# AI intensity dataset

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

*with expert number*
EX2018D335821
*for contract number*
CT-EX2018D335821-102
*regarding*
Deliverable 3:
AI intensity dataset

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

October 29, 2024

# D3: AI intensity 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 [15, 12] 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 [14]. 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 LLMs.

This third deliverable focuses on measuring the intensity of AI development, particularly in the context of progress in LLMs. By identifying and synthesising new and existing AI benchmarks, the project aims to provide a comprehensive overview of current research trends and to recalibrate measures of AI intensity. This will involve the use of extensive data collection techniques, including web scraping of scientific publications to map progress in AI capabilities and their impact on different occupational tasks. This output will lay the groundwork for assessing the impact of AI on the future workplace, providing valuable insights for researchers, policymakers and industry leaders to better understand and prepare for the evolving landscape of AI-driven occupational change.

## Introduction

This third deliverable of the AIM-WORK project focuses on measuring the intensity of AI development, with particular emphasis on advances in Large Language Models (LLMs) from 2020 to 2024. LLMs [2, 16, 4, 18], such as GPT-4 [1] and its successors, have significantly improved the ability of AI to process and generate human-like text, reshaping the landscape of AI capabilities and their applications. Consequently, understanding these advances and their implications for different occupational tasks is crucial for assessing the future impact of AI on the workplace.

Building on the foundation laid in the first deliverable[1], which categorised a comprehensive selection of AI benchmarks from PapersWithCode[2] (PwC) by modality, this report aims to explore into the intensity of AI research. The first deliverable provided an organised and structured overview of AI benchmarks, highlighting advances in AI capabilities across different data types such as text, images, audio and more. It established a baseline by identifying the top benchmarks per modality, based on the number of research papers addressing them.

In this third deliverable, we extend the initial findings by focusing specifically on measuring the *research intensity* or the activity level associated (in terms of production) with these

[1] https://github.com/nandomp/AIM-WORK-AI-impact/blob/main/JRC_Report_D1_Catalogue_of_AI_benchmarks.pdf

[2] https://paperswithcode.com

benchmarks per year, from 2020 to 2024. Our aim is to track the evolution of AI capabilities and correlate them with shifts in labour demand. To achieve this, we will employ data collection techniques, including web-scraping of open-access academic publications in the AI field . This comprehensive dataset will serve as the basis for updating and recalibrating the AI intensity measures introduced in the original 2021 framework by Songül Tolan and colleagues [15]. This analysis will also provide a more detailed understanding of the current trends in AI research, reflecting those areas of highest activity and innovation.

For readers seeking a thorough understanding of the methodologies and initial findings, we recommend referring to the first deliverable. This report will summarise those key points while extending the analysis to focus on the research intensity.

## AI progress

Evaluating progress in a particular AI discipline requires a focus on objective tools to measure the elements and objects of study, evaluate the prototypes and artefacts being developed, and examine the discipline as a whole [8]. Depending on the discipline and task, there is typically a loose set of criteria for how a system should be evaluated.

Several questions arise when trying to compare results or progress across different metrics: How do we compare results from different benchmarks for the same task (e.g., CIFAR-10 [9] vs. MNIST [7] vs. ImageNet [6]; see Figure 1) or from different tasks within the same benchmark? Even more challenging is the comparison of results from completely different tasks in different domains (e.g. image classification using ImageNet vs. language modelling using WikiText-103 [13]; see Figure 2).

While there may be a general perception of progress based on trends in evaluation metrics (increasing for accuracy or decreasing for error rates), it is misleading to assume that AI progress can be fully understood by looking at isolated tasks. There may be a lack of insight into the relationships between different tasks. For example, high performance in natural language processing tasks versus perception tasks cannot usually be easily integrated into a single agent with both perceptual and linguistic capabilities [3]. In addition, it is difficult to determine whether observed progress is due to improvements in hardware, data, computation, software, or the AI methods themselves [11].

Furthermore, the specialisation of many metrics to specific domains, overfitting of evaluations [17], and lack of continuity in some evaluation procedures are additional limitations in assessing AI progress [8]. Given these difficulties, **instead of using the rate of progress, we focus on the level of research activity around specific benchmarks**, which indicates the research intensity in a task by measuring the production outputs (e.g. research publications) from the AI community related to these benchmarks. Benchmarks with an increasing trend in production rates indicate greater research effort and intensity, which may eventually lead to progress. This has been observed in areas such as machine translation and object recognition, where high research intensity has led to undeniable progress and applications.

Overall, by tracking the research intensity of AI benchmarks, we may gain insights into the

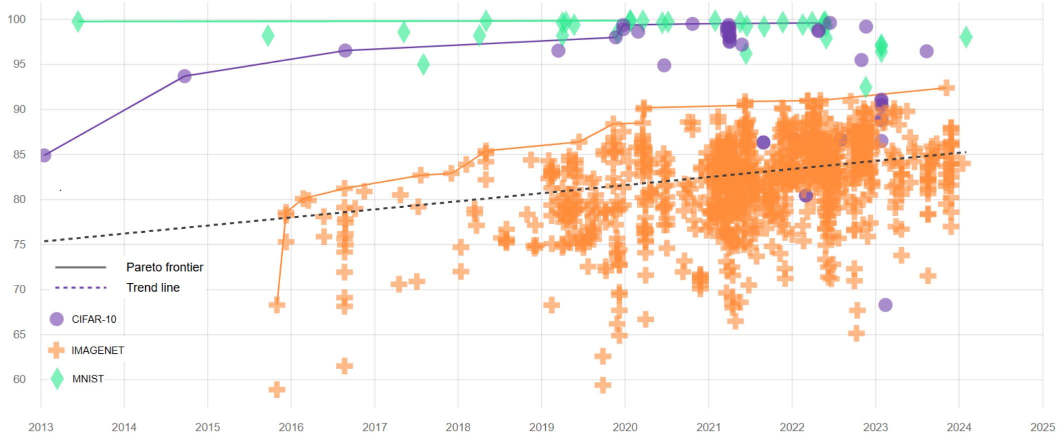

Figure 1: Top-1 accuracy results for AI systems addressing the task of Image Classification using three different benchmarks: CIFAR-10 (purple), ImageNet (orange) and MNIST (green). Plot generated using the AIcollaboratory framework [10].

