AI Benchmark Annotation Dataset

Fernando Martínez-Plumed – fmartinez@dsic.upv.es

with expert number
EX2018D335821
for contract number
CT-EX2018D335821-102
regarding
Deliverable 2:
AI Benchmark Annotation Dataset

for the
European Commission
Joint Research Center
Unit JRC/B/06

August 4, 2025

D3: AI Benchmark Annotation Dataset

Background

The contract for which this report is produced exists in the context of the **AIM-WORK project** which focuses on understanding the impact of artificial intelligence (AI) and machine learning (ML) technologies on the workplace, particularly through the lens of large language models (LLMs) developed between 2020 and 2024. Previous research [7, 3] has explored how AI can automate a wider range of job functions compared to previous automation technologies, challenging the boundaries previously set by Polanyi's paradox [6]. This project aims to build on these findings by revisiting and extending the AI impact framework developed by Songül Tolan and colleagues in 2021, incorporating the advanced capabilities of recent Large Language Models (LLMs).

This deliverable compiles an updated dataset of 352 AI benchmarks mapped to cognitive abilities and tasks, encompassing the 328 benchmarks analysed in [7] and 25 new benchmarks that have gained prominence in the past five years. For each benchmark we extract the number of documents mentioning it in the AITopics[1] archive for each year between 2008 and 2024 (both raw counts and a normalised measure relative to the yearly total). AITopics continually ingests new papers, meaning that even historical counts for 2008–2019 have grown substantially since the JAIR article was published. The dataset therefore offers a richer view of research intensity and enables a comparison of pre-LLM and post-LLM trends.

Introduction

Over the last two decades AI research has been driven by community benchmarks that provide a shared basis for measuring progress. Public leaderboards such as ImageNet, SQuAD and GLUE have steered both academic and industrial research, with models achieving new state-of-the-art results receiving widespread recognition [5]. Such benchmarks not only evaluate performance but also implicitly define the tasks that attract attention and resources[4]. The recent emergence of LLMs [1, 8, 2, 9], built through large-scale pre-training on massive corpora, has enabled cross-domain generalisation and sparked the development of new evaluation suites (e.g., BIG-bench, HELM) that test reasoning, alignment and multi-step planning. The AIM-WORK project seeks to understand how this shift affects the relative intensity of research across cognitive abilities and, by extension, which occupational tasks may be impacted next.

Deliverable D2 extends our earlier benchmark catalogue by automatically annotating both existing and new benchmarks with the cognitive abilities they require. Rather than relying exclusively on human experts and Delphi rounds, we complement the annotations with large language models (GPT-4) to scale the process. This deliverable therefore serves two goals: (i) to present the expanded benchmark dataset with updated intensity measures and annotations, and (ii) to provide an initial analysis of trends across abilities before and after the

[1] https://aitopics.org

arrival of LLMs.

Methodology

Following [7], we adopt a three-layer framework:

1. **Tasks**: units of work activity, derived from labour surveys and occupational taxonomies, spanning physical, intellectual and social domains.
2. **Cognitive abilities**: an intermediate layer that captures the mental faculties required to perform tasks. The fourteen abilities include memory processes (MP), sensorimotor interaction (SI), visual processing (VP), auditory processing (AP), attention and search (AS), pattern analysis (PA), causal and environmental reasoning (CE), communication (CO), emotion comprehension (EC), navigation (NV), conceptualisation (CL), quantitative and logical reasoning (QL), motor skills (MS) and meta-cognition (MC). While some abilities, such as MP or VP, are widely represented in AI research, others (e.g., MC) have few associated benchmarks.
3. **AI benchmarks**: publicly available datasets, competitions or tasks used to evaluate AI systems. We integrate 352 benchmarks by matching their names to AITopics queries and automatically assigning abilities based on keyword analysis of the benchmark description, expert annotations and GPT-4 outputs.

For each benchmark the dataset includes (i) the total number of AITopics documents per year from 2008–2024 that mention the benchmark, and (ii) a normalised value obtained by dividing the benchmark's document count by the total number of AI documents in that year. The normalised measure enables comparisons across years despite growth in the overall literature.

To determine pre-LLM versus post-LLM trends we split the series at 2018/2019: the years up to and including 2018 capture the period prior to transformer-based LLMs (e.g., GPT-2), while 2019 onwards represents the era of rapid LLM development. New benchmarks are identified as those with zero recorded documents prior to 2019 but non-zero counts thereafter.

