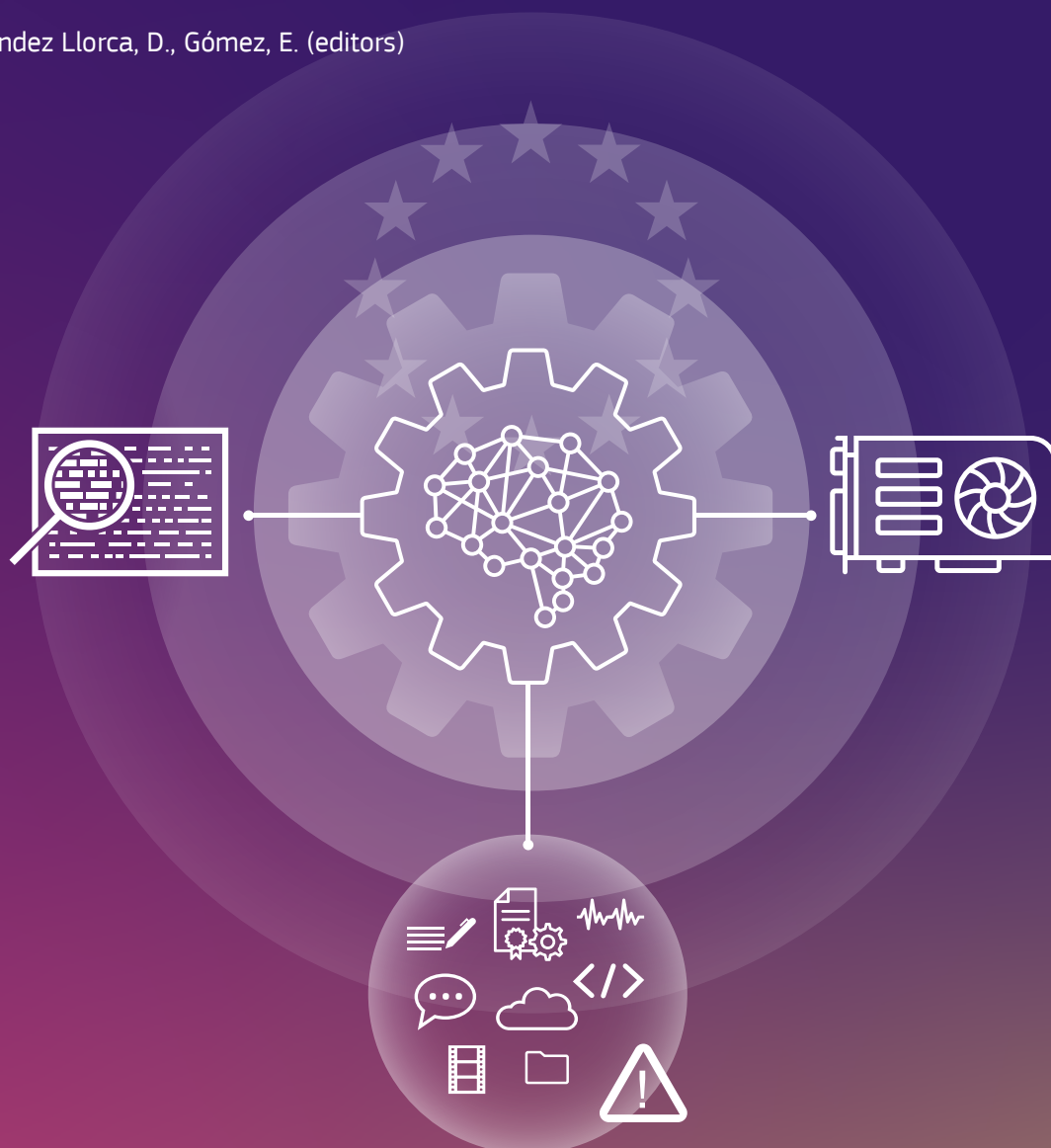
A Framework for General-Purpose AI Model Categorisation

Collection of External Scientific Studies on General-Purpose AI Models under the EU AI Act

Burden, J., Pacchiardi, L., Martínez-Plumed, F., Hernández-Orallo, J.

Fernández Llorca, D., Gómez, E. (editors)

2025



This publication is an External Study report prepared for the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The contents of this publication do not necessarily reflect the position or opinion of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Contact Information

Name: David Fernández Llorca

Address: European Commission, Joint Research Centre (JRC) Edificio Expo, c/Inca Garcilaso, 3, 41092 Seville - Spain

Email: david.fernandez-llorca@ec.europa.eu

The Joint Research Centre: EU Science Hub

<https://joint-research-centre.ec.europa.eu>

JRC143256

PDF ISBN 978-92-68-31431-9 doi:10.2760/5330387 KJ-01-25-459-EN-N

Luxembourg: Publications Office of the European Union, 2025

© European Union, 2025



The reuse policy of the European Commission documents is implemented by the Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Unless otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.

For any use or reproduction of photos or other material that is not owned by the European Union permission must be sought directly from the copyright holders.

How to cite this report: Burden, J., Pacchiardi, L., Martínez-Plumed, F., Hernández-Orallo, J. *A Framework for General-Purpose AI Model Categorisation*, Fernández Llorca, D., Gómez, E. (editors), Publications Office of the European Union, Luxembourg, 2025, <https://data.europa.eu/doi/10.2760/5330387>, JRC143256.

Contents

Abstract.....	2
Acknowledgements.....	3
Note from the Editors.....	4
Executive summary.....	6
1 Introduction.....	7
2 Background: definitions and considerations.....	8
2.1 AI model and AI system.....	8
2.2 Approach development considerations.....	9
2.3 System-level considerations for GPAI model categorisation.....	9
3 Operationalising the definition of a GPAI model.....	11
3.1 Identifying cognitive domains.....	11
3.2 Testing Each Domain.....	14
3.2.1 Domains and modalities.....	14
3.2.2 Annotating Demands and Measuring Capabilities.....	15
3.3 Competently performing in a domain.....	17
3.3.1 Preliminary considerations.....	17
3.3.2 Practicalities and potential issues with human baselines.....	17
3.3.3 Testing conditions.....	19
3.4 Wide range/generalality.....	19
3.5 Putting it all together.....	20
4 Empirical illustration of the proposed approach.....	22
4.1 How to analyse LLM performance with respect to human difficulty scores.....	22
4.2 Correlation of capability levels with model size/compute.....	22
4.3 Sensitivity analysis of classification thresholds and averages.....	23
4.3.1 Effect of aggregation function.....	24
4.3.2 Effect of threshold value.....	26
4.3.3 Domain pass/fail policy.....	27
5 Conclusions.....	32
References.....	33
List of abbreviations and definitions.....	34
List of figures.....	35
List of tables.....	36

Abstract

This report proposes a framework for categorising AI models as General-Purpose AI (GPAI) models as defined in the EU AI Act, based on their capabilities and generality. It breaks down the core components of the GPAI model definition into measurable elements, focusing on four primary cognitive domains: (1) Attention and Search, (2) Comprehension and Compositional Expression, (3) Conceptualisation, Learning and Abstraction, and (4) Quantitative and Logical Reasoning. The report suggests using the Annotated Demand Levels (ADeLe) procedure to evaluate AI models' capabilities in these domains, and provides 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.

Acknowledgements

The authors thank the Joint Research Centre and the AI Office team for valuable feedback and suggestions on the draft.

The editors would like to thank the JRC colleagues who have helped us develop the work resulting in this collection, the AI Office for its inputs, as well as those who have kindly agreed to review the drafts and the final external study reports.

Authors

Burden, John

Pacchiardi, Lorenzo

Martínez-Plumed, Fernando

Hernández-Orallo, José

Editors

Fernández Llorca, David

Gómez, Emilia

Note from the Editors

The EU AI Act entered into force on 1 August 2024, with the aim of promoting innovation in and uptake of AI in the Union, while ensuring a high level of protection of health, safety and fundamental rights, including democracy and the rule of law. Chapter V of the AI Act outlines obligations for the providers of general-purpose AI (GPAI) models, which are AI models *"trained with a large amount of data using self-supervision at scale, 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"*. Moreover, the chapter specifies additional obligations for the most advanced GPAI models, those that pose systemic risks, which are classified as such according to criteria established in Article 51 and Annex XIII. From 2 August 2025, the obligations for providers of GPAI models and GPAI models with systemic risk enter into application.

The European Commission's Joint Research Centre (JRC) has been providing scientific support throughout the legislative process of the AI Act since 2020. After the Council and Parliament reached a final agreement in December 2023, the JRC initiated an internal study, which included two external experts, focusing on the technical aspects of Chapter V that likely required further clarification. This study generated an internal report titled "General Purpose AI Models under the AI Act" (Hernández-Orallo et al., 2024), which provided preliminary insights into compute, generality, capabilities, and systemic risks. One of the clear conclusions of this preliminary study was that further scientific work was necessary.

Between September 2024 and June 2025, the JRC setup and managed, in close collaboration with the EU AI Office, a pool of 15 external experts with diverse expertise and backgrounds. This expert pool produced further technical scientific input on key aspects of Chapter V, through the development of methodologies for categorising AI models as GPAI models and for classifying GPAI models as GPAI models with systemic risk, to inform implementation of the EU AI Act. The experts also provided input on the recently published Commission guidelines on the scope of obligations for providers of GPAI models (European Commission, 2025), as part of the public multi-stakeholder consultation. The primary outcome of this expert pool is this **Collection of External Scientific Studies on General-Purpose AI Models under the EU AI Act**, which comprises a total of six external scientific study reports.