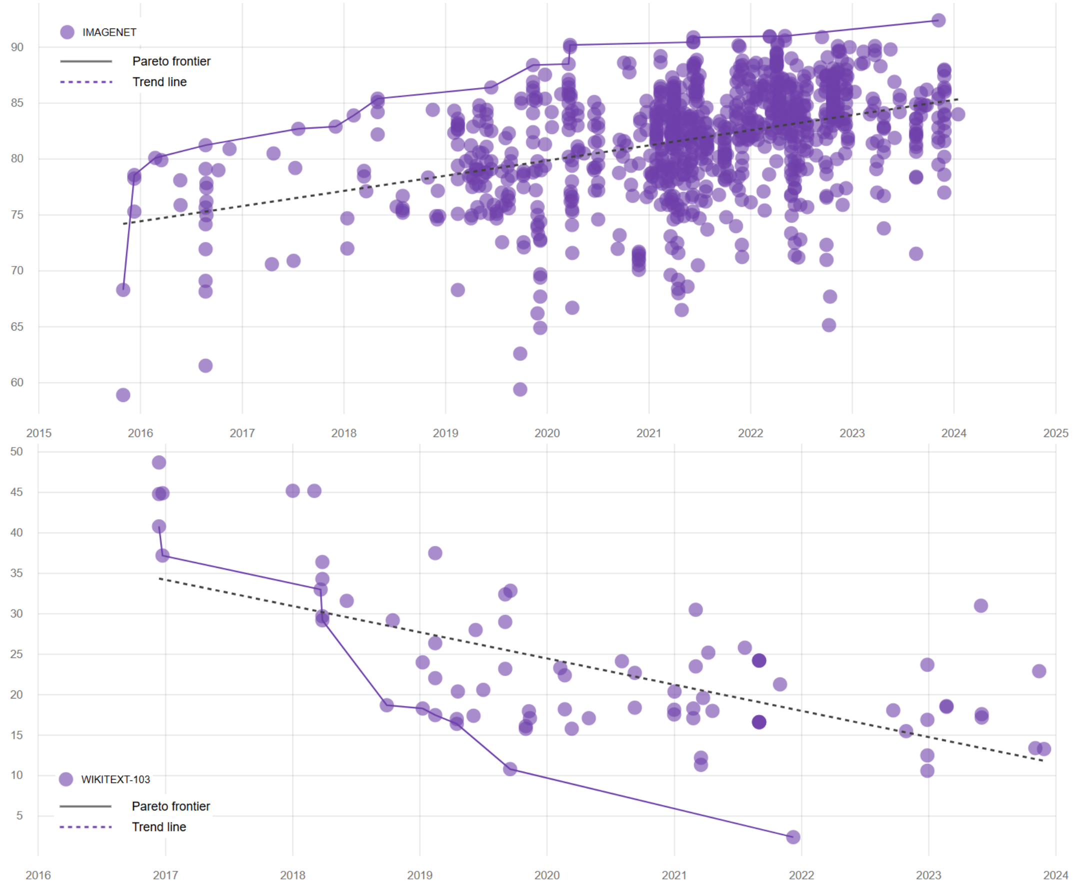

Figure 2: Results for AI systems addressing different tasks, benchmarks and evaluation metrics: (top) Image Classification and ImageNet, (bottom) Language Modelling and WikiText-103. Plot generated using the AIcollaboratory framework.

evolving landscape of AI research and identify areas where significant attention and resources are being directed.

## Methodology

### Scrapping Scientific Papers

To extract the relevant data for our analysis, we used web scraping techniques to obtain information from the PwC platform. Specifically, we wanted to collect details about the papers mentioning different AI benchmarks in different modalities. The following is an overview of the procedure, which was implemented using the R programming language [5] and relevant libraries.

The web scraping process started by accessing the PwC dataset website, which contains filters for different data modalities such as images, text, video and audio. By parsing the HTML content of this page, we identified the section representing these filters and extracted the modalities along with their respective counts. Libraries such as rvest[3] for reading HTML, httr[4] for working with HTTP and dplyr[5] for data manipulation were instrumental in this stage.

Each modality was handled separately. For each modality, the script navigated through the list of associated datasets and benchmarks, working page by page to cover all entries. The primary objective was to retrieve benchmark names, descriptions and the number of associated scientific papers and benchmarks (see D1[6] for further details). This required careful traversal of the DOM structure to locate and extract the required textual information and attributes from the relevant HTML nodes.

After collecting the initial data, we refined it by scraping the detailed page of each benchmark. This step involved retrieving time series data on the number of papers mentioning each benchmark over the years 2020 to 2024. The JSON data embedded in the page was parsed to extract these time series, detailing the intensity of research activity over the specified period. Finally, the data were normalised to facilitate comparisons between different benchmarks and modalities, allowing us to generate summary statistics and visualise trends.

