Catalogue of AI benchmarks

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

with expert number
EX2018D335821
for contract number
CT-EX2018D335821-102
regarding
Deliverable 1:
Catalogue of AI benchmarks

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

August 27, 2024

D1: Catalogue of AI benchmarks

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 [25, 20] 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 [23]. 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 first deliverable focuses on cataloging new and existing AI benchmarks to capture the current state of AI research. By identifying and synthesising these benchmarks, the project aims to map the progress in AI capabilities and their implications for various occupational tasks, laying the groundwork for assessing AI's impact on the future workplace.

Introduction

The rapid development of Artificial Intelligence (AI) and Machine Learning (ML) technologies has led to tremendous advances in various domains [22], especially with the emergence of Large Language Models (LLMs)[3, 26, 7, 27]. The AIM-WORK project aims to explore the transformative impact of these innovations on workplace activities and the wider professional environment. As part of this project, our first deliverable presents a comprehensive catalogue of AI benchmarks, which are crucial for measuring the capabilities and progress of AI models in different tasks.

AI benchmarks [21] are essentially standardised assessments designed to evaluate the performance of AI systems on specific tasks, similar to how academic tests evaluate student performance. They play a critical role in guiding and accelerating AI research and development by providing a clear, objective basis for comparison. Consistent benchmarking ensures that progress is not only tracked and measured, but also strategically directed to address gaps and improve existing capabilities.

This deliverable categorises AI benchmarks by input data modality, referring to the different types of data that can be processed and generated by AI systems, including text, image, video, audio, and more [1]. This approach simplifies the complex landscape of AI evaluation into more manageable segments, allowing for focused analysis and better insights into AI's strengths and limitations in handling different types of data. By establishing clear categories, we mitigate the complications that can arise from classifications based on tasks, languages or applications, all of which can be more convoluted and hinder straightforward analysis.

Using data from the widely recognised PapersWithCode[1] (PwC) platform, we present a

[1] https://paperswithcode.com/

curated list of AI benchmarks for each modality (around 10K in total). In our future analysis we will perform a selection of benchmarks per modality based on the volume of research papers addressing these benchmarks, highlighting those that are the most influential and widely recognised in the AI research community. This ensures that the benchmarks we focus on are both relevant and reflective of the broader research trends and innovations within each modality.

The importance of this deliverable goes beyond simply cataloguing benchmarks. By providing a structured overview of the current state of AI capabilities across different data types, we lay the groundwork for deeper analysis and understanding of how these advances may reshape occupational tasks and the future workplace.

In the following sections, we will go through the methodology followed to gather all the data and detailed descriptions of these benchmarks, categorised by modality, also providing a preliminary analysis to delineate the research interest and activity surrounding each benchmark.

What are AI Benchmarks?

AI benchmarks are standardised tests used to evaluate the performance of AI systems on specific tasks. These benchmarks serve as a way to measure and compare the capabilities of different AI models, providing a common basis for assessing their strengths and weaknesses. Think of them as 'report cards' for AI systems, providing insight into how well these systems perform in areas such as language understanding, image recognition and more.

For example, in text processing, an AI benchmark might involve understanding and generating human language, such as answering questions or translating text. In image processing, benchmarks often test an AI's ability to recognise objects or faces in images. By providing a consistent way to measure performance across tasks and domains, AI benchmarks play a key role in ensuring that AI technologies advance in a reliable and predictable way. This helps both scientific research and the practical application of AI in real-world scenarios.

The Role of AI Benchmarks

AI benchmarks play a crucial role in the development and evolution of artificial intelligence systems. They **set performance standards** that AI models must meet or exceed, much like a set of academic criteria. For example, a benchmark for natural language processing might involve complex tasks such as answering questions based on a given text, with the AI's performance rigorously compared to human responses.

This benchmarking process not only establishes clear performance criteria, but also **drives research and innovation** by highlighting gaps in current AI capabilities. When benchmarks reveal shortcomings, such as an AI's difficulty in distinguishing between similar objects in image recognition tasks, researchers and developers are directed to those specific areas for improvement. Benchmarks also **enable objective comparisons** between different AI models and approaches, providing a transparent view of which methodologies are most effective and why.

The insights gained from these benchmarks do not just stop at identifying current strengths and weaknesses. They actively **guide the development of new models and algorithms**. For example, if a benchmark reveals an AI's poor understanding of context in language processing, developers can focus on refining that particular skill. In this way, the continuous cycle of benchmarking, evaluation, and refinement accelerates technological progress and ensures that AI systems evolve to meet increasingly sophisticated standards.

Through benchmarking, the AI community can collectively push the boundaries of what's possible, ensuring that progress is not only tracked and measured, but also strategically directed for future innovation [21, 16].

AI Benchmarking Platforms

In AI there is often a loosely defined set of criteria for evaluating systems, which poses several challenges. How do we compare results on the same benchmark using different metrics? How do we compare different benchmarks designed for similar tasks, such as COCO [13], MNIST [4], and ImageNet [9], or different tasks within the same benchmark? Moreover, how do we evaluate the results of completely different tasks in different domains? These challenges highlight the limited value of analysing AI solely through the performance of specific systems on specific tasks, leading to a fragmented understanding of AI capabilities.

This issue is further compounded by the distinction between specialised and general AI systems [18, 12]. Specialised systems are optimised for specific tasks, while general systems are designed to handle a wide range of tasks. With the advent of general AI systems, such as the GPT series [5], which demonstrate broad capabilities across multiple tasks, it becomes even more important to have a consistent and comprehensive benchmarking framework.

