Final Report: Findings

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

with expert number
EX2018D335821
for contract number
CT-EX2018D335821-102
regarding
Deliverable 4:
Final Report: Findings

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

August 4, 2025

D4: Final Report: Findings

Background

The AIM-WORK project investigates the ways in which advances in Artificial Intelligence (AI) and Machine Learning (ML) technologies affect labour markets. A key component of this work is the linkage between occupational tasks and the cognitive abilities that AI systems exercise. Previous research by [5] established a three-layer framework connecting jobs to tasks, tasks to cognitive abilities, and abilities to AI benchmarks. By measuring the intensity of research activity associated with each benchmark, their framework provided an indicator of which cognitive abilities might experience rapid progress and hence impact labour sooner. Recent studies, however, have argued that many widely used benchmarks have become saturated, potentially obscuring real progress and underscoring the need for new evaluation tasks.

Deliverable D2 of this project compiled an updated dataset of 352 benchmarks mapped to cognitive abilities, collected document counts from the AITopics archive for each benchmark between 2008 and 2024, and introduced LLM-assisted annotation methods. Since then, the rapid proliferation of large language models (LLMs) after 2019 has led to the creation of numerous new benchmarks designed to probe abilities such as long-context reasoning, multimodal integration and instruction following. The present deliverable (D4) extends the analysis by (i) normalising research intensity across periods of different length, (ii) computing growth factors for each ability, and (iii) identifying and examining the top new benchmarks that have emerged in the post-LLM era.

Introduction

Benchmarks lie at the heart of AI progress. They provide shared tasks and datasets on which researchers compete and collaborate, thereby steering community attention and resource allocation [3]. Classic benchmarks such as ImageNet and SQuAD catalysed leaps in computer vision and natural language processing performance [4]. More recently, the advent of transformer-based LLMs (e.g., GPT-2, GPT-3, PaLM, GPT-4) [1, 6, 2, 7] has spurred the design of evaluation suites like BIG-bench and HELM to assess broad language understanding, instruction following and safety. Understanding how research activity shifts across cognitive abilities in response to such innovations is critical for anticipating their impact on work tasks.

This deliverable aims to provide a comprehensive picture of these shifts. We analyse annual research intensity for fourteen cognitive abilities from 2008–2024, compute pre-LLM (2008–2018) and post-LLM (2019–2024) averages, visualise temporal dynamics through line charts and heatmaps, quantify growth factors, and examine the distribution of research across original and new benchmarks. In addition, we present a curated list of 24 benchmarks that were not included in the original [5] study but have gained significant attention in recent years, along with the rationale for their inclusion.

Analysis and results

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.

Temporal dynamics of ability shares

We first examine how the distribution of research intensity across abilities evolves over time. Figure 1 displays the normalised share of research intensity for each ability from 2008 through 2024. Each curve is the ratio $R_{j,t} / \sum_{k=1}^{14} R_{k,t}$, so the shares sum to one in every year. Visual processing (VP) dominates throughout but gradually declines after 2015 as new tasks emerge. Attention and search (AS) and quantitative/logic (QL) rise steadily, reflecting the importance of transformer architectures and language understanding tasks. Memory processes (MP) exhibit the sharpest post-2019 increase, consistent with the rise of retrieval-augmented LLMs. Abilities related to robotics—sensorimotor interaction (SI), navigation (NV) and motor skills (MS)—decline in relative share, suggesting a shift in research focus away from physical embodiment.