### AI intensity

For each benchmark, we calculate intensity by averaging the normalised number of scientific documents per year over a given period of time. This results in a benchmark intensity vector with normalised values between 0 and 1.

To ensure a non-redundant dataset, we removed repeated entries where benchmarks appeared in more than one modality (e.g. images and text). In such cases, we retained the benchmark in the modality deemed most relevant, resulting in the removal of 258 duplicate entries (see Table 1 for the number of benchmarks per modality after this filtering). It is important to

[3] https://rvest.tidyverse.org/

[4] https://httr.r-lib.org/

[5] https://dplyr.tidyverse.org/

[6] https://github.com/nandomp/AIM-WORK-AI-impact/blob/main/JRC_Report_D1_Catalogue_of_AI_benchmarks.pdf

| **Modality** | **#Benchmarks** | **Modality** | **#Benchmarks** | **Modality** | **#Benchmarks** |
|---|---|---|---|---|---|
| Images | 2831 | Biomedical | 10 | Physics | 11 |
| Texts | 2269 | LiDAR | 19 | Dialog | 1 |
| Videos | 501 | RGB Video | 7 | Midi | 5 |
| Audio | 256 | Tracking | 9 | 6D | 2 |
| Medical | 102 | Biology | 11 | Ranking | 4 |
| 3D | 143 | 3d meshes | 5 | Replay data | 3 |
| Graphs | 188 | Actions | 3 | Interactive | 3 |
| Time series | 139 | Stereo | 1 | Financial | 1 |
| Tabular | 104 | EEG | 28 | Parallel | 1 |
| Speech | 96 | Tables | 10 | | |
| RGB-D | 35 | Hyperspectral images | 13 | | |
| Environment | 82 | Music | 9 | | |
| Point cloud | 22 | MRI | 7 | | |

Table 1: Modalities and their respective number of benchmarks (datasets) after filtering repeated benchmark entries having more than one modaliy asigned.

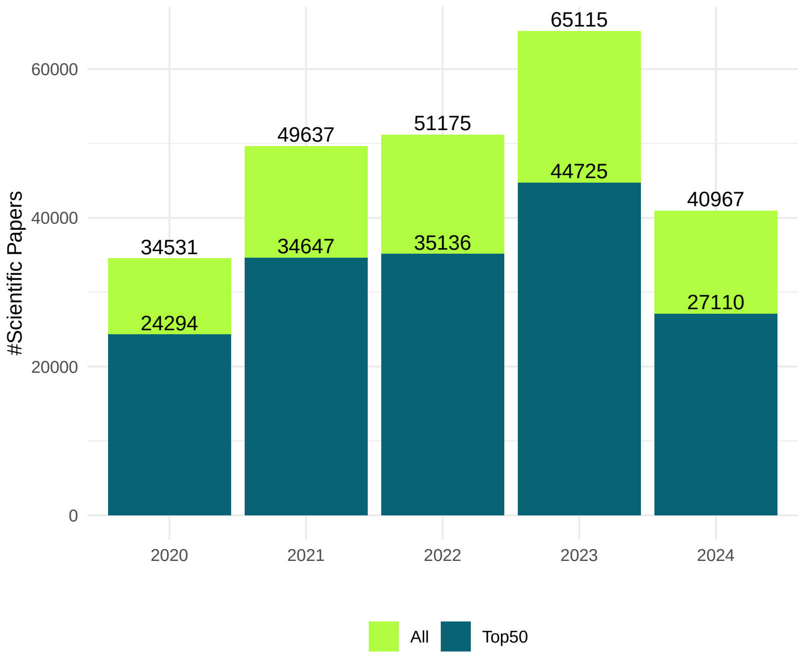

Figure 3: Increasing number of papers mentioning each benchmark over the years 2020 to 2024, for all the benchmarks in PwC (”All”) and the Top-50 per modality (”Top50”). Data for 2024 are incomplete (extracted as of August 2024).

note that this procedure has an impact on the analysis of intensity trends when analysed by modality. However, it does not affect our understanding of the impact of AI advances on occupational tasks, as the modality is primarily used as a means of organising benchmarks. Furthermore, we focus on the top 50 benchmarks per modality, or the maximum number available if there are fewer than 50 benchmarks for a given modality, resulting in a total of 1,493 benchmarks. This selection strategy ensures that we focus on the most influential and widely recognised benchmarks within the AI research community, as these top benchmarks are frequently cited and used in numerous studies, providing a robust basis for our analysis.

All the data can be downloaded from: `https://github.com/nandomp/AIM-WORK-AI-impact`.

## Data Overview

Below is a preliminary overview of the assessment of the interest of benchmarks aggregated by modality. Figure 9 shows the intensity trends for different AI modalities from 2020 to 2024 sorted by total number of papers addressing the benchmarks in the modality. Each subplot shows the normalised level of research activity (y-axis) for a given modality over the years.

High activity modalities such as **Image**, **Text** and **Video** show a generally stable intensity, with **Text** showing a notable peak in 2024. The **Audio** modality shows a stable intensity with a slight upward trend, while **Medical** shows a steady increase, indicating growing research interest in AI applications in healthcare. Moderately active modalities such as **3D**, **Graphics**, **Time-series**, **Tabular** and **Speech** maintain relatively stable research intensities, reflecting their continued relevance in various AI domains.