Given these complexities, it is clear that ad hoc evaluation methods are insufficient. Instead, we need comprehensive, standardised platforms facilitate cross-benchmark, cross-domain and cross-metric comparisons. In this regard, AI benchmarking platforms play a key role in standardising the evaluation of AI systems, providing a basis for comparison and driving progress through consistent and transparent performance metrics. These platforms offer diverse datasets and tasks, allowing researchers and developers to systematically assess the capabilities and limitations of their models [11, 15].

Below, we outline several notable AI benchmarking platforms.

- **PapersWithCode**[2]:Recognised as the largest, most up-to-date and open repository of machine learning papers and their experimental results [2], Papers with Code allows users to explore AI research through various metrics and leaderboards. It serves as an invaluable resource for tracking state-of-the-art performance on various AI tasks, and provides a comprehensive view of research progress.
- **OpenML**[3]: OpenML is an online platform for sharing and organising data, machine learning algorithms, experiments and results from predictive tasks. Unlike other repos-

[2] https://www.paperswithcode.com/
[3] https://www.openml.org/

itories, OpenML's database includes instance-level results, providing a granular view of model performance.

- **AI Index Annual Report**[4]: Published annually, the AI Index report visualises and analyses data on AI trends and progress. Designed for a wide audience, including policymakers, researchers, executives, journalists and the general public, it provides a comprehensive overview to help understand the multifaceted landscape of AI.
- **AI Collaboratory**[5]: Developed as part of the European Commission's *AI Watch* project[6], the AI Collaboratory serves as a collaborative platform for the evaluation, comparison and classification of AI systems [19, 17]. It fosters collaboration between AI researchers and institutions to improve the benchmarking process.
- **BigBench**[7]: The Beyond the Imitation Game Benchmark (BIG-bench) is a collaborative effort to evaluate the capabilities and predict the future performance of large language models (LLMs). With over 200 tasks, BIG-bench provides a comprehensive evaluation framework for language models [24].
- **HELM**[8]: The Holistic Evaluation of Language Models (HELM) aims to improve transparency in the evaluation of language models and foundational models. Using a holistic evaluation approach, HELM evaluates models on multiple dimensions such as accuracy, calibration, robustness, fairness, bias, toxicity and efficiency [14].
- **LM Evaluation Harness**[9]: This project provides a unified framework for evaluating causal language models across different NLP tasks and model frameworks. It emphasises reproducibility through a standard interface and task versioning, facilitating rapid evaluation and comparison of new LLMs with previous studies.
- **Mosaic Model Gauntlet**[10]: Developed by MosaicML, this project introduces a taxonomy for benchmarking pre-trained models against 34 different tests, organised into six competencies. Using data from LLM publications and open source frameworks such as EleutherAI's Eval Harness and Stanford CRFM's HELM, it provides a comprehensive evaluation landscape.

It is important to note that the above list represents some of the key platforms contributing to AI benchmarking, but is by no means exhaustive. Each initiative offers different insights and tools for assessing the state of AI, providing valuable resources for researchers, developers and policymakers to navigate and advance the field. Throughout this deliverable, we extensively use data from the PWC platform to catalogue AI benchmarks across different modalities, ensuring that our selection is both relevant and reflects the latest research trends.

[4] https://hai.stanford.edu/ai-index-2021
[5] https://ai-collaboratory.jrc.ec.europa.eu/
[6] https://knowledge4policy.ec.europa.eu/ai-watch_en
[7] https://github.com/google/BIG-bench
[8] https://crfm.stanford.edu/helm/latest/
[9] https://www.eleuther.ai/projects/large-language-model-evaluation
[10] https://www.mosaicml.com/llm-evaluation

Methodology

A robust and systematic methodology was used to accurately catalogue and evaluate the current state of AI research. This section outlines the procedures and tools used to extract, categorise and analyse AI benchmarks. By using data from the PwC platform, we ensured that the benchmarks selected for our study were both comprehensive and reflective of current trends in AI development. The detailed steps below describe our approach to identifying modalities, extracting relevant data and categorising benchmarks to facilitate meaningful comparative analysis.

Extraction and categorisation of AI benchmarks

As a first phase of the project to analyse the impact of AI on the labour market, we extracted detailed data on AI benchmarks from the PwC platform. As stated above, this platform serves as a comprehensive repository providing insights into the most influential benchmarks within the AI research community. Note that the exact number of most relevant benchmarks to focus on for each modality in our subsequent analyses, such as the top 10%, top 50, etc., is still to be determined. In general, the selection needs to be based on the number of research papers addressing each benchmark, ensuring that our catalogue highlights widely recognised and influential benchmarks. See table 1 for a summary of all modalities provided and the total number of papers in the literature that address them.

Modality	**#Benchmarks**	**Modality**	**#Benchmarks**	**Modality**	**#Benchmarks**
Images	2826	Biomedical	105	Physics	25
Texts	2731	LiDAR	80	Dialog	23
Videos	913	RGB Video	71	Midi	19
Audio	418	Tracking	61	6D	13
Medical	348	Biology	56	Ranking	10
3D	330	3d meshes	55	Replay data	9
Graphs	253	Actions	52	fMRI	8
Time series	226	Stereo	42	Financial	7
Tabular	189	EEG	39	Cad	6
Speech	178	Tables	39	Parallel	6
RGB-D	170	Hyperspectral images	38	Lyrics	2
Environment	129	Music	37	PSG	2
Point cloud	114	MRI	33		

Table 1: Modalities and their respective number of benchmarks (datasets)