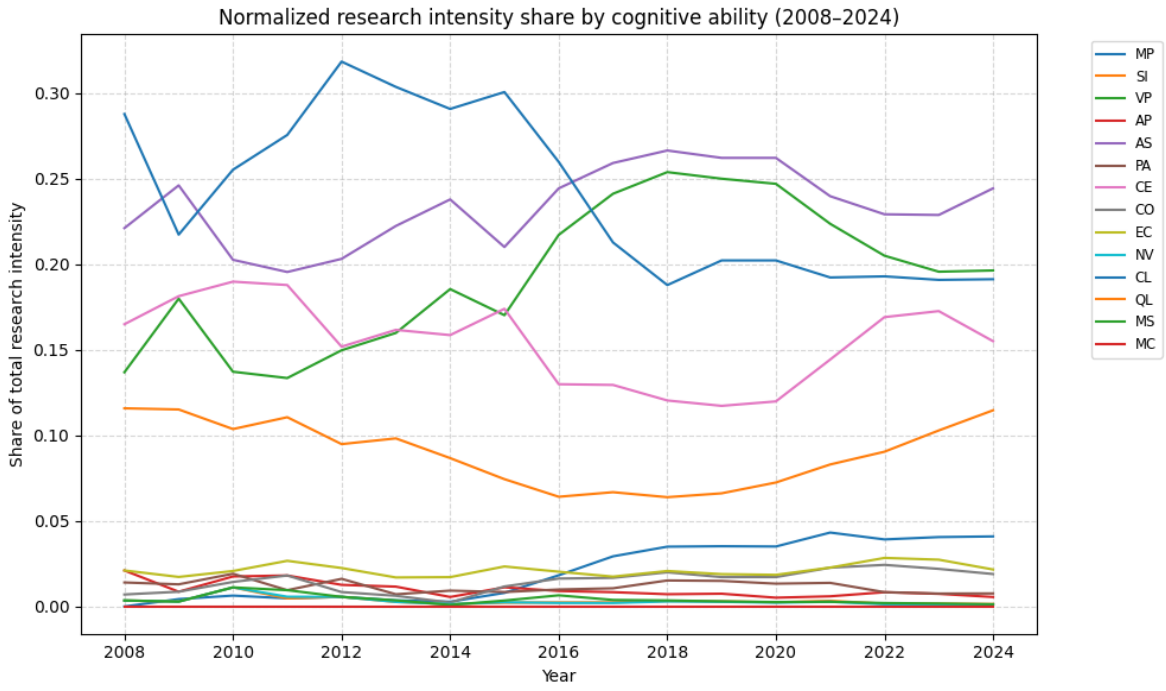

Figure 1: Normalised share of research intensity across cognitive abilities (2008–2024). Each ability's intensity is divided by the total intensity across all abilities in the same year. The curves reveal long-term trends in research emphasis.

Pre-LLM versus post-LLM comparisons

To assess the impact of LLMs, we compare average yearly intensities before and after 2019. Figure 2 presents a bar chart of $\bar{R}_j^{\text{pre}}$ versus $\bar{R}_j^{\text{post}}$ for each ability. Visual processing remains the largest contributor in both periods but its average intensity declines slightly. Memory processes more than double in average intensity, underscoring the importance of memory and context handling in LLM architectures. Communication (CO) grows modestly, while sensorimotor interaction, navigation and motor skills decline. Quantitative/logic (QL) and attention and search (AS) maintain robust levels across both periods.

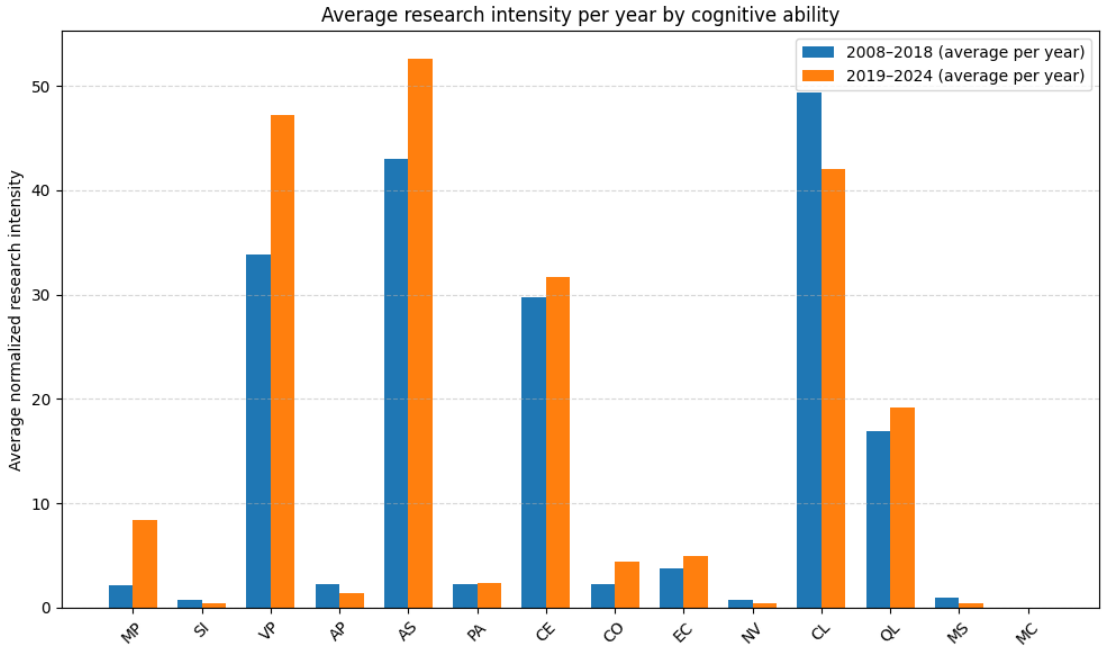

Figure 2: Average yearly research intensity by cognitive ability for the pre-LLM period (2008–2018) and the post-LLM period (2019–2024). Normalising by the number of years allows direct comparison between periods of different length.