Although more documents may be added to the collection in the future, as of the writing of this editorial, the titles of the external reports included in the collection are:

- Training Compute Thresholds - Key Considerations for the EU AI Act
- A Framework for General-Purpose AI Model Categorisation
- A Framework to Categorise Modified General-Purpose AI Models as New Models Based on Behavioural Changes
- A Proposal to Identify High Impact Capabilities of General-Purpose AI Models
- The Role of AI Safety Benchmarks in Evaluating Systemic Risks in General-Purpose AI Model
- General-Purpose AI Model Reach as Criterion for Systemic Risk

The overall objective of this collection is twofold. On the one hand, it aims to contribute to broadening the understanding and discussion of the technical and scientific issues related to GPAI models and the identification of systemic risks. On the other hand, it seeks to provide a solid scientific basis for informing the implementation of Chapter V of the EU AI Act, which has recently entered into application. It is clear that we are dealing with complex issues, where a clear scientific consensus has yet to be established, and which require a certain degree of flexibility. Nevertheless, this is part of the

necessary effort to promote innovation and the uptake of AI, while ensuring protection for human health, safety, and fundamental rights in Europe.

These external scientific studies cover aspects regarding the presumption of having high impact capabilities based on cumulative amount of computation used for training (Article 51(2)), notification conditions (Article 52), the definition of a GPAI model (Article 3(63)), and considerations for GPAI models being classified as GPAI models with systemic risk based on capability benchmarks, safety benchmarks and reach (Article 51(1) and Annex XIII).

These studies reflect the outcome of the scientific and technical analysis of a series of external experts to the Commission. In some cases, they present a state-of-the-art review, while in others, they propose methodologies based on solid scientific evidence, while acknowledging significant uncertainty, as many of these problems still lack a widely accepted solution. The content, analysis, recommendations, and suggestions should then not be interpreted in any way as the position of the Commission, nor of the JRC editors in particular, but rather as the opinion of the authors.

Executive summary

The EU AI Act introduces the concept of general-purpose AI (GPAI) models. This report is guided by the following question:

- When should an AI model be categorised as a GPAI model on the basis of capabilities and generality?

Our contribution is to propose an operationalisation of the concept of a GPAI model that balances scientific rigour with practical feasibility, ensuring accurate reflection of evolving AI capabilities.

The operationalisation breaks down an AI model's core capabilities into key cognitive domains (such as attention and search, comprehension and compositional expression, conceptualisation and abstraction, and quantitative and logical reasoning). Then, it relies on instance-level analysis as well as human performance comparisons to derive meaningful performance metrics. These metrics are then combined across domains to lead to a final metric which could be used as part of an assessment of whether an AI model should be considered to be a GPAI model or not.

This report presents a potential methodology, without delving into setting numerical thresholds nor into the extent to which system-level components should be considered. These questions will need to be separately addressed to align the proposed methodology with legal, technical and regulatory developments. Moreover, answers to these questions should be updated over time to account for the evolving technical and regulatory landscape.

1 Introduction

This report presents a method for operationalising the definition of General-Purpose AI (GPAI) models as introduced in the EU AI Act, building on the preliminary framework presented in (Hernández-Orallo et al., 2024). Our goal is to ensure this operationalisation is both scientifically rigorous and practically implementable, striking a balance between fidelity to cognitive theory and feasibility for model developers. To achieve this, we draw on foundational work from cognitive psychology and psychometrics.

We propose an explicit set of cognitive capabilities that an AI model should demonstrate to a suitable level to be considered “general-purpose.” This selection is 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 (Carroll, 1993, Keith and Reynolds, 2010) along with more recent adaptations for the AI domain (Tolan et al., 2021). 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 or other entities who may need to assess the relevant models against these capabilities, we reduce this to a core set of four domains we believe to be most pertinent: (1) Attention and Search, (2) Comprehension and Compositional Expression, (3) Conceptualisation, Learning and Abstraction, and (4) Quantitative and Logical Reasoning. We outline a concrete procedure for measuring these capabilities using existing AI benchmarks and model evaluations.

We then examine how these measured abilities could be used to design a binary GPAI model categorisation. We consider various decision rules and aggregation functions, such as arithmetic, geometric, and harmonic means, and explore the implications of these choices. While we do not prescribe a fixed aggregation function and threshold ourselves, we illustrate how different choices of thresholds affect the classification and argue that such choices must remain context-sensitive and be subject to ongoing revision.

This report offers what we see as a gold standard assessment procedure: one that allows capabilities to be measured accurately, and supports principled decision-making about whether a model should be categorised as a GPAI model.

2 Background: definitions and considerations

We begin with background and definitions that are useful to keep in mind when deciding how to categorise AI models as GPAI models.

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’.***

A number of recitals of the EU AI Act clarify how the concept of a 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.

2.1 AI model and AI system

The definition of an AI model is not given in the AI Act. The glossary of the preliminary report (Hernández-Orallo et al., 2024) gives the following definition of an 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.*

Section 2.1 of the preliminary report (Hernández-Orallo et al., 2024) further specifies the definition in the following way:

- an ‘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.

By contrast, the EU AI Act (Article 3(1)) defines the related notion of 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.’

The European Commission has recently released guidance for the interpretation of the definition of AI systems (European Commission, 2024).

2.2 Approach development considerations

The following aspects were considered when drafting our approach and criteria for guiding the categorisation of AI models as GPAI models:

- The approach must **be aligned with the definition of a GPAI model** as outlined in Article 3(63) of the EU AI Act and follow the recitals as closely as possible to ensure it is technically sound and consistent.
- The approach must enable **easy determination of whether a given AI model qualifies as a GPAI model** or not. This implies clear and objective criteria that are straightforward to apply.
- The approach should be **difficult to circumvent** or manipulate to ensure that it is effective in practice.
- The approach should be designed to remain **relevant for approximately the next two years**, accounting for the current state of AI technology, and any anticipated developments.

2.3 System-level considerations for GPAI model categorisation

The preliminary report (Hernández-Orallo et al., 2024) argues that: *“although the focus of the analysis should be at the model level (policy requirement), the assessment of capabilities, generality, systemic risks, and safety requirements may benefit from the evaluation of the standalone model in conjunction with system-level components”*. This is motivated by the fact that the definition of a GPAI model includes the requirement of being *“capable of competently performing a wide range of different tasks”* and having the potential to *“be integrated into a variety of downstream systems or applications”*. Therefore, the generality of a model can be evaluated in the context of the different systems in which it can be embedded.

Thus, in determining whether an AI model qualifies as a GPAI model, it is important to look beyond the model itself and consider the system-level components that influence its capabilities.

In practice, system-level components affect the model performance (and therefore the determination of its generality) in the following ways:

- The way **users interact** with an AI model, by means of user interfaces (UI) and user experience (UX), can have a significant impact on its performance. For example, a LLM without

a proper user interface is essentially an "inert matrix" because the user cannot give instructions. Conversely, well-designed interfaces and API calls that allow for multiple interactions and growing context windows can improve an LLM's ability to perform tasks by allowing users to correct errors and guide the model.

- The **availability of tools** can dramatically extend the capabilities of an AI model. For example, by teaching a model how to use external tools, implementing prompts to guide its reasoning, or using scaffolding to structure its thinking, the model can generate and evaluate multiple candidate solutions and even benefit from data enhancements (Davidson et al., 2023). In practice, giving a model access to a web search, for example, allows it to gather information that it would not otherwise have, enabling it to perform tasks that it could not before. Similarly, tools such as calculators can help models overcome limitations such as basic arithmetic, enabling them to perform more complex mathematical reasoning.

It could therefore be important for policy-makers to establish the set of system-level components to which an AI model is given access when assessing its capabilities and generality.

3 Operationalising the definition of a GPAI model

To effectively categorise AI models as GPAI models, it is necessary to break down the core components of the GPAI definition and make them measurable. According to the definition in Article 3(63) of the AI Act, a GPAI model:

- **displays significant generality,**
- **is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market,**
- can be integrated into a variety of downstream systems or applications,
- 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 2.3), while the fourth is purely of legal nature. The document focuses primarily on the first two elements of the definition, and assumes for simplicity that when an AI model has these two elements it should be considered to be a GPAI model. However, note that there may be technical solutions (such as some form of cryptographic verification) to ensure that a model, which would be categorised as a GPAI model based on its capabilities (i.e., the first two parts), is used only in a specific system or application (such as inside a specific physical device or robot).

Building on the preliminary report (Hernández-Orallo et al., 2024), we suggest framing generality in terms of **abstract capability domains**. These broadly correspond to distinct constructs that are required for some kinds of cognitive tasks but not others. At a high level, we take the view that consistent competent performance across a wide range of domains entails that the system is sufficiently capable and general-purpose to be categorised as a GPAI model¹. To address this, four key questions are considered:

1. Which cognitive domains should be investigated?
2. What tests and methodology should we use to evaluate each of the considered domains?
3. What does it mean to perform consistently and competently in a particular domain?
4. How should performance across domains be combined to determine generality? In particular, how "wide" a range of domains is needed to demonstrate generality?

We start with the first question; identifying meaningful cognitive domains that we can use to determine GPAI-status.