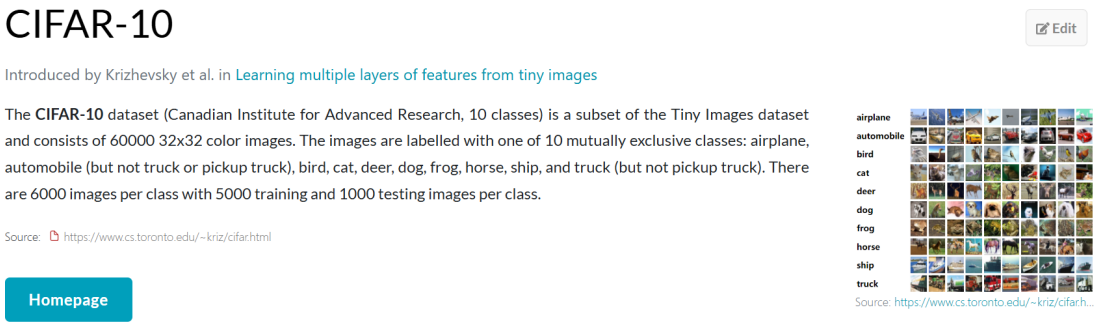

Figure 1: Specific benchmark (CIFAR-10) and its description in PwC.

Apart from the total number of benchmarks (the datasets, see Figure 1) for each modality, we also gather the **number of papers**, and the **number of tasks**. The former refers to the approximate number of research publications that mention or use the benchmarks (see Figure 2. This metric serves as an indicator of research interest and activity around a particular benchmark. A high number of publications indicates that a benchmark is widely recognised and frequently used by the research community, reflecting its importance and relevance in advancing the field.

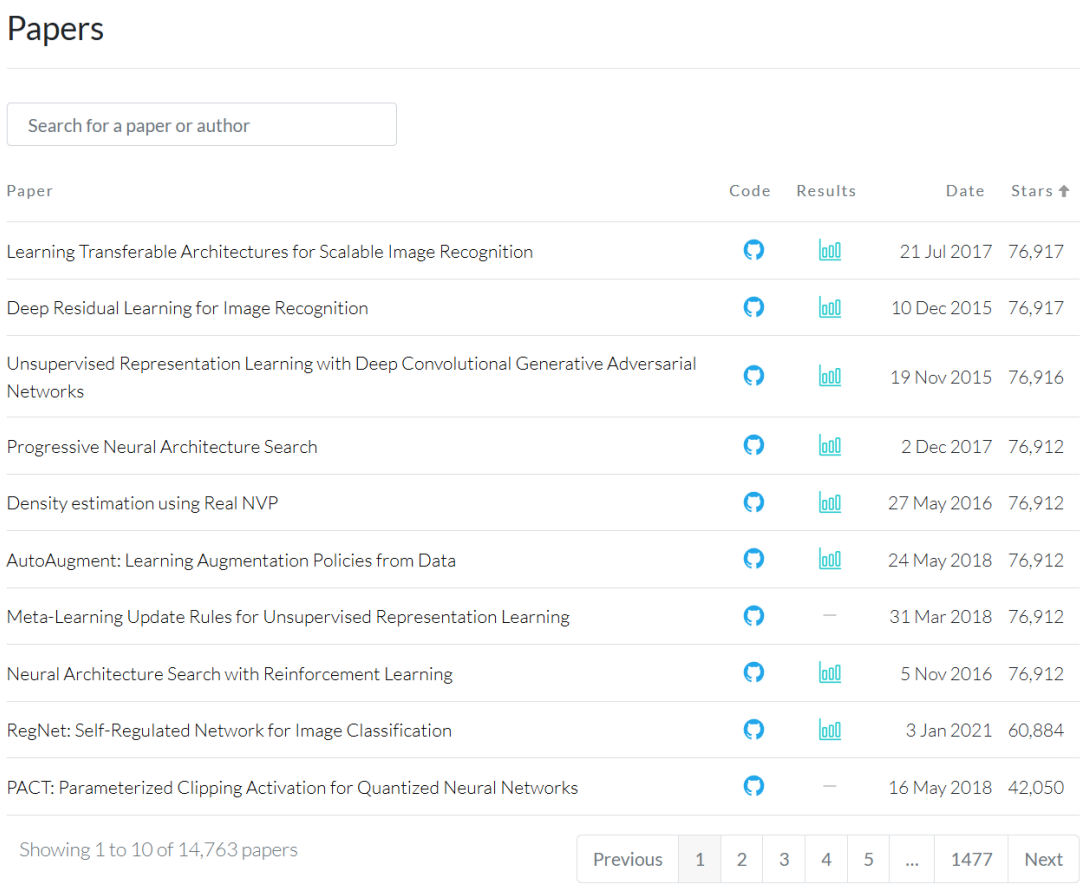

Figure 2: List of papers referencing or using specific benchmarks. The figure shows a list of papers addressing CIFAR-10.

The latter includes the various AI tasks and leaderboards that use a particular benchmark (see Figure 3). Tasks refer to the specific challenges or problems that researchers are trying to solve using a benchmark. Each task often has its own leaderboard, where different models are ranked according to their performance. For example, tasks in the text modality might include sentiment analysis, machine translation, or text summarisation. A benchmark with a large number of tasks indicates its versatility and broad applicability to different problem domains.