In addition, modalities such as **RGB-D**, **Environment**, **Point-Cloud**, **Biomedical**, and **LiDAR** show stable or slightly increasing trends, indicating continued and expanding research efforts in these areas. On the other hand, some modalities show fluctuating or more pronounced changes in intensity. For example, **Tracking** and **RGB Video** show moderate but steady growth, while Biology shows a sharp peak in 2023, indicating a burst of research activity. **3D Meshe**s and **Actions** also show increasing trends towards 2024. Fluctuating trends are observed in modalities such as **Stereo**, **EEG**, **Tables** and **Music**, with periods of higher intensity followed by declines, reflecting a potential shift in research focus. **Interactive** and **Physics** show peaks in certain years, suggesting episodic interest. Less active modalities such as **Financial**, **Cad**, **Parallel**, **Lyrics** and **PSG** show relatively low and stable intensities.

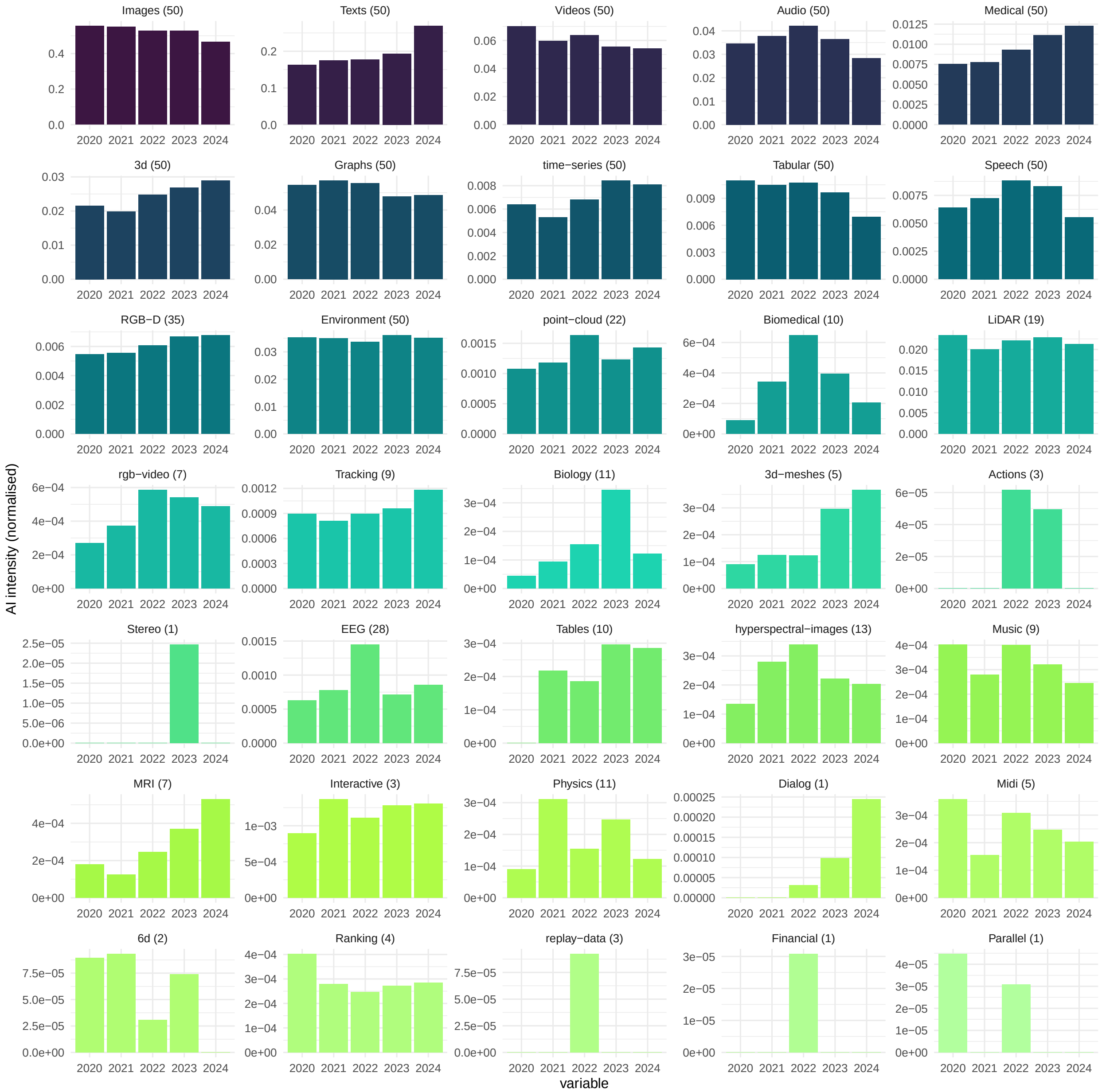

Figure 4: Evolving landscape of AI research intensity across modalities. Top 50 benchmarks per modality as selected (sorted by number of papers). Note that some modalities have very few benchmarks, so the results are not conclusive.

For illustrative purposes, the results are broken down by modality below. We focus on the top 50 benchmarks for the five most prominent modalities: images, text, video, audio and medical. These modalities were selected based on their importance and impact within the AI research community. The rest can be found in the repository.