Figure 3 shows the share of total intensity attributable to each ability in the pre- and post-LLM periods. Shares sum to one within each period. Visual processing's share falls from around 42% to 33%, while memory processes increase from 1.5% to 4.5%. Communication's share grows from 2% to 3.4%. These shifts demonstrate a diversification of research focus in the LLM era.

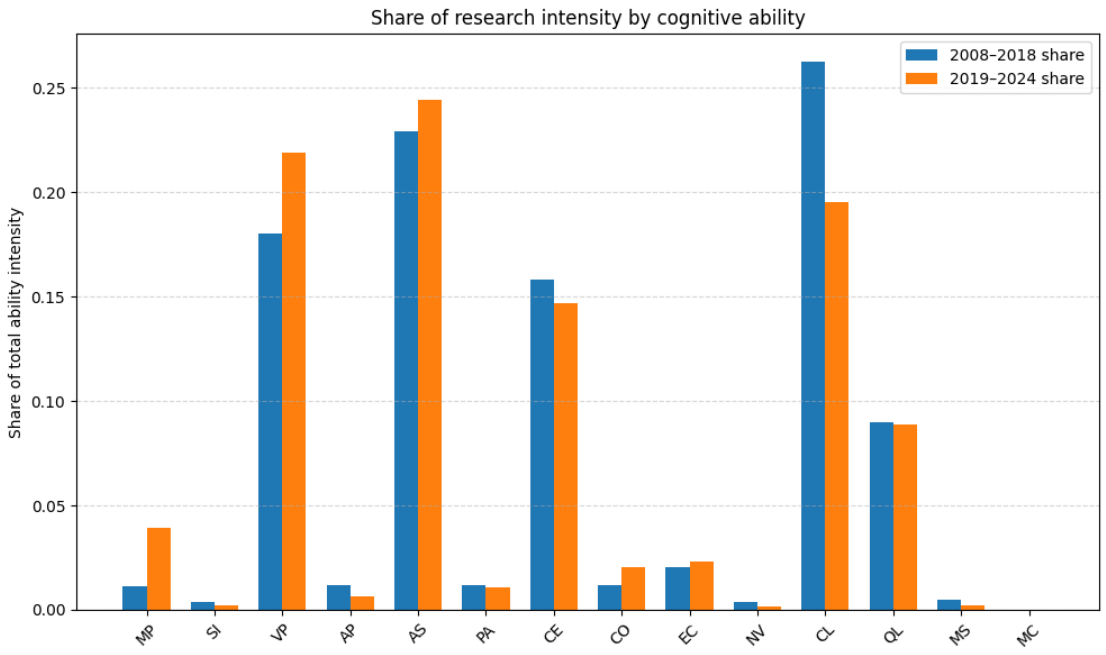

Figure 3: Share of total research intensity by cognitive ability in the pre-LLM and post-LLM periods.

Heatmap of ability shares

While line and bar charts show aggregated trends, a heatmap provides a more detailed year-by-year view. Figure 4 presents a two-dimensional heatmap of normalised ability shares. Darker colours correspond to larger shares. The sustained dominance of visual processing and the rise of memory and quantitative reasoning after 2019 are clearly visible. The heatmap also reveals that sensorimotor interaction and navigation intensities peak around 2012–2014 and decline thereafter, whereas pattern analysis and emotion comprehension maintain relatively constant low shares.