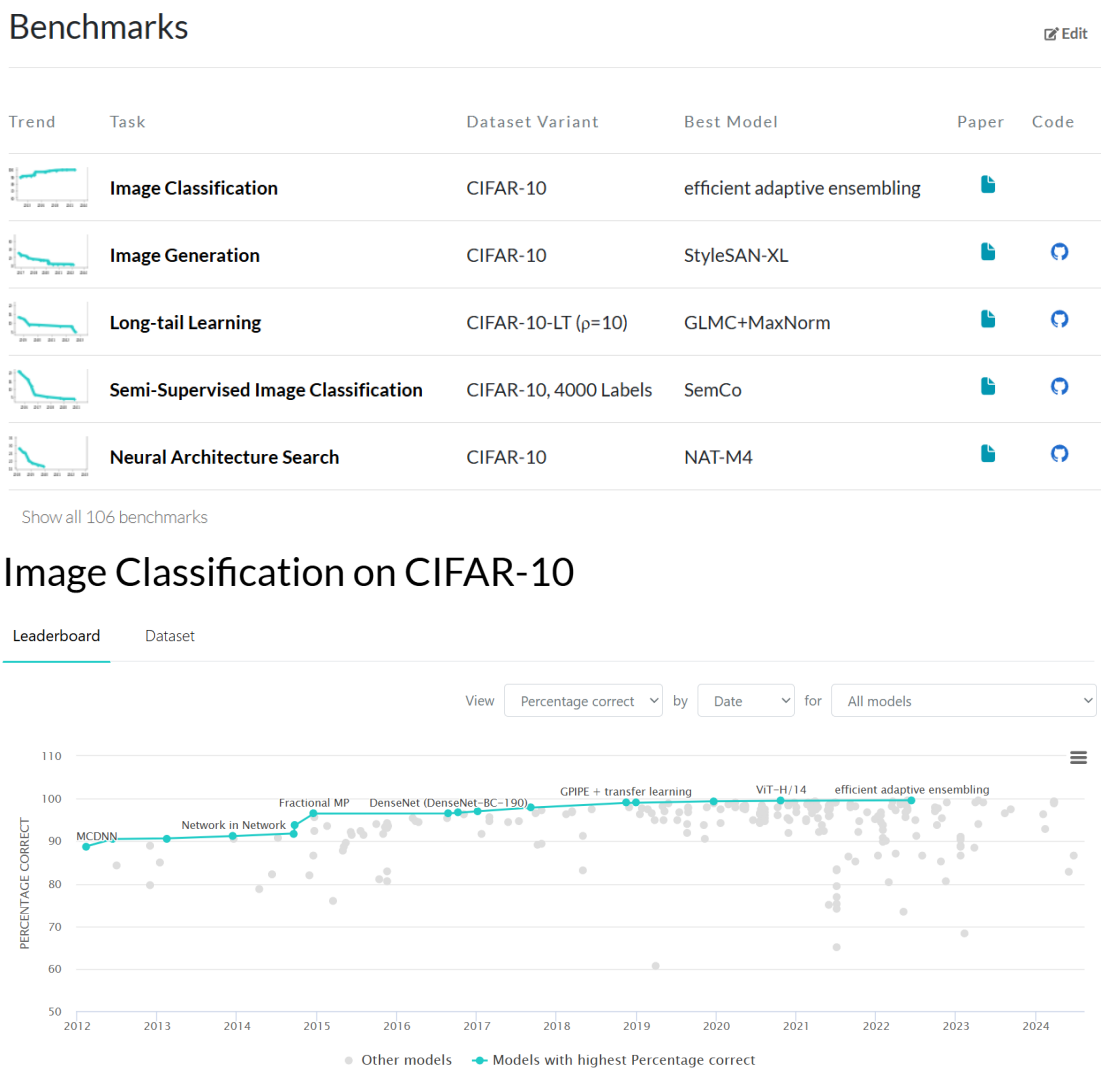

Figure 3: Number of tasks and leaderboards associated with each benchmark for different AI modalities. The figure shows these details for CIFAR-10.

To extract the data on benchmarks, number of papers and number of tasks, we followed a systematic web-scraping procedure. This process involved visiting the PwC website, navigating to the relevant pages for each AI modality, extracting the required information and saving it for further analysis. The following steps describe the methodology in detail:

- **Identify modalities**: Create a dataframe containing different AI modalities (e.g. text, images, audio) and the number of benchmarks associated with each.

- **Scrape data**: For each modality, navigate to the corresponding web page containing the list of data sets (benchmarks). Then parse the web page to extract key information about each dataset, including its name, description, the number of papers mentioning or using it, and the number of tasks associated with it.
- **Store the data**: Store the extracted data in a structured format (e.g. dataframe) for later analysis and visualisation.

Several R libraries [8] are used to perform the web scraping procedure. The rvest[11] library serves as the primary tool for web scraping, allowing HTML content to be read and information to be extracted from web pages. Key functions within rvest include read_html(), html_node(), html_nodes(), html_text(), and html_attr(). The stringr[12] library is used for string manipulation and extraction tasks with functions like str_replace_all() and str_split_fixed(). For data manipulation, the dplyr[13] library provides a consistent grammar for data frame operations, including bind_rows() for combining data frames, mutate() for adding variables, and select() for selecting columns. The httr[14] library, which facilitates working with HTTP, is used to download and read HTML content from URLs using the GET() function.

Data Modalities

Data modalities [1] refer to the different types of data that AI systems can process and generate, and are critical to understanding the breadth and scope of generative AI applications. These modalities include text, image, audio, video, and other kinds of data, each of which has unique characteristics and requires specialised techniques to handle effectively. In the context of generative AI [10] , text data is used for tasks such as content creation and customer support, using natural language processing [6] to generate coherent and meaningful written material. Image data includes algorithms that can create high-quality images from text descriptions, helping industries such as design and entertainment. Audio includes the generation of realistic sounds and speech, with applications in virtual assistants, audio books and music composition. Video data, which combines the complexities of image and audio, is used for dynamic content generation, facilitating media creation and enhancement. Finally, we can also find multimodal AI systems whoch are capable of processing and generating data across multiple modalities simultaneously, enabling rich applications such as interactive virtual assistants and enriched multimedia experiences.