## Images

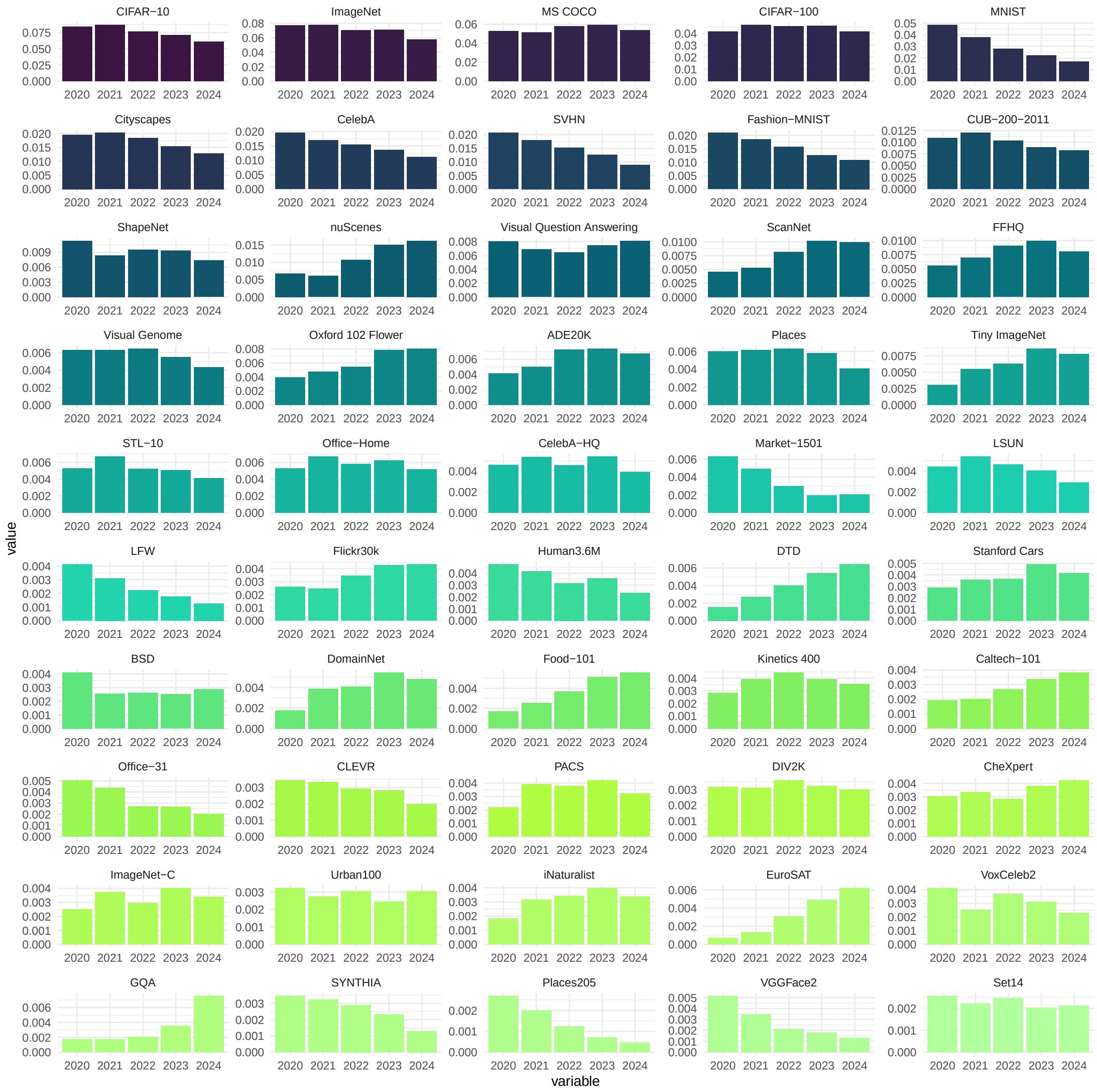

Figure 5: Evolving landscape of AI research intensity across modalities.