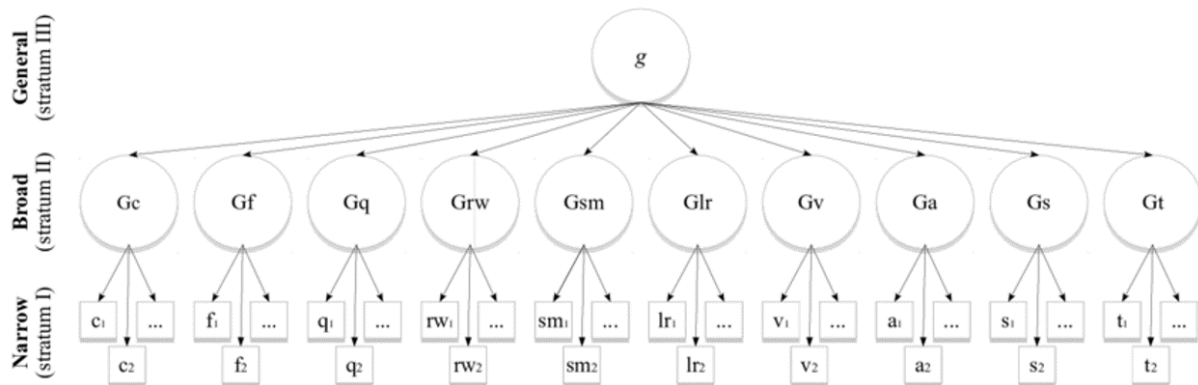
3.1 Identifying cognitive domains

The Cattell-Horn-Carroll (CHC) (Carroll, 1993, Keith and Reynolds, 2010) theory of intelligence has been a cornerstone of psychometrics and cognitive psychology for several decades. It is widely accepted as a comprehensive model because it organises cognitive abilities into a hierarchical structure, with a general intelligence factor (g) at the top, broad abilities in the middle and more specific abilities at the bottom. The most recent version of CHC theory includes 10 broad cognitive

¹Notice how the choice of what to interpret as "tasks" goes hand in hand with determining what "wide range" means, with the consequence that the same effect can be obtained by considering tasks=domains and tasks=specific problem types, as long as "wide range" is suitably specified.

abilities and over 70 narrow abilities (see Figure 1). Many widely used intelligence tests and assessment batteries are based on the CHC framework, giving it considerable empirical support and recognition in educational and clinical settings.

Figure 1: Cattell-Horn-Carroll's three stratum model. The broad abilities are Crystallised Intelligence (Gc), Fluid Intelligence (Gf), Quantitative Reasoning (Gq), Reading and Writing Ability (Grw), Short-Term Memory (Gsm), Long-Term Storage and Retrieval (Glr), Visual Processing (Gv), Auditory Processing (Ga), Processing Speed (Gs) and Decision/Reaction Time/Speed (Gt).



Source: Modified version from Tim bates, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=24564663>.

However, the theory is not without its critics. Some researchers argue that CHC relies too heavily on factor analysis and may oversimplify the nuances and dynamic nature of cognition. Alternative models, such as Gardner's multiple intelligences (Gardner and Hatch, 1989) or Sternberg's triarchic theory (Sternberg, 1985), have been proposed to capture aspects such as creativity, practical problem solving or emotional intelligence that the CHC model may not fully address. Moreover, as the theory originates from studies of human cognition, applying its framework to artificial intelligence systems requires a careful translation of these domains.

In this context, we use CHC theory primarily as an inspiration to identify and operationalise the core cognitive domains relevant for evaluating whether an AI model might be a GPAI model. In particular, expanding on the CHC, in (Tolan et al., 2021) the authors derive a list of 14 cognitive abilities (domains) that are relevant to AI capabilities. These are:

- Memory processes (MP)
- Sensorimotor interaction (SI)
- Visual processing (VP)
- Auditory processing (AP)
- **Attention and search (AS)**
- Planning and sequential decision-making and acting (PA)
- **Comprehension and compositional expression (CE)**
- Communication (CO)
- Emotion and self-control (EC)
- Navigation (NV)
- **Conceptualisation, learning and abstraction (CL)**

- **Quantitative and logical reasoning (QL)**
- Mind modelling and social interaction (MS)
- Metacognition and confidence assessment (MC)

These domains cover a wide range of capabilities and tasks. However, mastery of all these domains is not necessarily required for a model to be categorised as a GPAI model. For instance, sensorimotor interaction can be incredibly important if needing to interact physically in the real world yet is not needed to perform many cognitively demanding tasks that we would associate with GPAI models. We have highlighted in bold the four domains (and their particular subdomains) which we think are most pertinent for categorising AI models as GPAI models, though a different selection may be defensible: *Attention and Search (AS)*, *Comprehension and Compositional expression (CE)*, *Conceptualisation, learning, and abstraction (CL)*, and *Quantitative and Logical Reasoning (QL)*.

For ease of reference, we report here the definitions of these concepts from (Tolan et al., 2021) and (Zhou et al., 2025). These definitions have been rephrased for simplicity and suitability to consideration for GPAI models outside of the workplace:

- **AS: Attention and search:** 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.

These domains and subdomains are not exhaustive, but they represent a minimum set of capabilities which we think are necessary for a model to be considered a GPAI model. It is possible that other capabilities could emerge from the presence of these domains, albeit in a limited way (e.g., modules without a directly trained vision system were able to create and understand ASCII art).

Notice that the capabilities identified above mostly refer to how a model parses and analyses incoming information and generates new information. However, the definition of a GPAI model also includes the ability to act and to use capabilities to achieve a goal. The operationalisation of this aspect is achieved through behavioural evaluation using benchmarks to ensure that a model is able to complete the tasks set for it.

3.2 Testing Each Domain

Now that we have identified the four primary domains we are interested in (noting that other reasonable choices could have been made), AI models will need to be tested on these domains. To do so, we adopt the technique outlined in (Zhou et al., 2025). Here, we annotate individual task-instances according to the demands they pose on the various domains and then analyse the performance of AI models at the level of individual instances to identify how the various demands impact performance. This does take some initial effort to calibrate demand levels and construct rubrics to annotate demands. However, once these are set up the process of annotating the demands of new instances is automatable, and the rubrics can be reused for new evaluations. We describe this method in more detail in Section 3.2.2.

Nevertheless, when selecting instances to evaluate a particular domain on, we must first consider what *modality* an AI model operates in, as discussed in the following section.

3.2.1 Domains and modalities

AI models can have different input and output modalities. These necessarily impact the set of tasks on which an AI model can be applied. At the same time, models with different output modalities can show proficiency in the same domain (e.g., models outputting text and audio are both in principle able to demonstrate reasoning abilities). In practice, this means that, when testing models for their competence in each domain, it is necessary to instantiate a test specific to the output (and input) modalities of the model². If a model can employ multiple modalities, it can be evaluated using tests specific to each modality; in this case, competence in a single modality is sufficient to claim the model is competent in that domain.

Moreover, some constrained output modalities do not plausibly allow the model to show competence in any domain. This is easy to demonstrate in extremes: a model that can only output a single bit may not be able to express logical reasoning outside of a multiple-choice context.

Below is a non-exhaustive list of modalities:

- Text (including code)
- Images
- Audio (including speech)
- Sensorimotor
- Tabular data

This list captures most modalities in which current models generally process stimuli. Of course, modalities can be mixed-and-matched as desired (e.g., a model could take as input both text and an image). At the time of writing, most commercially available frontier models take text, images, and audio as both input and output modalities. Yet this could change with advances in robotics (where sensorimotor modalities become a large factor) or novel modalities (such as olfactory) being utilised.

AI models process all modalities by converting them into binary strings. Therefore, there is no limit to the types of data they can process, as long as the data can be represented digitally.

The focus should be on matching the modality of evaluation to the modality for which the model was developed: it is possible to subject a text-only AI model to a test developed for audio models, as both

²For instance, tests testing common-sense reasoning could be instantiated both in written and spoken form, which can be respectively applied to text-to-text (e.g. LLMs) and audio-to-audio (e.g. Alexa) models

text and audio are converted to binary strings before being inputted to the model. However, with all probability, the text-only AI model will not be able to interpret the binary representation of the audio string, as it was not developed to do so.

3.2.2 Annotating Demands and Measuring Capabilities

Introduction to the Annotated Demand Levels (ADeLe) procedure

At the core of the ADeLe³ procedure for annotating task-instances⁴ 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 Search (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.

The ADeLe process for annotating demands and evaluating capabilities has two components: the "System Process" (SP) and the "Task Process" (TP). The SP is performed once per AI model, whereas the TP is performed once for each new task or benchmark. We detail these processes in the next subsections.

Measuring Capabilities

In the SP, an AI model is tested on a test battery with known demands. The responses from the AI model can be "sliced" across individual demands, giving rise to so-called subject-characteristic curves that detail how an AI model's response varies with the level of a particular demand. In practice, this is done by considering all instances whose demand for the considered domain is at a given level and computing the average score of the AI model. Moreover, the ADeLe procedure employs 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 (Thurstone, 1937). Given the simultaneous annotation of tasks with different demands, we get a demand profile per instance and per benchmark. A single pass through the questions of a benchmark by an LLM can measure all capabilities simultaneously. Both profiles can be represented by a polar plot, as we see in Figure 2 taken from (Zhou et al., 2025).

Annotating Demands

We recommend using the ADeLe battery for the SP step, as it includes modern benchmarks and covers a wide range of key capabilities, and its automated annotation has been independently validated by human reviewers through inter-rater analysis and the Delphi method (Linstone, 1975). However, the same approach could be used to annotate a new dataset with more refinement, or to cover specific areas.