Categorising benchmarks by data modality rather than other potentially more complex and convoluted schemes, such as task or language, provides a clear and organised framework for understanding AI capabilities. Each modality involves different techniques, algorithms and evaluation metrics. By focusing on modality, we can more easily track progress, address specific challenges, and apply insights across domains within the same type of data. This straightforward categorisation simplifies analysis, and helps readers quickly grasp the state of AI progress across different data types and tasks.

[11] https://rvest.tidyverse.org/
[12] https://stringr.tidyverse.org/
[13] https://dplyr.tidyverse.org/
[14] https://httr.r-lib.org/

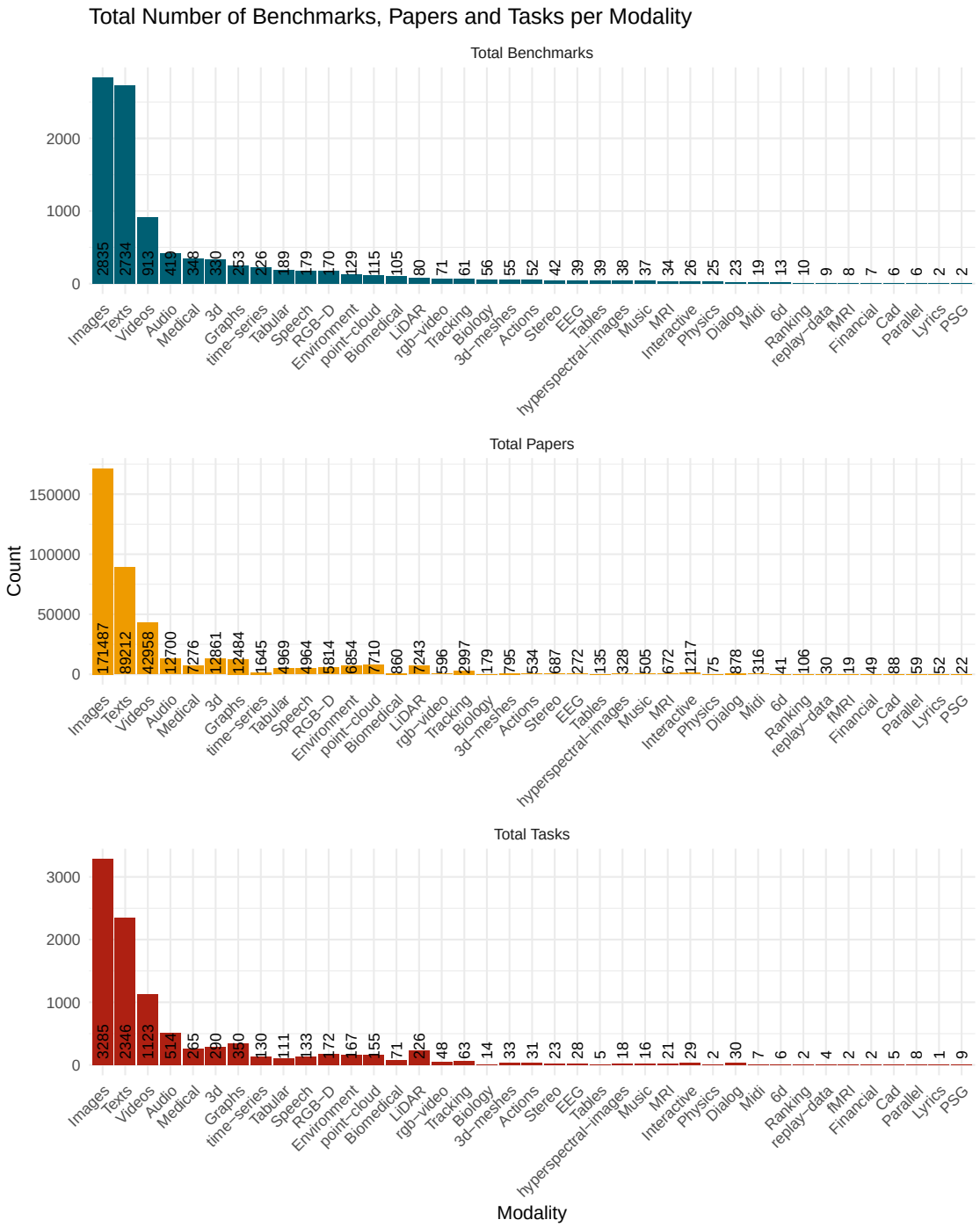

Figure 4: The figure provides an overview of the total number of benchmarks, papers and tasks per modality. It highlights the focus of AI research by showing the distribution of benchmarks, papers and tasks across different data types. In particular, the image and text modalities dominate across all three categories, indicating significant research activity and development.

Figure 4 illustrates the distribution of benchmarks, papers, and tasks across the defined data modalities (from PwC) in Table 1, providing a comprehensive overview of the AI research landscape, identifying focal points of innovation and revealing the breadth of tasks within each modality. The top plot shows the total number of benchmarks available for each modality, emphasising significant variation among them. For instance, images and texts have the highest number of benchmarks, indicating extensive research attention and development in these areas. The middle plot highlights the total number of papers published for each modality, reinforcing the trend seen in benchmarks, with images and texts again leading significantly. The bottom plot presents the total number of tasks associated with each modality, demonstrating that the focus areas with the most benchmarks and papers also tend to have the greatest variety of tasks.