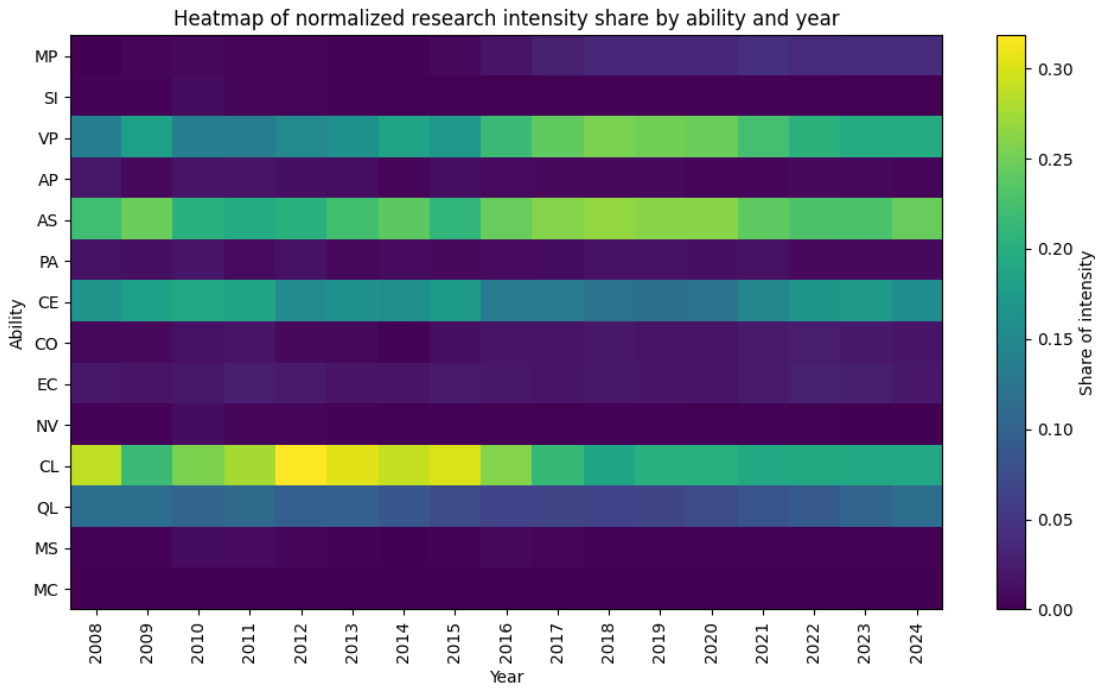

Figure 4: Heatmap of normalised research intensity share by ability (rows) and year (columns). Darker colours indicate higher shares.

Growth factors across abilities

Figure 5 plots the growth factor G_j for each ability. Bars above the dashed line (value 1) indicate abilities whose average intensity increased after 2019, while bars below suggest decline. Memory processes register the highest growth factor (approximately 2.5), reflecting their emergence as a core component of LLM architectures. Communication, emotion comprehension and quantitative reasoning show moderate growth. Conversely, sensorimotor interaction, navigation and motor skills have growth factors below 0.5, signifying relative decline in research emphasis on physical embodiment tasks.