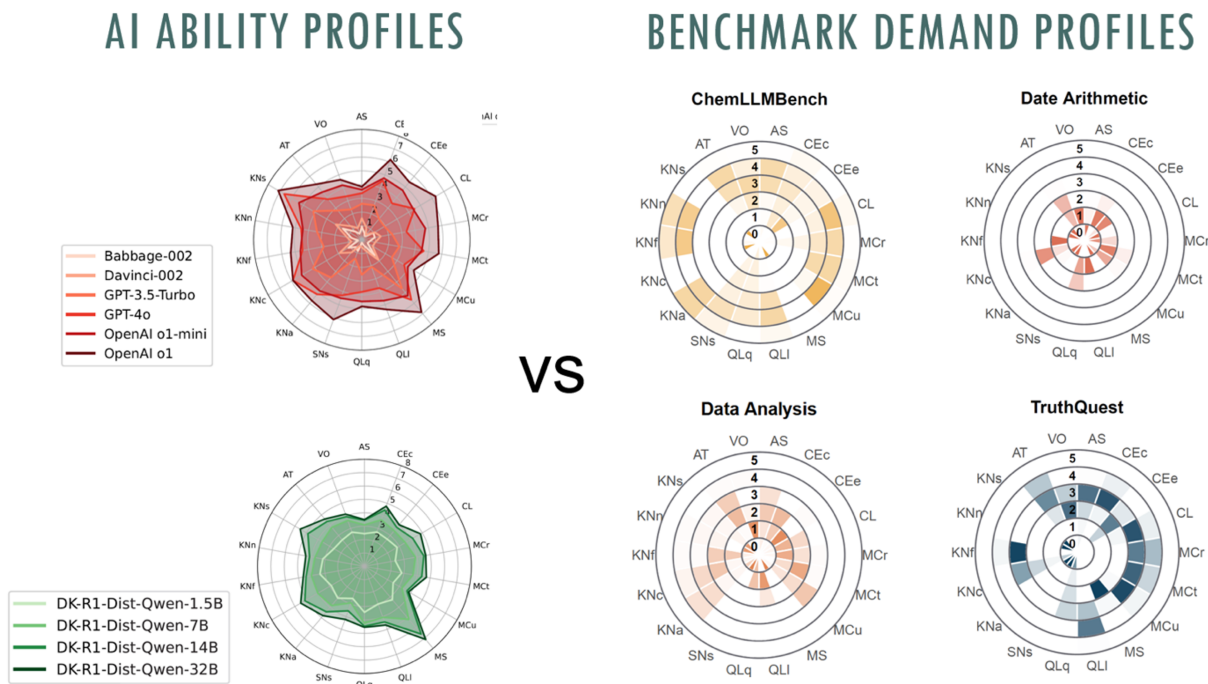
Instances are annotated with demands on a ratio scale. This means the scale contains an *absolute zero*, where demands are not present at all. Ratio scales also require the differences between levels to be consistent across the scale. Ratio scales are important as they ensure comparability across different capabilities. ADeLe adopts the convention that doubling the demand halves the log odds of success.

To practically achieve this for a large dataset, rather than having humans assign the demands manually, it is instead prudent to take advantage of existing LLMs' capability to robustly apply

³ADeLe v1.0: A battery for AI Evaluation with explanatory and predictive power. <https://kinds-of-intelligence-cfi.github.io/ADELE/>

⁴A task-instance refers to a particular example, instantiation, item or question of a task.

Figure 2: Ability profiles (radial plots) for OpenAI and DK LLMs using the ADeLe battery 1.0 (left). Demand profiles for a selection of four benchmarks from the ADeLe battery v.1.0 (right). The colour intensity illustrates the frequency or number of issues for each domain and level of difficulty.



Source: elaborated from (Zhou et al., 2025).

rubrics. By carefully developing precise rubrics identifying what the demands of a question are, an LLM can assign those demands automatically. This greatly reduces the burden of applying such a method. The ADeLe dataset takes over 16,000 task instances from high-quality benchmarks and annotates them using such rubrics.

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 $1/10^{(-l)}$ humans being able to solve the task).

Anticipating Performance

In the TP, the learnt capabilities of the model can be used to anticipate performance on new unseen tasks. The same rubrics used to annotate the original dataset (either ADeLe or a custom dataset) are used to annotate the new instances. Then, using a black-box assessor (Hernández-Orallo et al., 2022)—a type of predictive model used to predict instance-level performance based on instance-features—that has been trained on the AI model and SP dataset, performance can be anticipated on the new tasks. With a large number of training instances, and carefully developed rubrics and carefully annotates demands, the predictive power of this approach can be very high. For example, ADeLe routinely achieved AUROC scores of 0.85 or higher.

Putting it all Together

Overall, the proposed ADeLe methodology provides two vitally important indicators that can be used to assess whether an AI model should be categorised as a GPAI model. The first is the capability profile of the model — the collection of capability values indicating the demand where the model will attain at least 50% success rate. The second is the predictive model of performance in new tasks. We anticipate that the first of these will likely be the primary indicator guiding GPAI categorisation, but the second validates its use. In the next section we look at what performing competently looks like.

3.3 Competently performing in a domain

Now that we have identified a set of domains and a methodology for assigning 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 capability scores into a decision. In practice, we suggest normalising the scales used in the ADeLe approach by using a human population to measure the model's capability on this human-normed scale. Additionally, we need to identify the test conditions under which AI models are subject to assessment on the dataset.

3.3.1 Preliminary considerations

We reject the idea that competently performing can be effectively understood via simplistic aggregated results (e.g., across a benchmark). Simply averaging aggregate performance, such as accuracy, across benchmarks is problematic because, among other issues (Eriksson et al., 2025), benchmarks vary in difficulty. Direct aggregation can lead to biased results. In addition, benchmarks with different random guess accuracy rates (e.g. 45% vs. 33%) are not directly comparable. These issues are the primary motivations we have for recommending the ADeLe methodology.

Instead, performance should be measured against a common **baseline** across benchmarks. Even with a common baseline, however, using it to calculate the difference or ratio of model accuracy can be misleading. For example, a 10% improvement when the baseline is 90% is more significant than the same improvement when the baseline is 30%. In essence, simply considering the performance in terms of improvement over a single baseline does not provide any insight on how significant that improvement was.

One possible baseline could be the average accuracy of a human population. Since the task-instances in the ADeLe battery are taken from existing datasets, it may be possible to use existing results to find human performance on these questions. Unfortunately, this is not likely as not all datasets directly compare against a human baseline or do not report the instance-level results. Alternative baselines could also be used, such as a population of AI models or a single AI model sampled multiple times (with high temperature, if applicable). However, these would be less representative of general-purpose intelligence and would likely make for less accurate model categorisations.

To establish a human baseline, several aspects still require further consideration. Factors such as the differences in the populations studied (e.g. experts vs. non-experts) in the papers that introduce the various datasets can make aggregation difficult if we rely on the reported numbers. Further, relying on mean, rather than median, performance can skew results due to the influence of outliers.

Finally, the median (or mean) accuracy of human population does not provide information on how the human performance is spread (something that is very desirable for us to know): on two different baselines, the average human performance may be very close but the spread could be very different (e.g., on the first benchmark all humans get 50%, while on the latter humans get evenly distributed accuracies between 0 to 100%). Therefore, comparing to a single human baseline is uninformative as, if a LLM achieves 75% of human performance, this may be very poor in terms of comparison to the human population on one benchmark, but acceptable on another.

3.3.2 Practicalities and potential issues with human baselines

Implementing human norming for ADeLe's demand level scales is logistically challenging. The core idea is to calibrate each demand level (e.g. level l) to a corresponding human probability of success. For example, that demand l corresponds to 1 in 10^l human participants responding correctly (e.g. "level 3 \approx 1 in 1,000 people can solve"). Note that this does not require the creation of any new question, or altering them in any meaningful way. In practice, human data collection may be

performed using platforms such as Prolific⁵; it would be sufficient to do this on ~ 100 s instances per domain (e.g., with the ADeLe Light battery), as lower demand levels (where perhaps 1 in 10 or 1 in 100 people succeed) can be reasonably estimated with dozens of human responses per item. However, 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 on the order of 10^5 people, clearly infeasible in terms of recruitment and cost. This raises the concern that the upper end of the ADeLe scales cannot be directly normed without extensive human sampling.

Several mitigation strategies can address this. First, a **selective sampling** approach can focus human evaluation on mid-range difficulty items and extrapolate the tail probabilities using statistical models. For instance, one could fit a logistic or item-response curve to human performance data collected for moderate demand levels and then extrapolate to predict success rates at higher levels, rather than trying to directly measure vanishingly small success probabilities. Second, a **hybrid LLM-human annotation** strategy could leverage language models to preliminarily gauge which instances are extremely challenging, and only a sparse set of those highest-demand items would then be presented to human experts. Finally, one can 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 when interpreted carefully it can at least confirm that someone can solve the item. Basically, while humans remain the ultimate reference for "ground truth" difficulty, it is acceptable to rely partly on extrapolation and expert judgment at the upper end.

Still, a key challenge is to determine the appropriate human reference population. The selected group should ideally reflect a consistent level of technical knowledge, education and domain expertise relevant to the cognitive domain being assessed. Humans' performance can vary dramatically depending on these factors. For example, a task involving complex scientific reasoning may yield higher performance scores when assessed against participants with advanced education or technical backgrounds, whereas assessments of general language comprehension may be better served by a broader representative sample of the population. If these differences are not carefully controlled, the resulting baselines of performance may be inconsistent across experiments, making comparisons between benchmarks unreliable.

To address these issues, we can follow some strategies:

- Implement standardised pre-screening procedures to ensure that participants meet specific educational or technical criteria to maintain consistency across experiments. For instance, platforms such as Prolific allows for demographic pre-screening.
- In cases where it is difficult to assemble a completely homogeneous group, consider stratifying the participant pool based on performance levels or relevant demographic factors that account for variability within the human baseline.
- Use identical instructions, test environments and scoring protocols across all experiments to minimise extraneous sources of variation so human performance reflect true differences in ability rather than artefacts of the testing process.
- Record detailed information about human participants (e.g., educational background, technical experience, and self-reported ability) to provide a basis for future adjustments should the baseline need to be normalised.
- Use equating methods if we have some common items for the two human populations or samples.

⁵<https://www.prolific.com/>

Even with standardised human population through the above strategies, it may be that, for a considered domain, the difficulty distribution is narrow (e.g. most instances have ~60% human success rates in the considered population). This issue can be addressed by expanding or adjusting the dataset with additional instances.