Data

Below we provide a detailed overview of the benchmarks organised by modality, highlighting the number of research papers addressing each of the top 50 benchmarks. The following subsections describe the key characteristics and relevance of AI benchmarks within different modalities, illustrated with histograms that show their respective research interest.

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

Texts

Text benchmarks evaluate AI systems on various natural language processing (NLP) tasks such as language understanding, text generation, sentiment analysis, and machine translation. These benchmarks assess how well AI can process, interpret and generate human language in written form.

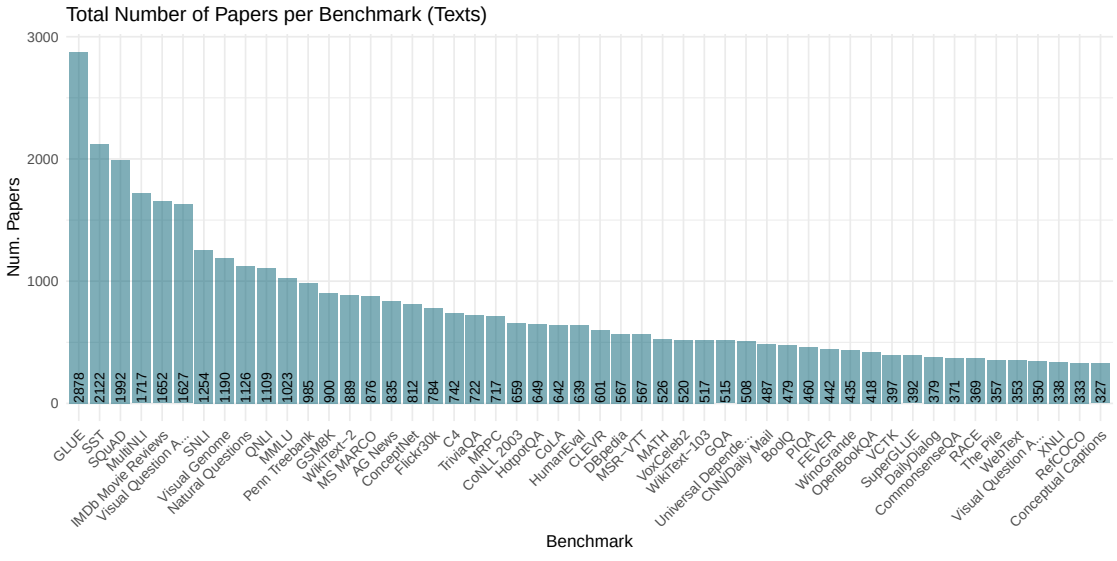

Figure 5: Histogram showing the top-50 benchmarks in the Text modality per number of papers addressing them.

Images

Image benchmarks focus on computer vision tasks, including object detection, image classification, and face recognition. They measure the ability of AI systems to understand and interpret visual data from static images.

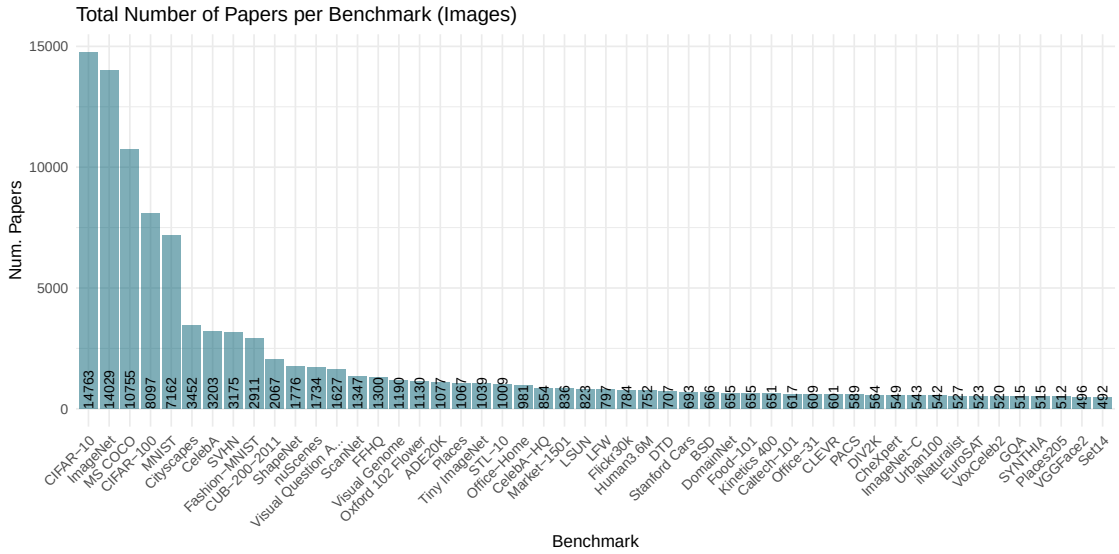

Figure 6: Histogram showing the top-50 benchmarks in the Images modality per number of papers addressing them.

Videos

Video benchmarks evaluate the performance of AI in tasks such as video classification, activity detection, and object tracking within video sequences. These benchmarks test how well AI can analyse and make sense of dynamic visual data.