Computing ability-level intensities

Let $I_{b,t}$ denote the normalised intensity of benchmark b in year t. Each benchmark b is associated with a binary vector $\boldsymbol{a}_b \in \{0,1\}^{14}$ indicating which of the fourteen cognitive abilities it exercises. The annual research intensity for ability j in year t is computed as

$$R_{j,t} = \sum_{b:\, a_{b,j}=1} I_{b,t},$$

where $a_{b,j}$ is the jth element of $\boldsymbol{a}_b$. Summing $R_{j,t}$ over all years yields the cumulative research intensity for each ability. We compare the totals for 2008–2018 and 2019–2024 to capture the effect of LLMs.

Automatic annotation via LLMs

While the original framework relied on expert judgement to map benchmarks to abilities, we leverage the contextual reasoning of large language models (GPT-4) to scale the annotation. For each benchmark we supply the official task description and rubric definitions of the abilities, and ask the model to estimate which abilities are required. These automated suggestions are then cross-checked by four human experts. This hybrid approach reduces annotation time while preserving accuracy. In the appended dataset we include both the human and GPT-4 annotations; disagreements were resolved through majority vote.

The framework in [7] defined a set of detailed rubrics that link benchmark tasks to cognitive abilities through observable task characteristics. Each ability is described in terms of the cognitive functions it entails and the types of tasks that require it. For example, *memory processes* encompass tasks that require recalling facts, storing information across context windows or retrieving relevant documents. The rubrics can be found in the Supplementary Material of [7].

To scale annotation beyond manual expert mapping, we employ a large language model (GPT-4) to act as a "virtual annotator." For each benchmark we construct a prompt that includes (i) the benchmark's official task description and, where available, its task hierarchy from AITopics or PapersWithCode, and (ii) the rubric definitions of the fourteen abilities. The model is asked to list the abilities that it believes are necessary to solve the task, along with a brief justification. An example prompt reads:

> *You are given the description of an AI benchmark: "MathVista is a visual question answering dataset with diagrams and mathematical plots." Based on the definitions below, which cognitive abilities are required to perform well on this benchmark? Abilities include: 1. Memory processes (requires storing and recalling information)... 2. Sensorimotor interaction ... 3. Visual processing ... 4. Quantitative/logic ... 5. Attention and search ...*

The model's output is parsed into a binary vector $\boldsymbol{a}_b$ indicating the presence or absence of each ability. These automated annotations are then reviewed by four human experts with experience in cognitive science and AI benchmarking. Experts consult the same rubric and accept, reject or modify the model's suggestions. In cases of disagreement between annotators, we adopt the majority vote. This hybrid procedure greatly reduces the time required to annotate hundreds of benchmarks while maintaining high quality. The final dataset includes both the expert and GPT-4 annotations, enabling future researchers to assess the consistency of annotations or to fine-tune the rubric. We also record the confidence levels returned by the model, which can be used to prioritise benchmarks for human review in subsequent updates.

Data overview

The compiled dataset contains 352 benchmarks. Table 1 reports the number of benchmarks per type and the number of benchmarks associated with each cognitive ability. Most entries (348) are datasets, with a handful of game environments and a single prize competition. The abilities with the largest number of associated benchmarks are attention and search (AS),

visual processing (VP), quantitative/logic (QL) and conceptualisation (CL). Meta-cognition (MC) currently has no dedicated benchmarks, highlighting a gap in evaluation resources.

Table 1: Counts of benchmarks by category and cognitive ability.

Ability	#benchmarks
MP (memory processes)	20
SI (sensorimotor interaction)	12
VP (visual processing)	162
AP (auditory processing)	15
AS (attention and search)	190
PA (pattern analysis)	17
CE (causal/environmental reasoning)	148
CO (communication)	28
EC (emotion comprehension)	17
NV (navigation)	11
CL (conceptualisation)	90
QL (quantitative/logic)	91
MS (motor skills)	9
MC (meta-cognition)	0

In Table 2 shows thelist of 24 benchmarks that were not part of the original set analysed by [7] and that have been added in the AIM-WORK update. We selected these benchmarks because they have emerged since 2019, are widely cited in the AITopics archive, and cover task types that were previously under-represented. Many of them evaluate new capabilities of large language models (LLMs) and multi-modal systems, reflecting the need for fresh benchmarks to avoid saturation. The rationale for inclusion is as follows:

- **Emergence of LLMs and multi-modal models**: Several new tasks (MMLU, TruthfulQA, MMLU-PRO, GPQA, IFEval, ARC-AGI) specifically target the broad knowledge, instruction following and reasoning abilities exhibited by modern LLMs. They provide a more comprehensive evaluation of cognitive abilities such as memory processes, reasoning and communication, which have grown in importance in the LLM era.
- **Multi-modal reasoning**: Benchmarks like MMBench, MMMU, MMVet, MathVista, HallusionBench and MMStar evaluate models that process both language and images, addressing the emerging trend of multi-modal transformers. These tasks were absent from the original set and help gauge abilities such as visual processing combined with logic and reasoning.
- **Robustness and safety**: TruthfulQA and HallusionBench test model robustness to hallucination and deception, while Disguised Faces in the Wild challenges models to recognise identities under disguise. Including these benchmarks responds to concerns about reliability and safety of AI systems.
- **Medical and scientific domains**: TCIA Pancreas CT, BUS 2017, PROMISE 2012, DIC HeLa and PhC-U373 extend coverage to 3-D medical imaging and cell segmen-

tation tasks. As AI adoption in healthcare grows, these datasets ensure that relevant abilities (e.g., visual processing and pattern analysis) are adequately represented.

- **Low-resource languages and specialised tasks**: Weibo NER adds coverage for Chinese social-media named-entity recognition, while LSUN Bedroom, CIHP, Occluded LINEMOD and OCRBench broaden the types of vision tasks beyond the well-studied ImageNet/COCO suites.

In a nutshell, these new benchmarks diversify the evaluation landscape, helping to mitigate the saturation of older datasets and ensuring that the framework remains relevant to emerging AI capabilities and application domains.

Table 2: New benchmarks added to the AIM–WORK dataset compared to Tolan et al.

Benchmark	Key task/category
MMLU	Multi-task language understanding (broad knowledge tasks)
TruthfulQA	Truthful question answering, hallucination detection
MMBench	Visual question answering (multi-modal reasoning)
MMMU	Multi-modal reasoning over images and mathematics
MMVet	Veterinary image visual question answering
MathVista	VQA with mathematical diagrams and graphs
GPQA	Goal-planning and physics question answering
IFEval	Instruction-following language understanding
MMLU-PRO	Professional domain multi-task language understanding
HallusionBench	Visual question answering detecting hallucinations
MMStar	VQA with astronomical charts (star recognition)
OCRBench	Optical character recognition evaluation
Weibo NER	Chinese social-media named-entity recognition
MuSR	Common-sense reasoning (multi-modal)
LSUN Bedroom 256×256	High-resolution image generation (bedroom scenes)
Disguised Faces in the Wild	Face verification under disguise/occlusion
TCIA Pancreas CT	3-D medical imaging segmentation (pancreas)
BUS 2017	Medical lesion segmentation (breast ultrasound)
CIHP	Human part segmentation (clothing/human parsing)
Occluded LINEMOD	6-D pose estimation with occluded objects
PROMISE 2012	Volumetric prostate MRI segmentation
ARC-AGI	Abstract reasoning puzzles for language models
DIC HeLa	Cell segmentation (HeLa microscopy images)
PhC-U373	Cell segmentation (glioma microscopy images)

Exploratory data analysis

While the main focus of this deliverable is to present the updated benchmark dataset, it is informative to explore several structural properties of the data before delving into temporal trends. The following figures summarise the distribution of benchmarks and intensities.

Benchmarks per ability. Figure 1 displays the number of benchmarks associated with each cognitive ability. Visual processing (VP) and attention/search (AS) have the largest number of benchmarks, reflecting the predominance of computer vision and language tasks

in public evaluation suites. In contrast, meta-cognition (MC) currently has no dedicated benchmarks, highlighting a gap in available resources.