3.3.3 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.

3.4 Wide range/generalizability

We now have four primary domains that can be used as part of assessing whether an AI model should be categorised as a GPAI model. Moreover, the procedure discussed in the previous section 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 interpret the results and categorise the model. The simplest way of doing this is to find a metric for the domain scores that combines them into a single score or set a threshold value for each score and a rule for interpreting different combinations of success.

Possible approaches include:

- An average over the human-normed capability scores on each of the domains. Once the average exceeds a certain threshold (e.g., a human-normed score of 50%) then the model is categorised as a GPAI model. Different mathematical averages can be used:
 - The *Arithmetic mean* if we want to allow majority-performance to lift-up / 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.
- Threshold values:
 - Identify a suitable threshold value for each domain (e.g., a human-normed capability of 50%) - then mark each domain as Pass vs. Fail. If there is a sufficient number of passes (e.g., 2+) then this indicates sufficient generality and capability and thus we categorise the model as a GPAI model.

With both approaches it is very easy to alter the threshold / scores needed for GPAI categorisation (e.g., if new guidance about when to categorise a model as a GPAI model is made into policy, or new understanding about generality becomes widely supported). However, in all cases it is non-trivial to account for small changes in performance near the threshold. These could potentially be gamed depending on the metric used.

In Section 4.3 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.

3.5 Putting it all together

With all of the above sections we can summarise the approach into an overall protocol that could be systematically followed to categorise a model as a GPAI model or not.

The overall protocol is the following:

1. **Identify subject to be evaluated.** 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). Further, we need to consider the system-level components that may affect generality and capability (and ultimately categorisation).
2. **Exclude subjects who are 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, apply the set of relevant tests to the AI model.** Currently we recommend the ADeLe battery but this can be expanded on and updated over time. These tests need to be presented in the appropriate modality for the model. If multiple modalities are available, consider the modality under which the model demonstrated strongest competence, measured as discussed in the point below.
 - (a) For each domain, analyse the results and obtain a domain-level score. We propose following the ADeLe methodology outlined earlier in the report: 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, report this as the domain capability. Further details of this method can be found in (Zhou et al., 2024) and an example is provided in Section 4.1.
4. **Combine the domain-level scores into a single measure of "generality".** 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 4.3.

4 Empirical illustration of the proposed approach

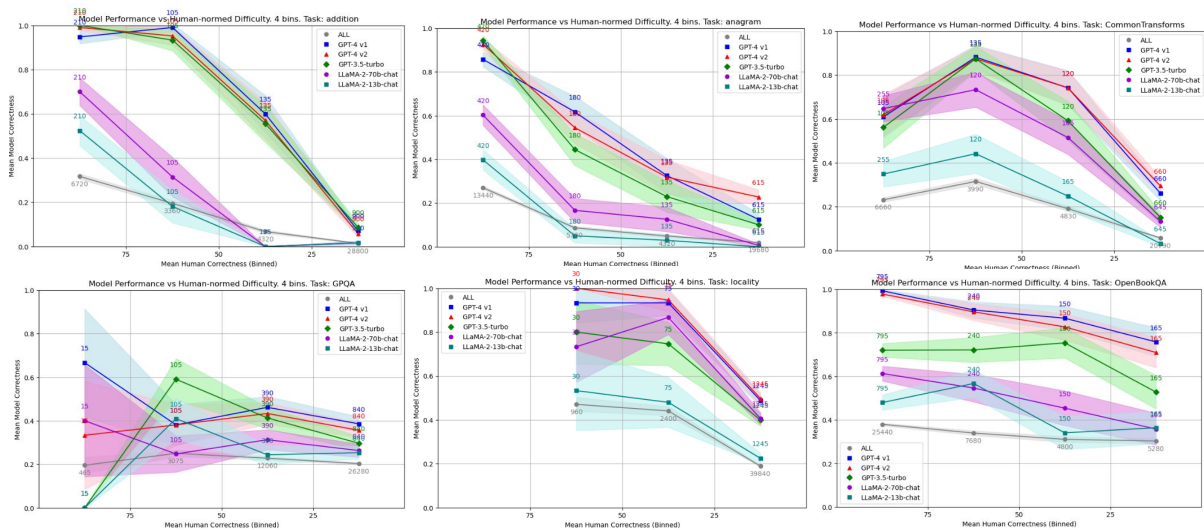
4.1 How to analyse LLM performance with respect to human difficulty scores

In this section we provide an illustration of the technique described in Section 3.2. The following two figures include results from real LLMs and real human performance.

The data is available at <https://github.com/wschella/llm-reliability> and is associated with (Zhou et al., 2024).

Figures 3 and 4 show 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 (QL), anagram (AS, CE), CommonTransforms (AS, CE), GPQA (AS, CE), locality (QL), OpenBookQA (AS, CE) for several models. The x-axis locates each example in bins depending on the percentage of human success.



Source: Own elaboration with data from (Zhou et al., 2025).

4.2 Correlation of capability levels with model size/compute

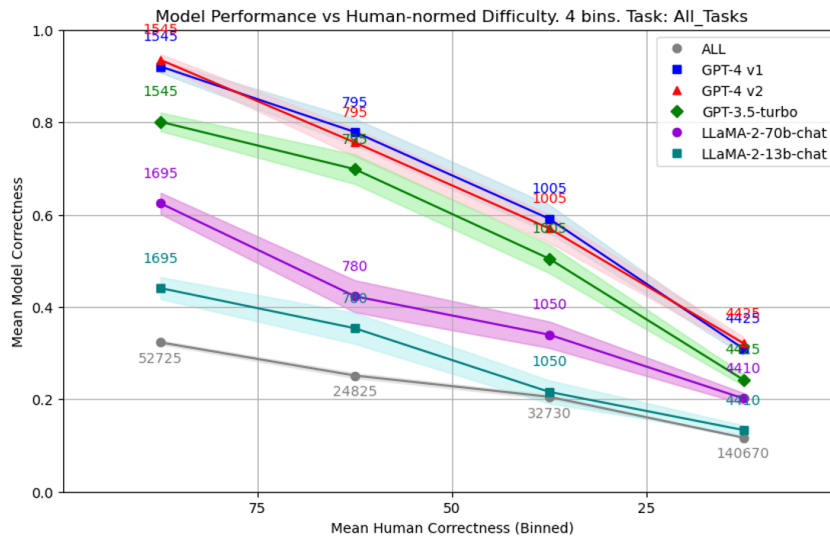
Here we analyse how a model's measured capabilities (its domain-wise capability levels from the ADeLe battery) correlate with the resources used to train that model, in particular, the number of parameters and total training FLOP.

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 5 and 6)⁶.

Traditional performance scaling analyses (such as the one shown in Figures 7 and 8, which aggregates results across 20 benchmarks from ADeLe) are prone to saturation effects. These arise

⁶LLaMA-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/>; <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 4: Aggregated results from Figure 3. Since results have been calibrated by human difficulty, then we can aggregate the results of several benchmarks in a meaningful way, even if for some of the tasks we did not have results for some bins.



Source: Own elaboration with data from (Zhou et al., 2025).

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 5 and 6), 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 models, e.g., while larger models still yield better performance, the rate of ability growth slows significantly beyond a certain size.

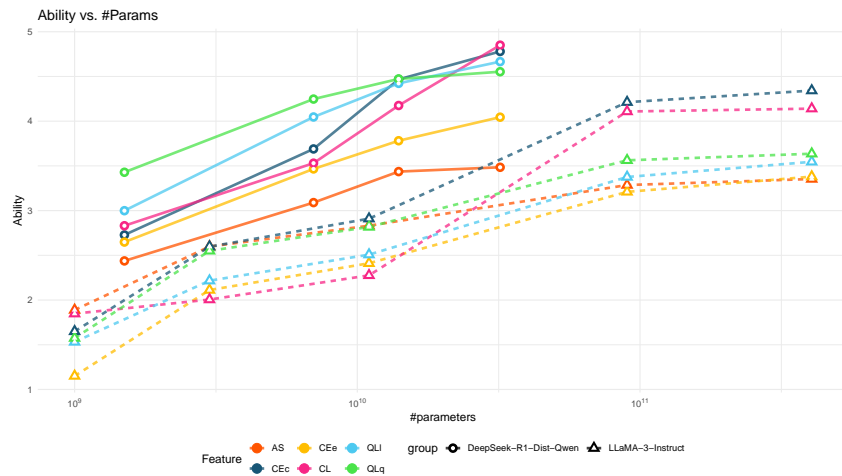
4.3 Sensitivity analysis of classification thresholds and averages

In this section, we empirically assess how changing the thresholds and aggregation methods for domain-level capability scores impacts the categorisation of AI models as GPAI models 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 (including LLaMA-3, DK-R1-Distilled-Qwen, GPT-3.5, and GPT-4 variants), as analysed in (Zhou et al., 2025). For each model, we obtain domain-level ability scores (specifically, the demand level at which the model achieves at least 50% success) across the four primary domains (and their subdomains) identified; Attention & Search (**AS**); Comprehension and Expression (with Verbal Comprehension (**CEc**) & Verbal Expression (**CEe**) as specific dimensions); Conceptualisation, Learning & Abstraction (**CL**); and Quantitative & Logical Reasoning (with Logical Reasoning (**QLI**) and Quantitative Reasoning (**QLq**) as specific dimensions). These scores are derived from the ADeLe battery (v1.0), as detailed in Section 3.1.

For the analysis, we explore:

Figure 5: The scaling curves (number of parameters) of actual abilities for LLaMA and DK-R1-Distilled-Qwen families across four broad demands: Attention & Search (AS); Comprehension and Expression (with Verbal Comprehension (CEc) & Verbal Expression (CEe) as specific dimensions); Conceptualisation, Learning & Abstraction (CL); and Quantitative & Logical Reasoning (with Logical Reasoning (QLI) and Quantitative Reasoning (QLq) as specific dimensions). Data from (Zhou et al., 2025).