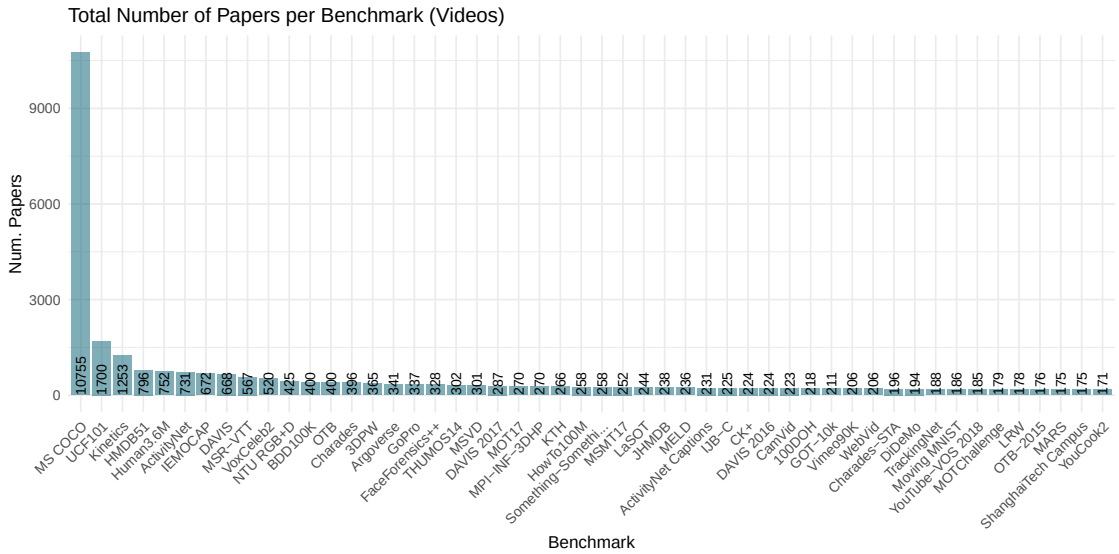

Figure 7: Histogram showing the top-50 benchmarks in the Videos modality per number of papers addressing them.

Audio

Audio benchmarks assess AI's ability to process and understand audio data, including tasks like speech recognition, sound classification, and music generation. They test the capability of AI systems to interpret and generate auditory information.

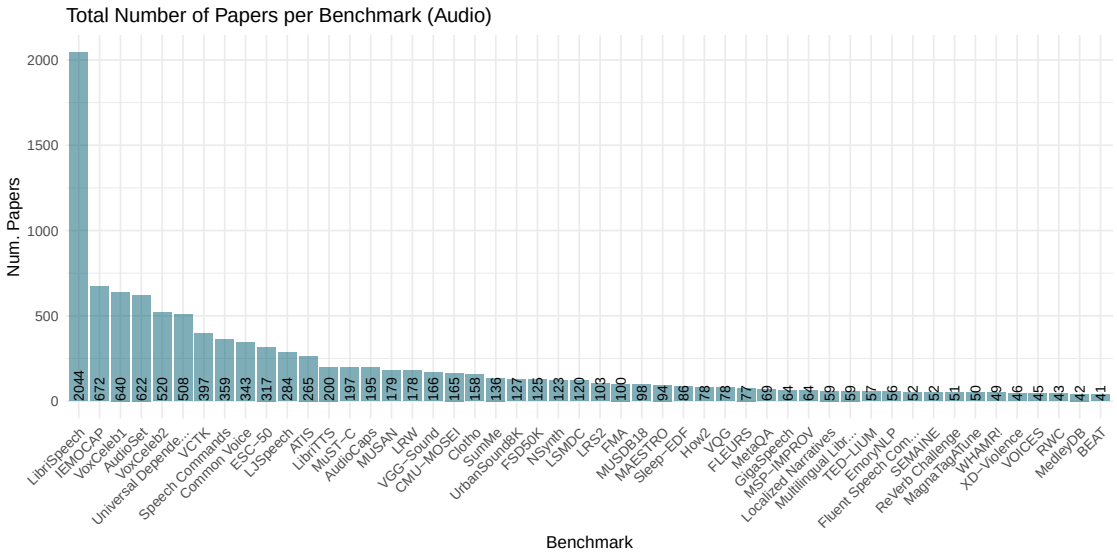

Figure 8: Histogram showing the top-50 benchmarks in the Audio modality per number of papers addressing them.

Medical

Medical benchmarks involve tasks such as medical image analysis, disease prediction, and healthcare data interpretation. These benchmarks evaluate how well AI can assist in medical and healthcare domains by processing various medical data types.