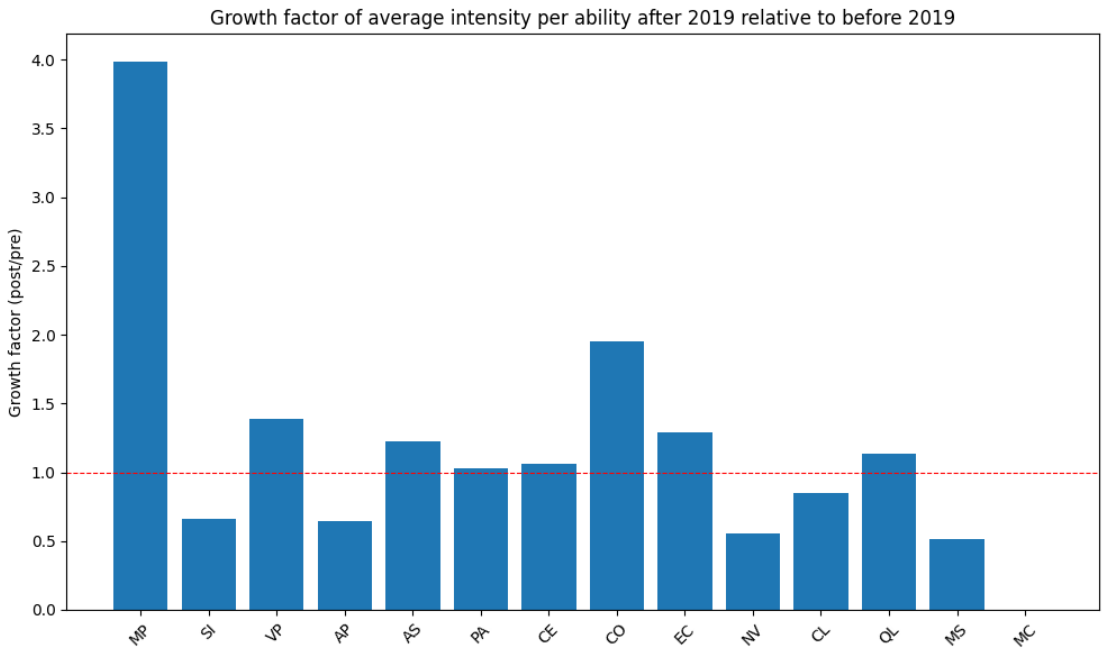

Figure 5: Growth factor of average research intensity per ability after 2019 relative to before 2019. The red dashed line marks unity (no change).

Top benchmarks and emerging tasks

Understanding which individual benchmarks dominate research can reveal concentration of effort. Figure 6 lists the ten benchmarks with the highest cumulative normalised intensity across 2008–2024. Classic vision datasets (ImageNet, COCO, KITTI) and language tasks (SQuAD, GLUE) continue to dominate, illustrating the inertia of established benchmarks. In contrast, Figure 7 shows the ten new benchmarks (introduced post-2019) with the highest cumulative intensity. Although their total contribution remains modest, we observe a mix of multi-modal (MMLU, MMBench), medical imaging (TCIA Pancreas CT) and robust face recognition (Disguised Faces in the Wild). This diversity indicates that the research community is exploring new domains rather than solely focusing on natural language processing.

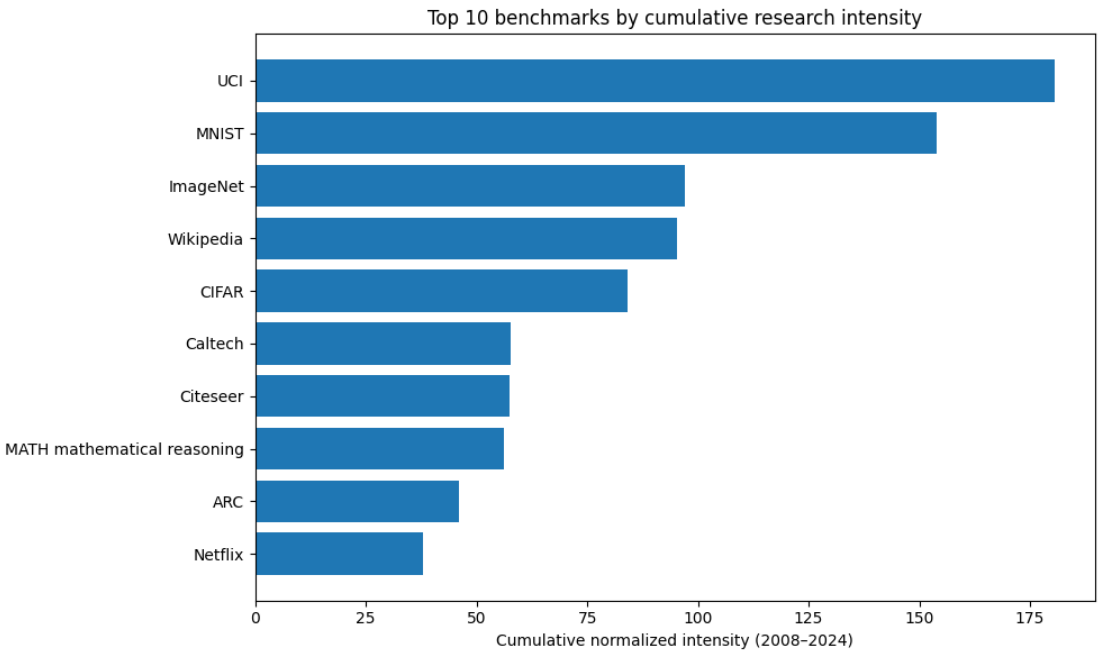

Figure 6: Top ten benchmarks by cumulative normalised research intensity (2008–2024).