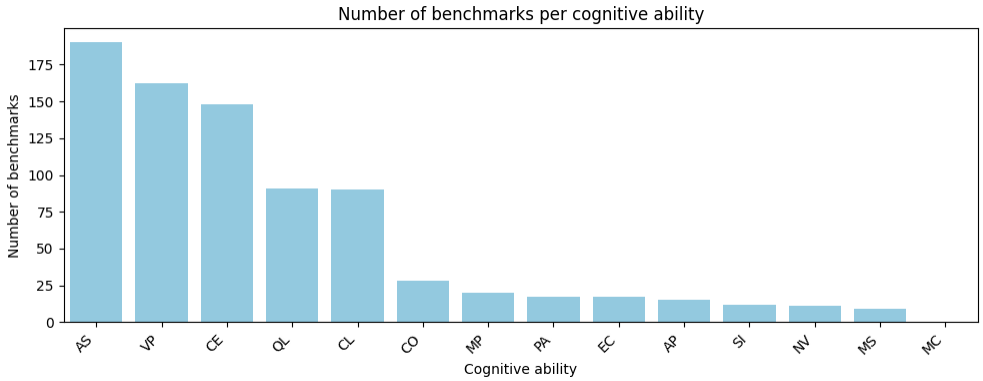

Figure 1: Number of benchmarks per cognitive ability. Bars show the count of benchmarks mapped to each ability. VP and AS dominate, while MC has no dedicated benchmarks.

Number of abilities per benchmark. Each benchmark may involve multiple cognitive abilities. Figure 2 plots the distribution of the number of abilities assigned to a benchmark. A majority of benchmarks involve one or two abilities, but a non-trivial fraction require three or more, illustrating the multi-faceted nature of some AI tasks.

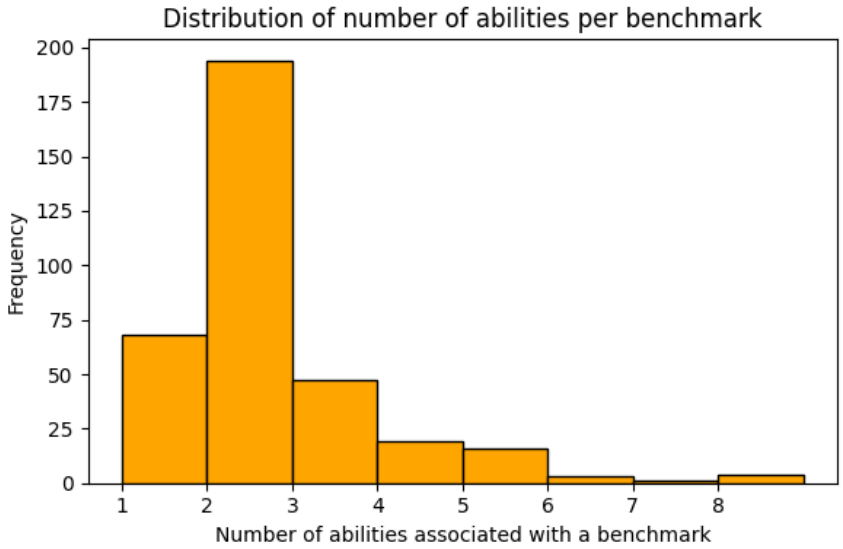

Figure 2: Distribution of the number of cognitive abilities per benchmark. The x-axis counts how many abilities are associated with a benchmark; the y-axis shows the number of benchmarks (log scale suppressed).

Intensity distribution across benchmarks. To gauge how research activity is distributed across benchmarks, we compute the average normalised intensity of each benchmark

across all years and plot the histogram in Figure 3. The distribution is heavy-tailed: a few benchmarks (e.g., ImageNet, COCO, SQuAD) accumulate high intensity, while most receive relatively little attention. The y-axis is shown on a log scale to highlight the long tail.