## Texts

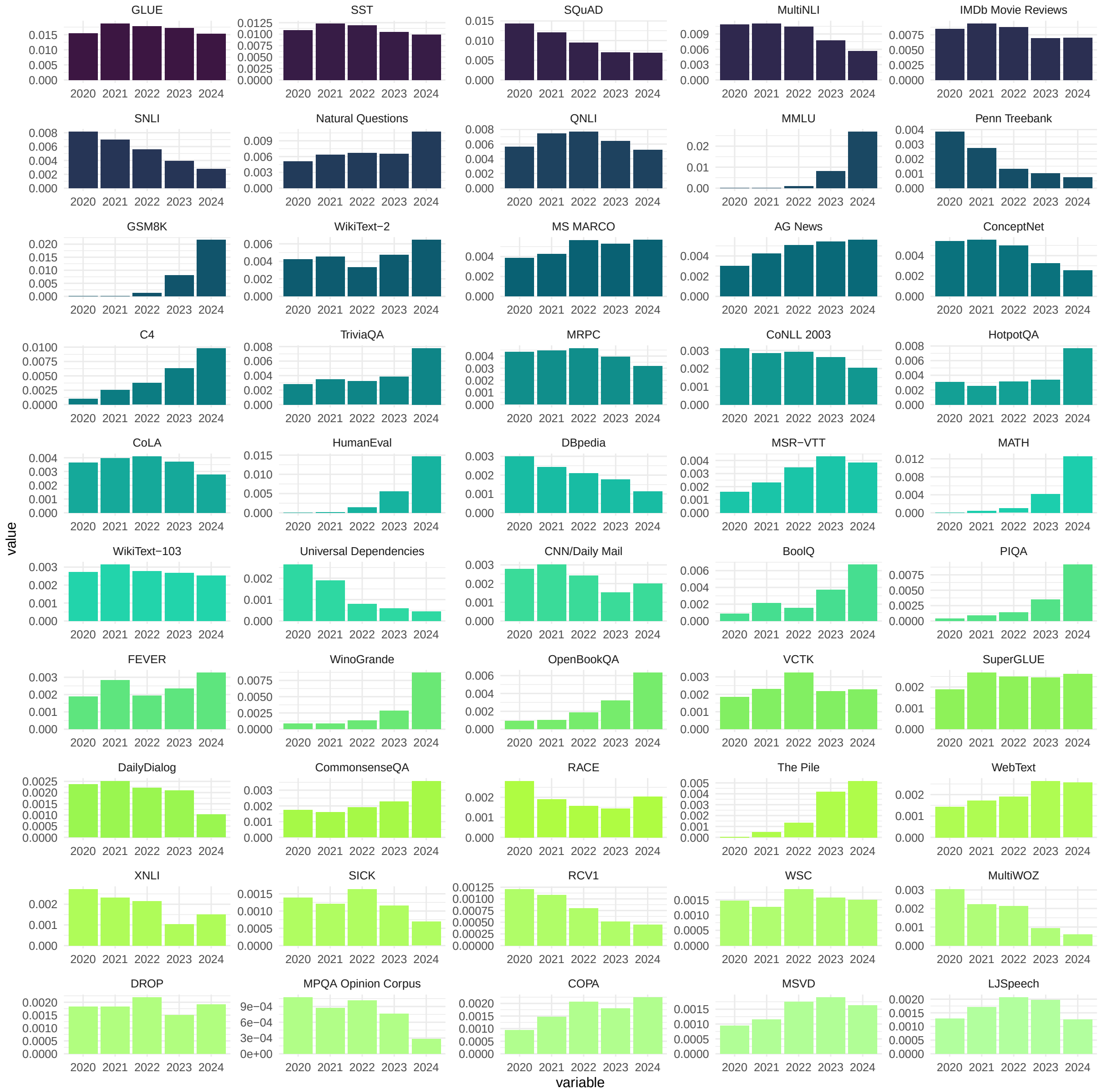

Figure 6: Evolving landscape of AI research intensity across modalities.

## Videos

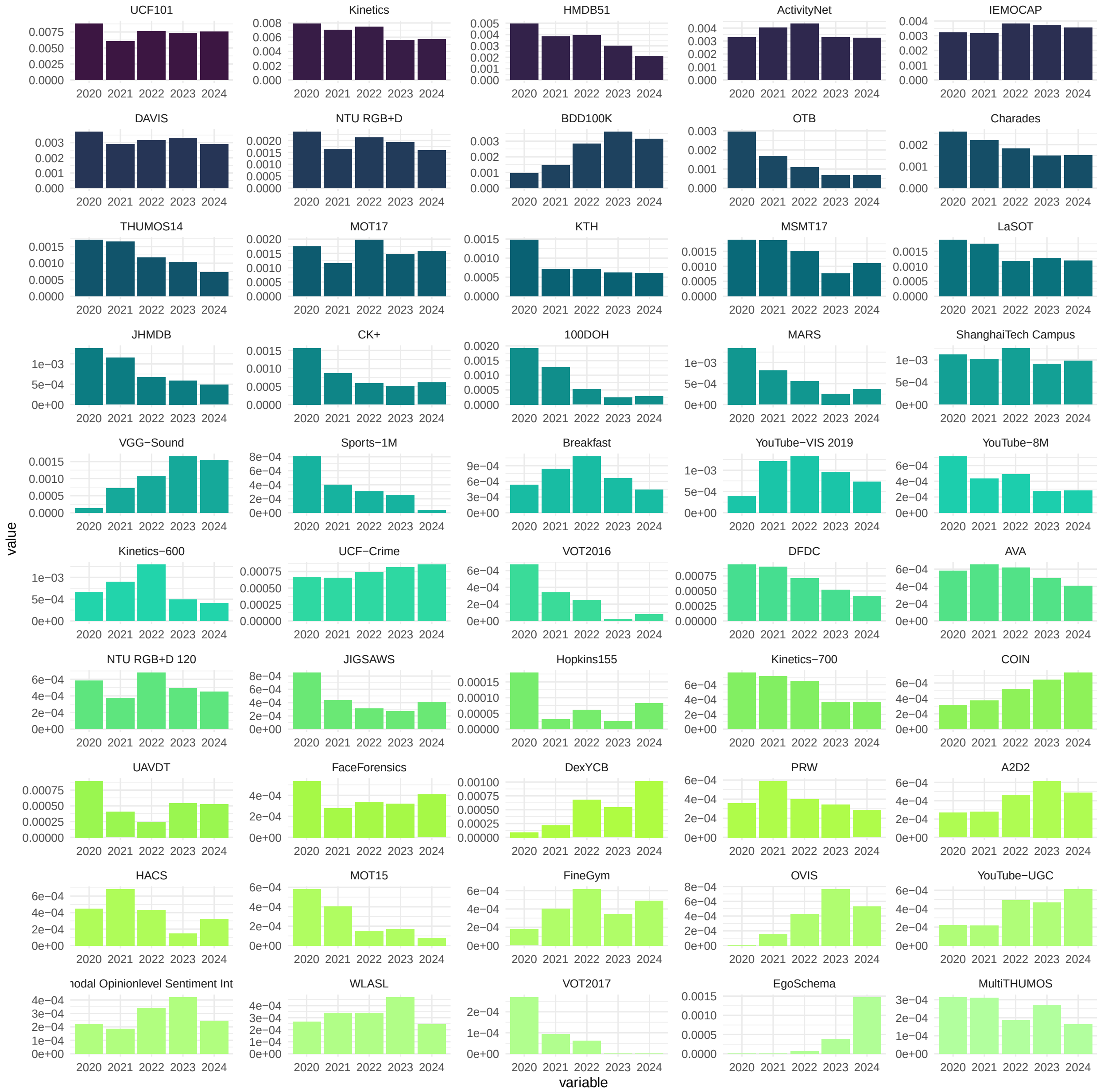

Figure 7: Evolving landscape of AI research intensity across modalities.