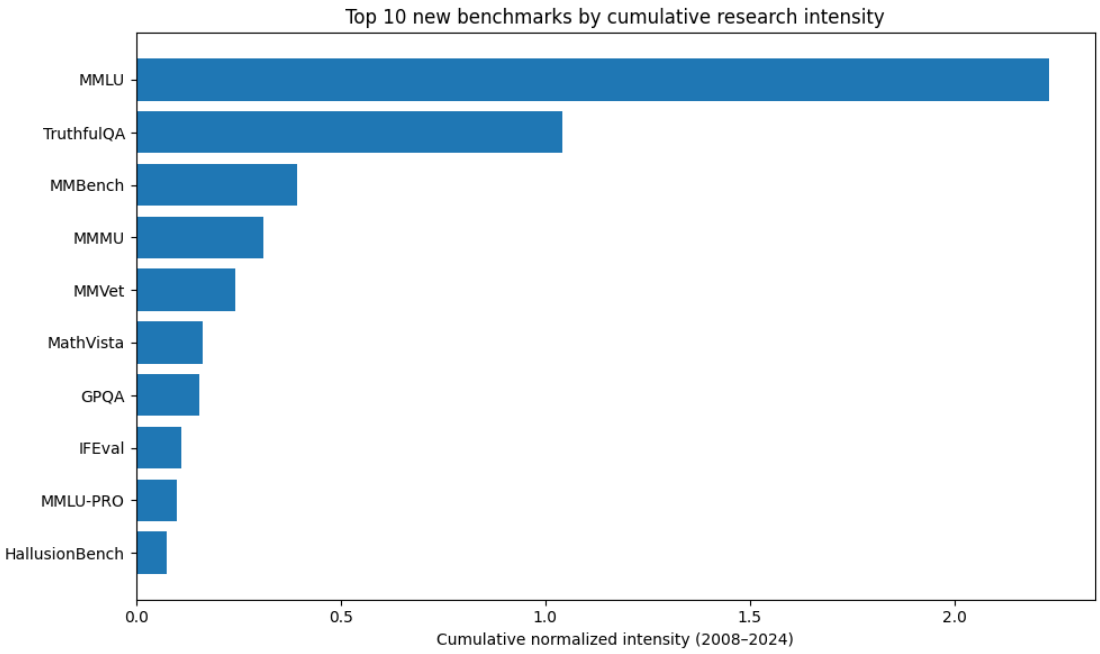

Figure 7: Top ten new benchmarks (introduced after 2019) by cumulative normalised research intensity.

Discussion

The analysis presented in this deliverable reveals several notable shifts in research focus across cognitive abilities. First, the dominance of visual processing persists but gradually recedes in relative terms, likely due to the maturity and saturation of vision benchmarks such as ImageNet and COCO. Second, memory processes, communication and quantitative reasoning show strong growth after 2019, aligning with the rise of LLMs and the need to handle long contexts, retrieve facts and generate coherent dialogue. Third, tasks associated with physical embodiment, including sensorimotor interaction, navigation and motor skills, decline relative to other abilities—possibly because research funding and attention have shifted towards language and multimodal modelling.

The emergence of 24 new benchmarks underscores the dynamism of the field. Many are designed to probe hallucination avoidance, multi-modal reasoning and medical image understanding, thereby addressing gaps in existing suites. Nevertheless, their cumulative intensity remains smaller than that of long-standing benchmarks. Ongoing monitoring will reveal whether they become established standards or suffer from rapid saturation. Our LLM-assisted annotation procedure proved effective for scaling benchmark mapping but also highlighted the absence of tasks explicitly targeting meta-cognition, which remains an open research challenge.

Conclusions

Deliverable D4 provides an illustrative analysis of research intensity across cognitive abilities, leveraging updated AITopics data through 2024 and introducing a curated set of new benchmarks. By normalising intensities, comparing pre- and post-LLM periods, computing growth factors and examining top benchmarks, we identify clear shifts in research emphasis: memory and communication abilities rise in prominence, while purely visual and sensorimotor tasks decline relatively. The dataset and figures serve as a resource for researchers and policymakers to anticipate which skills AI is advancing most rapidly and to design interventions accordingly. Future work could incorporate benchmark performance metrics, analyse cross-country differences in research activity, and develop benchmarks that explicitly target meta-cognition and other under-represented abilities.

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. “Research community dynamics behind popular AI benchmarks”. In: *Nature Machine Intelligence* 3.7 (2021), pp. 581–589.

[4] Ray Perrault and Jack Clark. “Artificial Intelligence Index Report 2024”. In: (2024).
[5] 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.
[6] Jason Wei et al. “Emergent abilities of large language models”.
In: *arXiv preprint arXiv:2206.07682* (2022).
[7] Wayne Xin Zhao et al. “A survey of large language models”.
In: *arXiv preprint arXiv:2303.18223* (2023).