Source: Own elaboration with data from (Zhou et al., 2025).

- **Aggregation functions:** arithmetic mean, harmonic mean, and geometric mean.
- **Thresholds:** Varying thresholds for categorising a model as a GPAI model. We assume, for the experiment, that the threshold for GPAI categorisation is set to values between 3 and 4 in each of the domains to limit the number of AI models that would be considered GPAI models.
- **Pass/fail rules:** Based on either (a) all domains exceeding a threshold, or (b) at least N out of 5 domains (with N varied from 1 to 6).

4.3.1 Effect of aggregation function

The choice of aggregation function materially affects a model’s aggregate score, and thus the set of models categorised as GPAI models at any fixed threshold. The arithmetic mean is most forgiving, as a high score 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, heavily penalising low scores in any domain.

However, as can be seen in Table 1, our results indicate that, in practice, the differences between these aggregation methods are relatively modest (generally less than 0.05–0.1) for the set of contemporary LLMs evaluated, especially for models that display reasonably balanced performance across all domains. In cases where a model is very strong in some domains but particularly weak in one, 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 classification for these models is typically robust to the specific averaging method chosen.

Figure 9, which shows radar plots of model ability profiles, further illustrates this point: the regular shapes of the high-performing models result in small differences between the arithmetic, geometric and harmonic means. Only models with pronounced bottlenecks in specific domains would show a more substantial discrepancy among the aggregation functions.

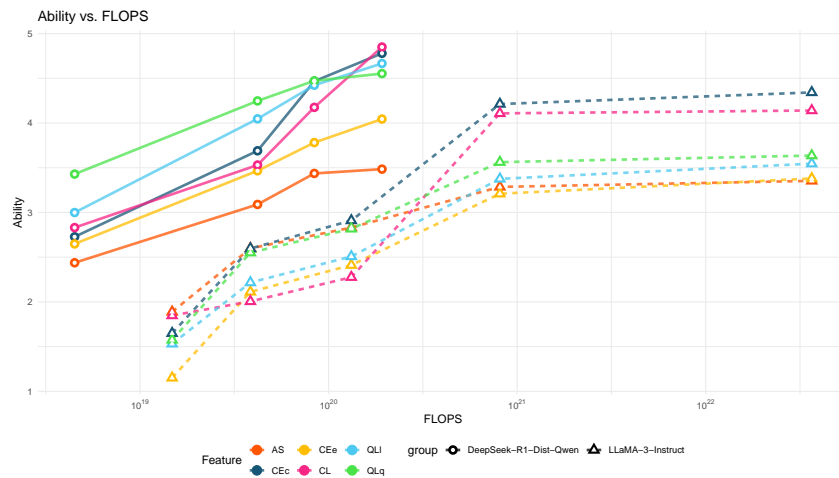
For this reason, the **arithmetic mean** may be a reasonable choice given it is the simplest, most

Table 1: 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 classification. 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

Source: Own elaboration with data from (Zhou et al., 2025).

Figure 6: The scaling curves (FLOP) of actual abilities for LLaMA and DK-R1-Distilled-Qwen families across four broad demands (as in Figure 5). Data from (Zhou et al., 2025). 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.



Source: Own elaboration with data from (Zhou et al., 2025).

intuitive aggregation method, and produces almost identical results to other aggregation methods based on current model profiles.

4.3.2 Effect of threshold value

Selecting the threshold value is also important in categorising AI models as 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 10, we observe the following trends:

- At **Lower Thresholds** (e.g., $\sim 3.0 - 3.5$): A broad range of models, including smaller and less recent 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" consideration 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 10) are categorised as GPAI models.
- At **High or Stringent Thresholds** (e.g., ≥ 4.5): Only the very top-performing models (e.g., OpenAI o1) retain GPAI status. Slight deficits in a single domain or minor regressions can exclude models that are otherwise reasonably general and highly capable.

The choice of threshold provides policymakers with an adjustable tool to tune the strictness of the GPAI 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 and sets a high bar, but can quickly shrink the pool of qualifying models, possibly below what the market or law intends.

Importantly, these trends hold across all three averaging strategies (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.

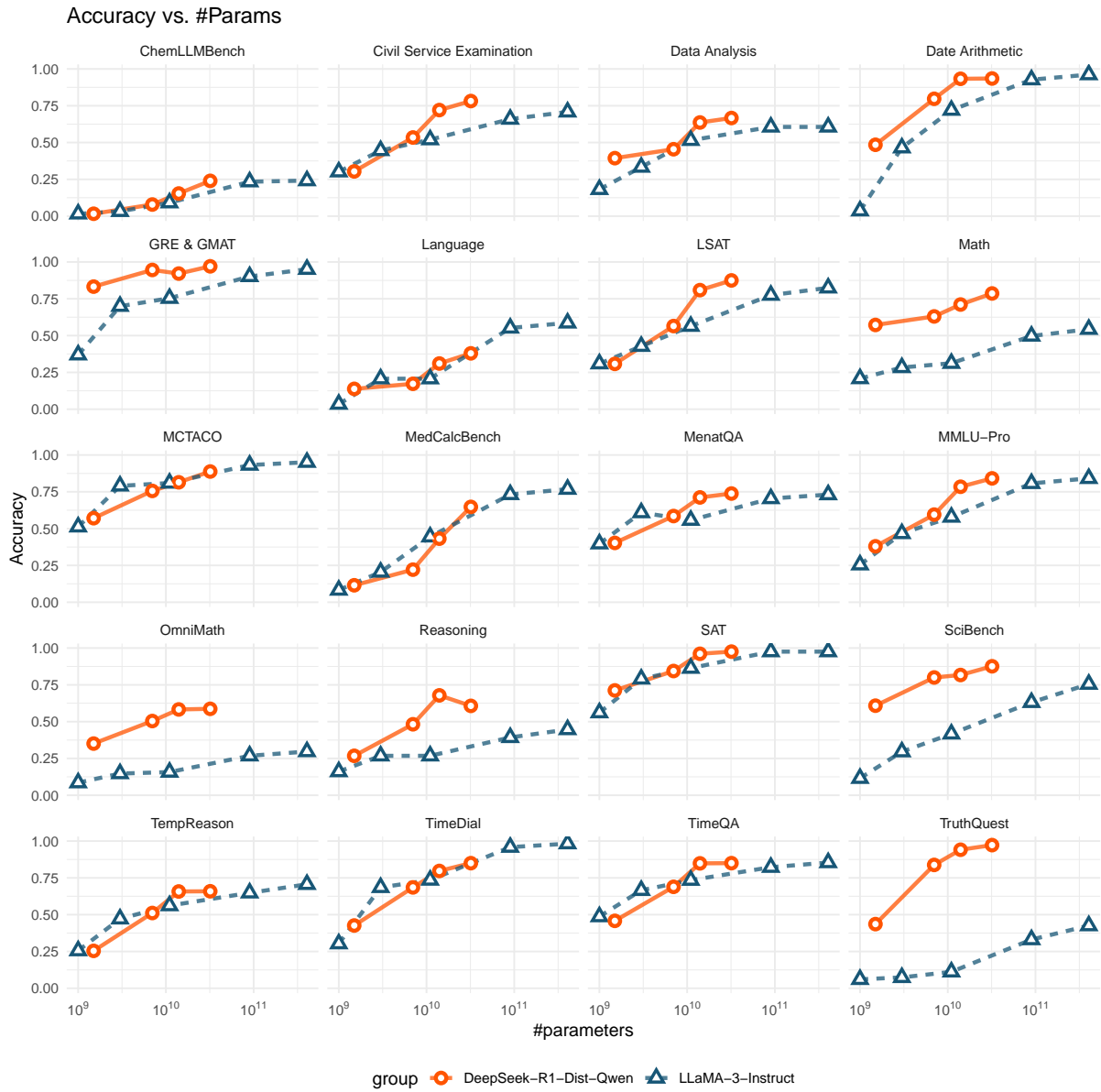
4.3.3 Domain pass/fail policy

Figure 11 illustrates the effects of domain-level policies. If the rule requires all domains to meet the threshold ('hard' rule), then only models with uniform capability are categorised as GPAI models, 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', perhaps due to architectural biases or gaps in their training data that lead to specific deficits. As a result, such policies may disqualify models that, by most practical standards, would still perform at a level consistent with what might be expected from GPAI models given the intention behind their definition.

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 whose abilities are on the cusp. 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.