# Audio

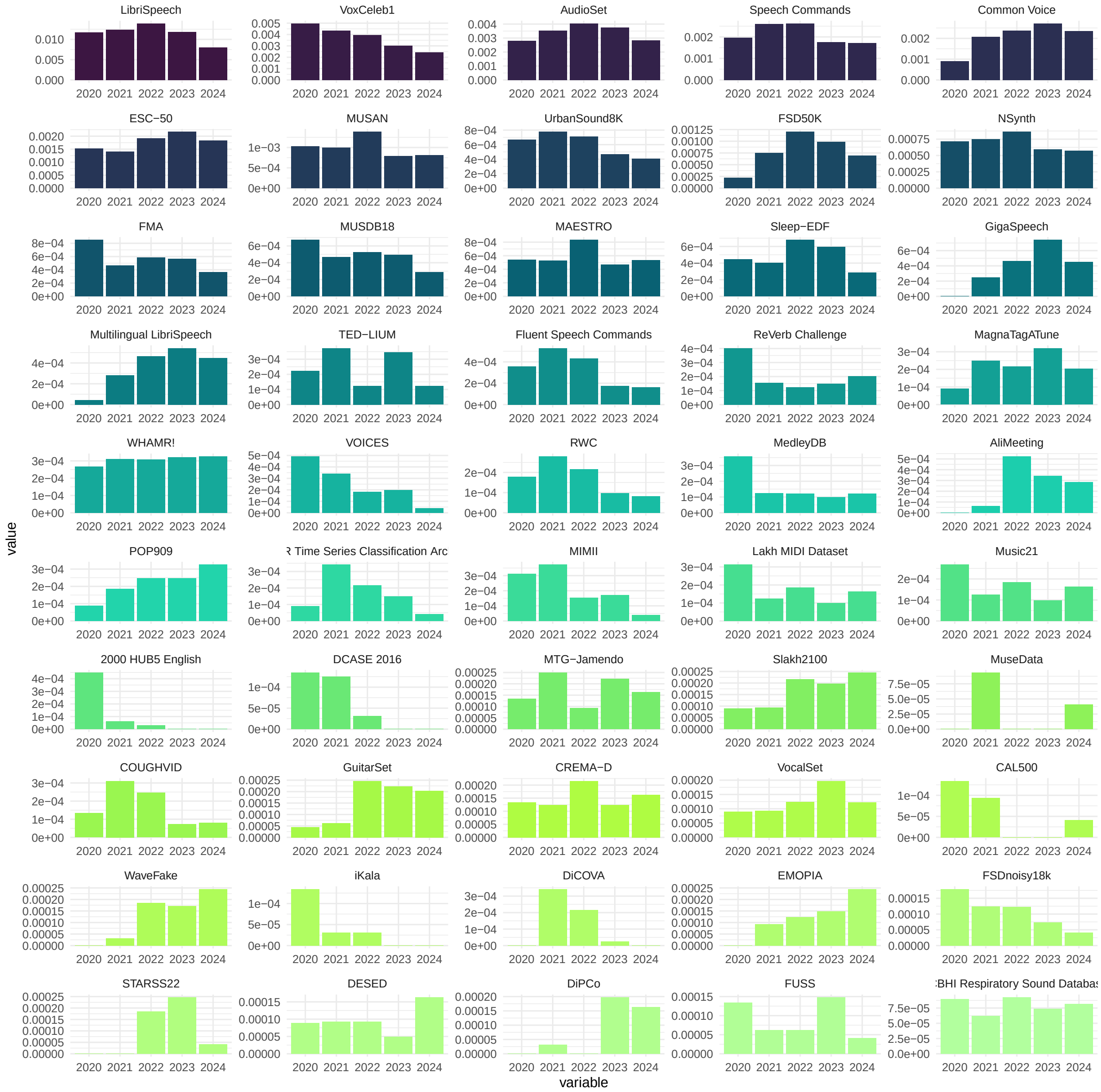

Figure 8: Evolving landscape of AI research intensity across modalities.