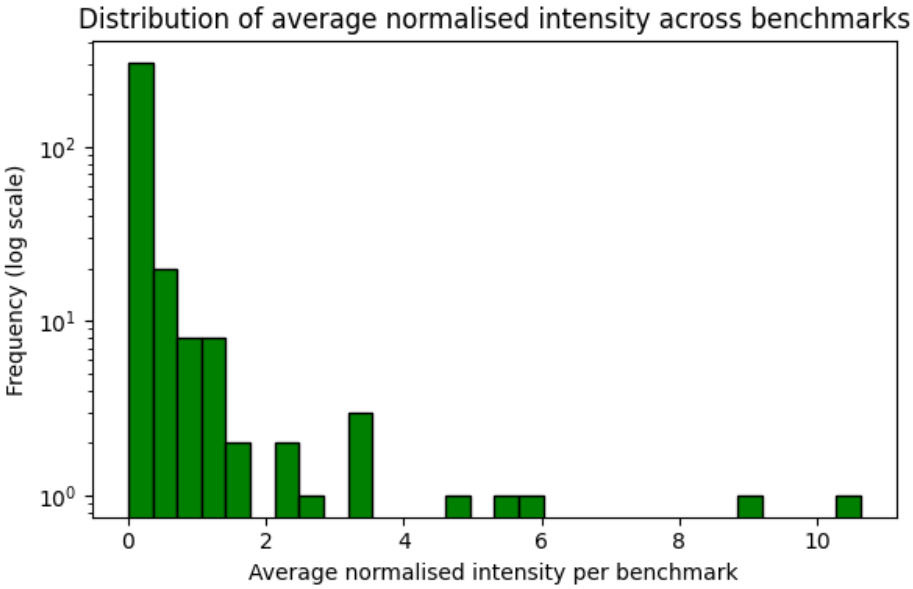

Figure 3: Distribution of average normalised intensity across benchmarks. Values are averaged over the years 2008–2024. The y-axis is on a log scale.

Introduction of new benchmarks. Figure 4 shows the number of benchmarks appearing for the first time in each year according to AITopics. The sharp increase after 2019 reflects the wave of benchmarks accompanying the development of LLMs and the broader diversification of evaluation tasks. This plot complements the more detailed analysis of pre- and post-LLM trends presented in Deliverable D4.

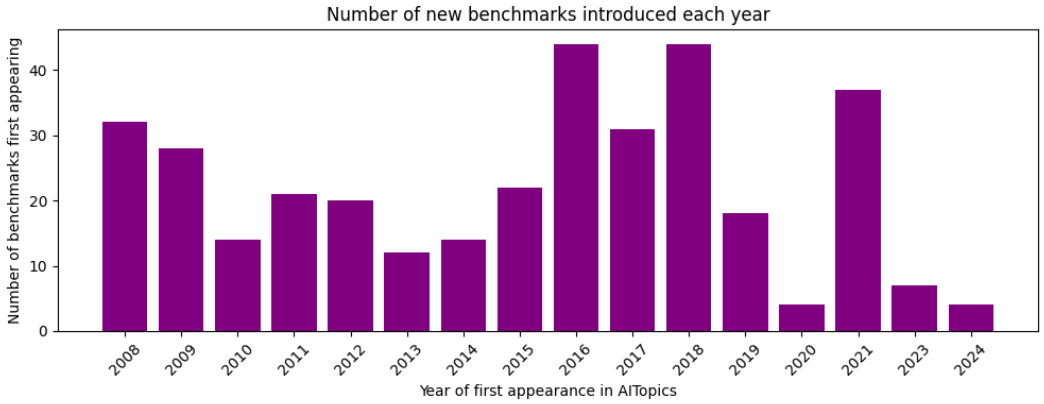

Figure 4: Number of benchmarks first appearing in the AITopics archive by year. The surge after 2019 corresponds to the introduction of numerous LLM-era benchmarks.

Conclusions

This deliverable presents an updated and expanded AI benchmark annotation dataset for the AIM-WORK project. By collecting AITopics intensity data up to 2024, adding 25 new benchmarks and introducing LLM-assisted annotations, we extend the framework of [7] to capture shifts in research focus brought about by modern language models. Future extensions could incorporate benchmark performance metrics, account for multi-modal models, and develop benchmarks for currently underrepresented abilities such as meta-cognition.

References

[1] Rishi Bommasani et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021).

[2] Hyung Won Chung et al. "Scaling Instruction-Finetuned Language Models". In: *arXiv preprint arXiv:2210.11416* (2022). DOI: 10.48550/arXiv.2210.11416. arXiv: 2210.11416 [cs].

[3] Fernando Martínez-Plumed et al. "Does AI qualify for the job? A bidirectional model mapping labour and AI intensities". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 94–100.

[4] Fernando Martínez-Plumed et al. "Research community dynamics behind popular AI benchmarks". In: *Nature Machine Intelligence* 3.7 (2021), pp. 581–589.

[5] Ray Perrault and Jack Clark. "Artificial Intelligence Index Report 2024". In: (2024).

[6] Michael Polanyi. "The logic of tacit inference". In: *Philosophy* 41.155 (1966), pp. 1–18.

[7] Songül Tolan et al. "Measuring the occupational impact of AI: tasks, cognitive abilities and AI benchmarks". In: *Journal of Artificial Intelligence Research* 71 (2021), pp. 191–236.

[8] Jason Wei et al. "Emergent abilities of large language models". In: *arXiv preprint arXiv:2206.07682* (2022).

[9] Wayne Xin Zhao et al. "A survey of large language models". In: *arXiv preprint arXiv:2303.18223* (2023).