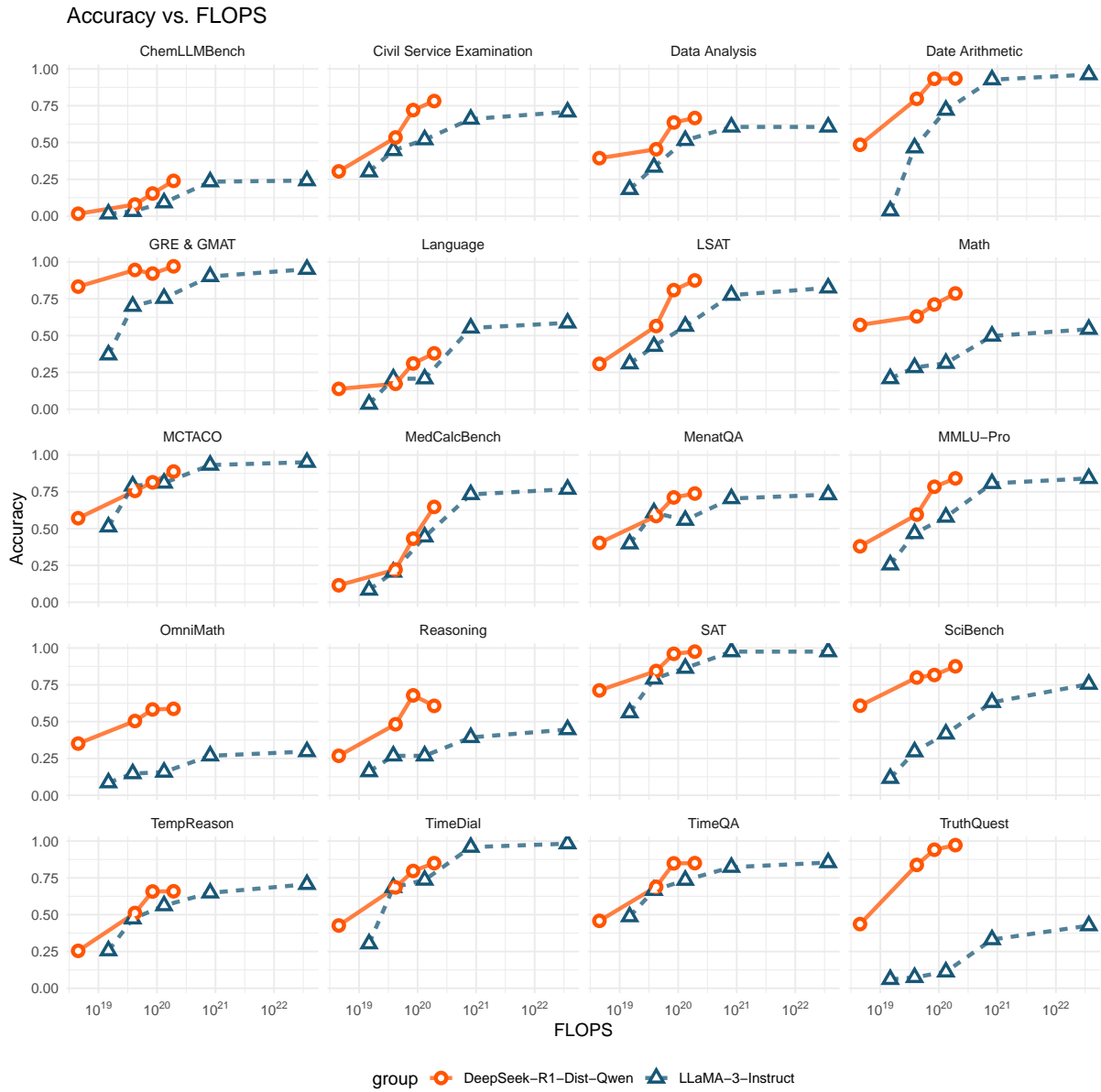
One possible approach in this regard would be to require models to exceed the competency threshold in at least three out of four (or a similar proportion) of the assessed domains to qualify as a GPAI model. Any chosen approach needs to balance the need for broad, reliable capabilities with realistic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

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



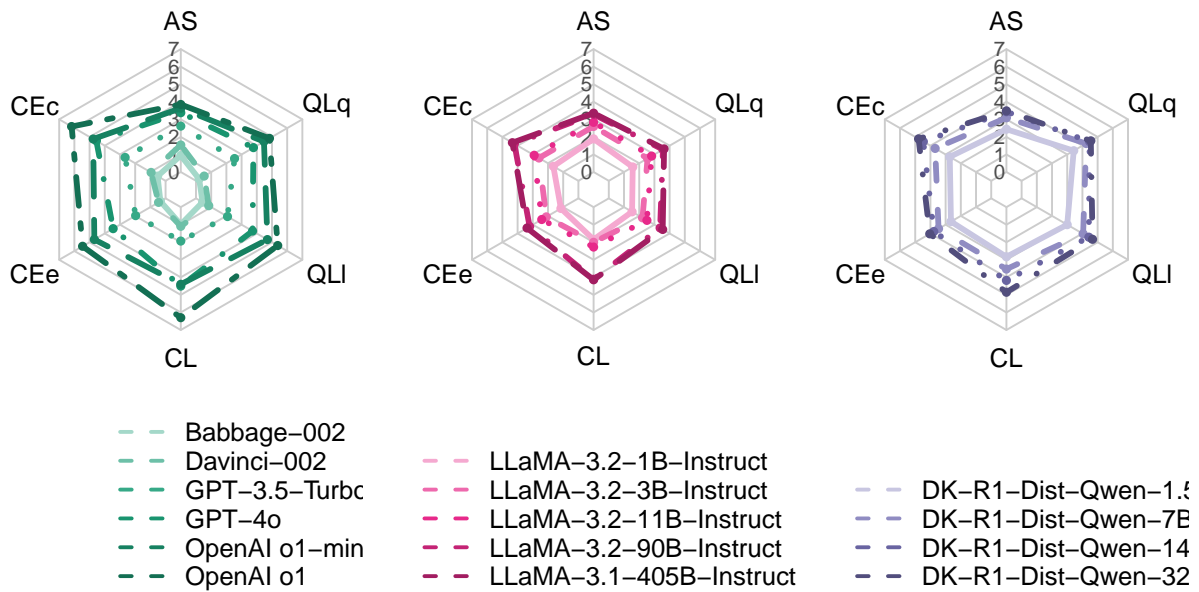
Source: Own elaboration with data from (Zhou et al., 2025).

Figure 8: The scaling curves (FLOP) of performance for LLaMA and DK-R1-Distilled-Qwen families across 20 different AI benchmarks.



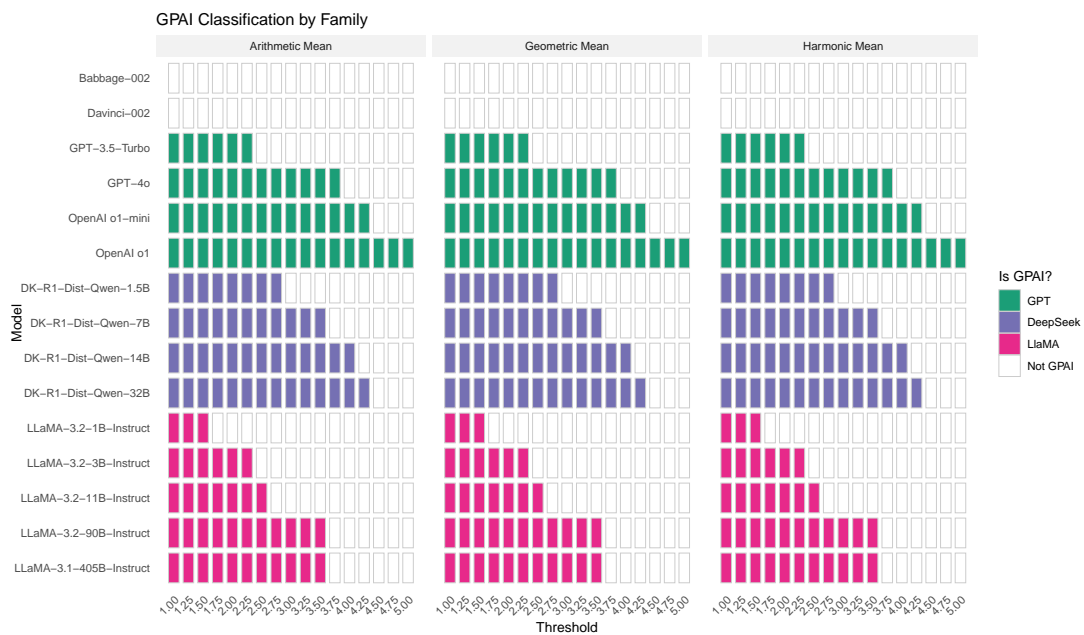
Source: Own elaboration with data from (Zhou et al., 2025).

Figure 9: 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.



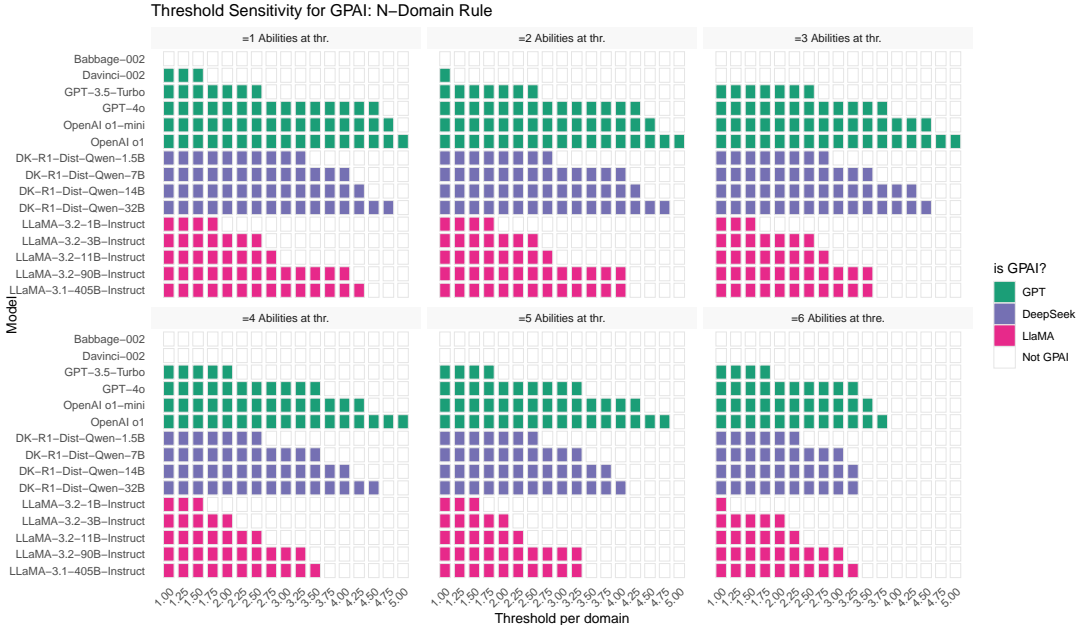
Source: Own elaboration with data from (Zhou et al., 2025).