## Medical

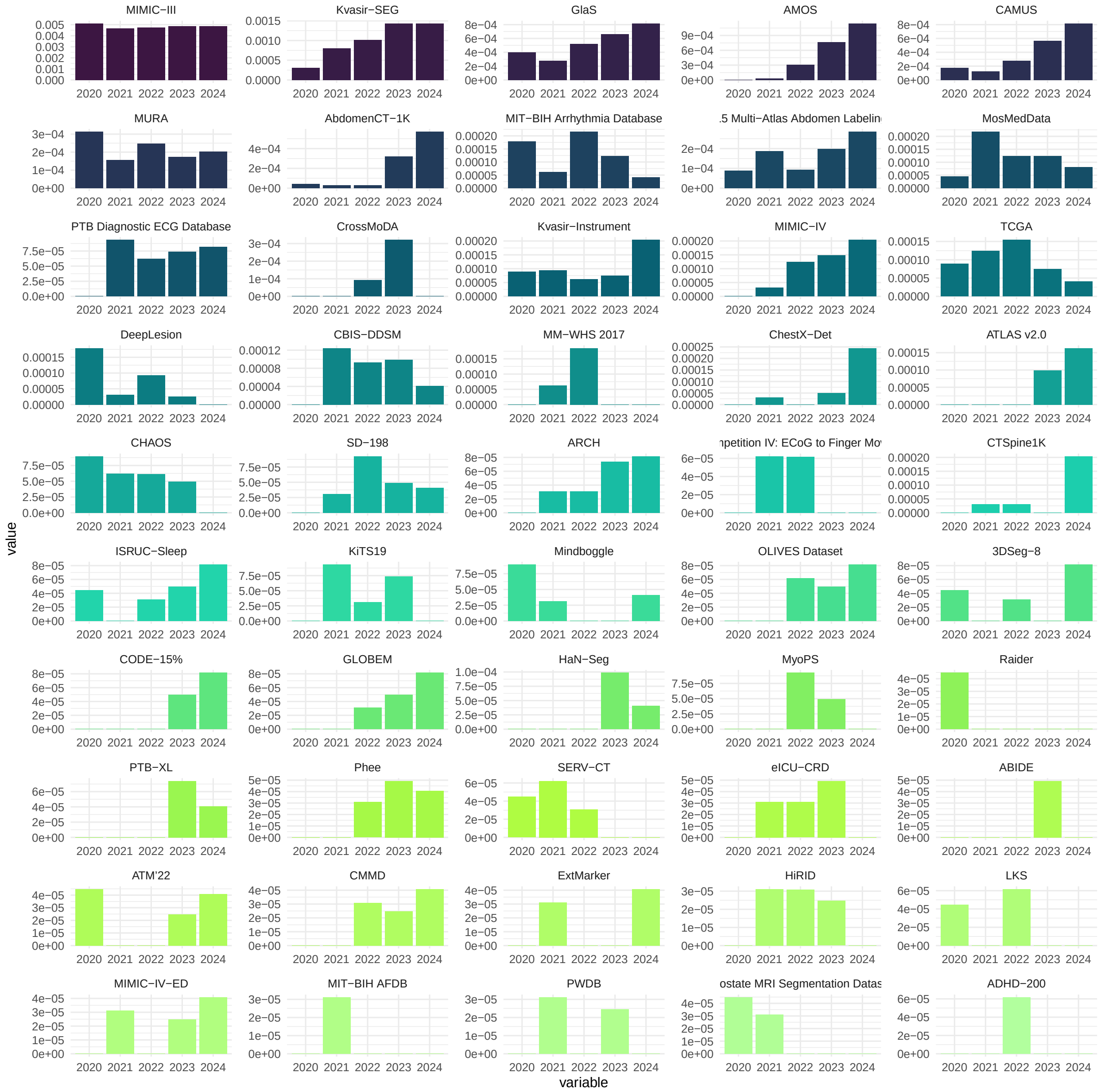

Figure 9: Evolving landscape of AI research intensity across modalities.

## Conclusions

This deliverable provides a comprehensive analysis of AI research intensity from 2020 to 2024. Using web scraping techniques on the PwC platform, we collected extensive data on AI benchmarks across different modalities, revealing key trends in the progress of AI research. Our findings indicate stable and growing research activity in high-impact areas such as text, image, and medical AI applications, while other modalities show varying patterns of research focus over the years.

The dataset and the insights it generates serve as a critical resource for understanding the evolving AI landscape, providing valuable guidance for researchers, policymakers, and industry leaders. By tracking research intensity, we not only identify the areas that are attracting significant attention, but also set the stage for future research into the impact of AI on the workplace. The trends captured in this report will inform strategic decisions and help prepare for the AI-driven transformation of work roles. The full dataset is available for further exploration at https://github.com/nandomp/AIM-WORK-AI-impact.

Going forward, our project will involve a thorough mapping and annotation of the selected benchmarks to the specific cognitive skills they are designed to assess. We will use a combination of automated tools, such as GPT-4, and manual review processes to ensure the accuracy and completeness of this mapping. By systematically categorising these benchmarks and identifying gaps in current AI capabilities, we aim to guide targeted research and development efforts.

# References

[1] Josh Achiam et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

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

[3] M. Brundage. "Modeling Progress in AI". In: *AAAI 2016 Workshop on AI, Ethics, and Society* (2016).

[4] 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].

[5] Michael J Crawley. *The R book*. John Wiley & Sons, 2012.

[6] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[7] Li Deng. "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[8] José Hernández-Orallo. "Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement". In: *Artificial Intelligence Review* 48.3 (2017), pp. 397–447.

[9] Alex Krizhevsky. *Learning multiple layers of features from tiny images.* https://www.cs.toronto.edu/~kriz/cifar.html. 2009.

[10] Fernando Martínez-Plumed, Jose Hernández-Orallo, and Emilia Gómez. "Tracking AI: The capability is (not) near". In: *ECAI 2020*. IOS Press, 2020, pp. 2915–2916.

[11] Fernando Martínez-Plumed et al. "Between Progress and Potential Impact of AI: the Neglected Dimensions". In: *arXiv preprint arXiv:1806.00610* (2018).

[12] 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.

[13] Stephen Merity et al. “Pointer sentinel mixture models”.
In: *arXiv preprint arXiv:1609.07843* (2016).

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

[15] 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.

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

[17] S. Whiteson et al.
“Protecting against evaluation overfitting in empirical reinforcement learning”.
In: *Adaptive Dynamic Programming And Reinforcement Learning (ADPRL), 2011 IEEE Symposium on*. IEEE. 2011, pp. 120–127.

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