Figure 10: Heatmap grid 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 categorised as a GPAI model at the given threshold under the specified aggregation. Colours denote model families (GPT, DeepSeek, LLaMA) and non-GPAI in white. Stricter thresholds/adverse means significantly reduce the number of models passing the GPAI bar.



Source: Own elaboration with data from (Zhou et al., 2025).

Figure 11: GPAI classification results as a function of the minimum number of domains required to exceed the threshold ("N-domain rule", from at least 1 to all 6 domains, including all subdomains). Rows correspond to models, columns to threshold values per domain. Coloured cells indicate GPAI status by model family as in previous figures.



Source: Own elaboration with data from (Zhou et al., 2025).

5 Conclusions

In this report, we introduced a principled framework that could be used to determine whether an AI model should be categorised as a general-purpose AI (GPAI) model. Our approach draws on established theories from cognitive psychology and foundational principles from psychometrics, with the goal of grounding GPAI designation in scientifically robust criteria. Central to this framework is a set of four core cognitive abilities, selected from a broader set of 14 abilities to reflect a diverse and representative range of domains that would jointly characterise general-purpose intelligence. To recap, these were (1) Attention and Search, (2) Comprehension and Compositional Expression, (3) Conceptualisation, Learning and Abstraction, and (4) Quantitative and Logical Reasoning.

To apply this framework, we require the ability to assign, for each task instance, a demand profile (a vector indicating the extent to which each of the 4 abilities –14 in the broader proposal– is required to succeed on that task instance). This enables us to treat capabilities as latent traits expressed to varying degrees across tasks. We adopt the ADeLe methodology (Zhou et al., 2025) for this purpose, leveraging large language models to annotate task demands using carefully constructed and validated rubrics. We also propose an extension to this methodology that calibrates model performance against human baselines, yielding interpretable and normed capability scores.

We have demonstrated the feasibility of this framework by evaluating several leading LLM families, including GPT, LLaMA, and Qwen. Our results show a general correlation between training compute (FLOP) and capability levels, but also reveal that some cognitive abilities emerge at a lower number of FLOP, suggesting non-uniform scaling behaviour across domains.

We proposed two complementary strategies for assigning GPAI status based on the final 4-dimensional capability profile. The first aggregates across dimensions using summary statistics such as the arithmetic, geometric, or harmonic mean—each of which weighs strengths and weaknesses differently. 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. This method offers interpretability and flexibility, especially when aligned with external stakes or deployment contexts.

While we have outlined two principled approaches for assigning GPAI status, we deliberately refrain from prescribing fixed thresholds or cut-offs ourselves. The appropriate standard for GPAI designation should be determined by the competent authority, and in accordance with legal interpretations of the EU AI Act. Moreover, as AI models continue to improve and the landscape of cognitive capabilities shifts, any thresholding scheme must be periodically re-evaluated to remain meaningful. We view our framework as a foundation: it provides a methodology for GPAI model categorisation, but does not dictate how the parameters that underpin the methodology, and that allow the categorisation in practice, should be set.

Together, these contributions lay the groundwork for a transparent, empirically grounded, and practically useful operationalisation of the definition of GPAI models: one that can evolve alongside the models it is designed to evaluate.

References

- Carroll, J. B., 'Human cognitive abilities: A survey of factor-analytic studies', 1. Cambridge university press, 1993.
- Davidson, T., Denain, J.-S., Villalobos, P. and Bas, G., 'Ai capabilities can be significantly improved without expensive retraining', [arXiv preprint arXiv:2312.07413](https://arxiv.org/abs/2312.07413), 2023.
- Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gómez, E. and Fernández-Llorca, D., 'Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation', [Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society \(AIES-25\)](https://proceedings.aai.acm.org/), 2025.
- 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.
- 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.
- Gardner, H. and Hatch, T., 'Educational implications of the theory of multiple intelligences', [Educational researcher](https://doi.org/10.1177/0013164489018008004), Vol. 18, No 8, 1989, pp. 4–10.
- Hernández-Orallo, J., Schellaert, W. and Martínez-Plumed, F., 'Training on the test set: Mapping the system-problem space in ai', In 'Proceedings of the AAAI conference on artificial intelligence', Vol. 36. pp. 12256–12261.
- Hernández-Orallo, J., Sevilla, J., Gómez, E. and Fernández-Llorca, D., 'General-Purpose AI Models in the AI Act: Capabilities, Generality, Systemic Risks and Compute', [European Commission, Joint Research Centre, JRC139341](https://www.jrc.ec.europa.eu/en/publications-and-communications/gp-ai-models-in-the-ai-act-capabilities-generality-systemic-risks-and-compute), 2024.
- Keith, T. Z. and Reynolds, M. R., 'Cattell–horn–carroll abilities and cognitive tests: What we've learned from 20 years of research', [Psychology in the Schools](https://doi.org/10.1177/0013164410377000), Vol. 47, No 7, 2010, pp. 635–650.
- Linstone, H., 'The delphi method: Techniques and applications in linstone ha, turoff m', [URL: http://www.is.njit.edu/pubs/delphibook](http://www.is.njit.edu/pubs/delphibook/), 1975.
- Sternberg, R. J., 'Beyond iq: A triarchic theory of human intelligence', CUP Archive, 1985.
- Thurstone, L. L., 'Ability, motivation, and speed', [Psychometrika](https://doi.org/10.1037/h0075558), Vol. 2, No 4, 1937, pp. 249–254.
- Tolan, S., Pesole, A., Martínez-Plumed, F., Fernández-Macías, E., Hernández-Orallo, J. and Gómez, E., 'Measuring the occupational impact of ai: tasks, cognitive abilities and ai benchmarks', [Journal of Artificial Intelligence Research](https://doi.org/10.1007/s10992-021-10000-0), Vol. 71, 2021, pp. 191–236.
- Zhou, L., Pacchiardi, L., Martínez-Plumed, F., Collins, K. M., Moros-Daval, Y., Zhang, S., Zhao, Q., Huang, Y., Sun, L., Prunty, J. E. et al., 'General scales unlock ai evaluation with explanatory and predictive power', [arXiv preprint arXiv:2503.06378](https://arxiv.org/abs/2503.06378), 2025.
- Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C. and Hernández-Orallo, J., 'Larger and more instructable language models become less reliable', [Nature](https://doi.org/10.1038/s41586-024-0468-4), Vol. 634, No 8032, 2024, pp. 61–68.

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 Risk

LLM Large Language Model

ROC Receiver Operating Characteristic (curve)

UI User Interface

UX User Experience

List of figures

Figure 1.	Cattell-Horn-Carroll’s three stratum model. The broad abilities are Crystallised Intelligence (Gc), Fluid Intelligence (Gf), Quantitative Reasoning (Gq), Reading and Writing Ability (Grw), Short-Term Memory (Gsm), LongTerm Storage and Retrieval (Glr), Visual Processing (Gv), Auditory Processing (Ga), Processing Speed (Gs) and Decision/Reaction Time/Speed (Gt).	12
Figure 2.	Ability profiles (radial plots) for OpenAI and DK LLMs using the ADeLe battery 1.0 (left). Demand profiles for a selection of four benchmarks from the ADeLe battery v.1.0 (right). The colour intensity illustrates the frequency or number of issues for each domain and level of difficulty.	16
Figure 3.	Results for Addition (QL), anagram (AS, CE), CommonTransforms (AS, CE), GPQA (AS, CE), locality (QL), OpenBookQA (AS, CE) for several models. The x-axis locates each example in bins depending on the percentage of human success.	22
Figure 4.	Aggregated results from Figure 3. Since results have been calibrated by human difficulty, then we can aggregate the results of several benchmarks in a meaningful way, even if for some of the tasks we did not have results for some bins.	23
Figure 5.	The scaling curves (number of parameters) of actual abilities for LLaMA and DK-R1-Distilled-Qwen families across four broad demands: Attention & Search (AS); Comprehension 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 (Zhou et al., 2025).	24
Figure 6.	The scaling curves (FLOP) of actual abilities for LLaMA and DK-R1-Distilled-Qwen families across four broad demands (as in Figure 5). Data from (Zhou et al., 2025). 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.	26
Figure 7.	The scaling curves (number of parameters) of performance for LLaMA and DK-R1-Distilled-Qwen families across 20 different AI benchmarks.	28
Figure 8.	The scaling curves (FLOP) of performance for LLaMA and DK-R1-Distilled-Qwen families across 20 different AI benchmarks.	29
Figure 9.	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.	30
Figure 10.	Heatmap grid 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 categorised as a GPAI model at the given threshold under the specified aggregation. Colours denote model families (GPT, DeepSeek, LLaMA) and non-GPAI in white. Stricter threshold-/adverse means significantly reduce the number of models passing the GPAI bar.	30
Figure 11.	GPAI classification results as a function of the minimum number of domains required to exceed the threshold (“N-domain rule”, from at least 1 to all 6 domains, including all subdomains). Rows correspond to models, columns to threshold values per domain. Coloured cells indicate GPAI status by model family as in previous figures.	31

List of tables

Table 1. 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 classification. The choice of aggregation function affects whether uneven domain performance is penalised (harmonic mean) or compensated (arithmetic mean). 25

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696,
- via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society



Scan the QR code to visit:

[The Joint Research Centre: EU Science Hub](https://joint-research-centre.ec.europa.eu)

<https://joint-research-centre.ec.europa.eu>



Publications Office
of the European Union