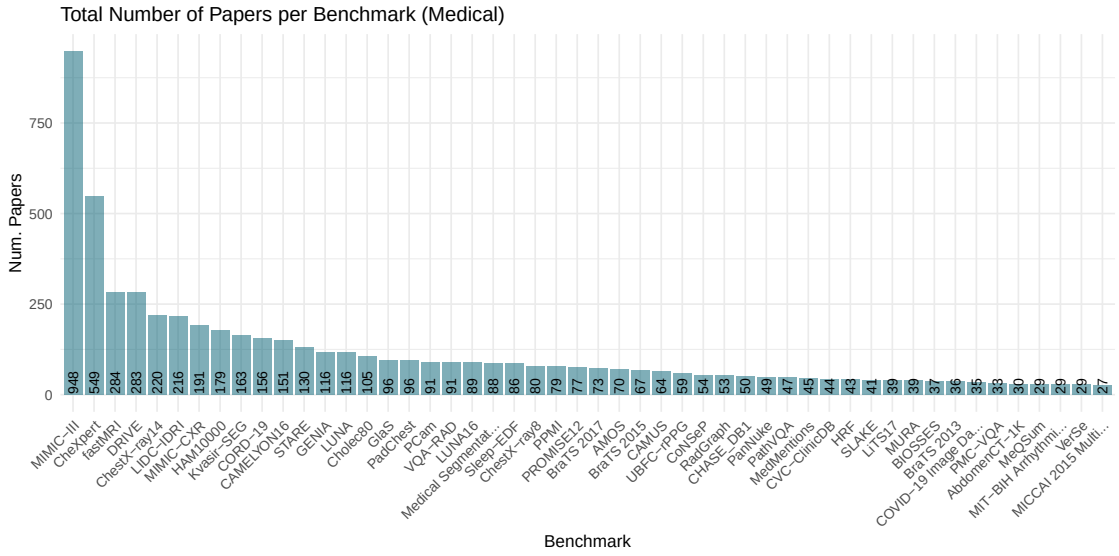

Figure 9: Histogram showing the top-50 benchmarks in the Medical modality per number of papers addressing them.

3D

3D benchmarks focus on tasks involving three-dimensional data, such as 3D object recognition, reconstruction, and segmentation. These benchmarks test AI's capability to understand and manipulate 3D structures and environments.

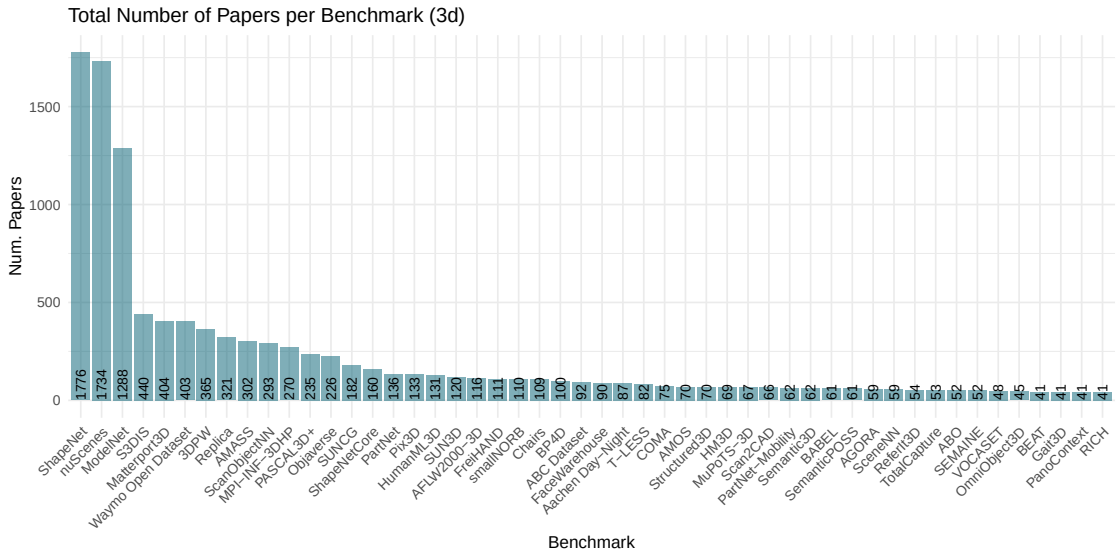

Figure 10: Histogram showing the top-50 benchmarks in the 3D modality per number of papers addressing them.

Graphs

Graph benchmarks assess AI systems' ability to perform tasks related to graph-structured data, like node classification, link prediction, and graph clustering. These benchmarks are crucial for applications in network analysis, social media, and biological data.

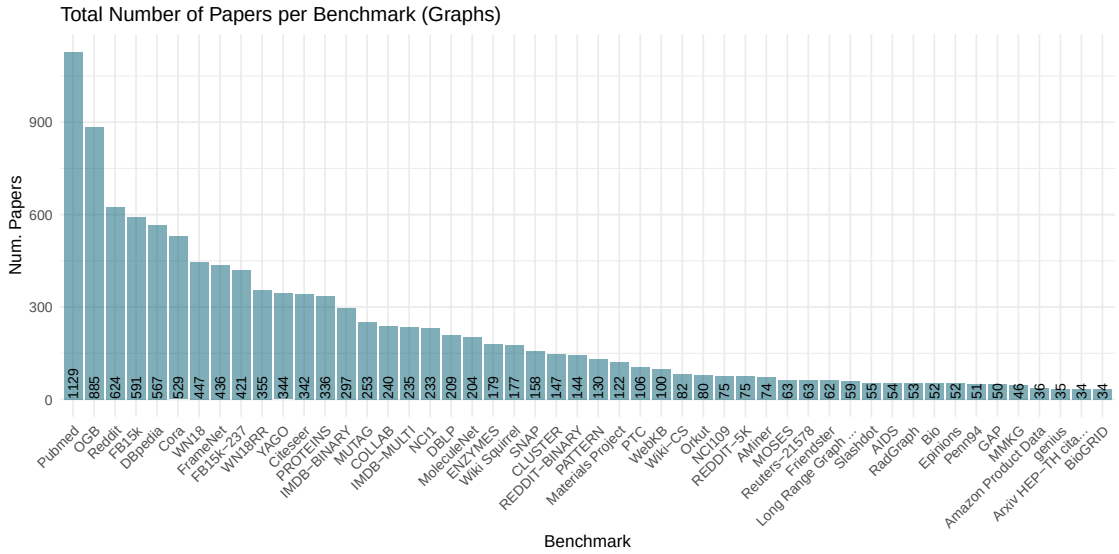

Figure 11: Histogram showing the top-50 benchmarks in the Graphs modality per number of papers addressing them.

Time Series

Time series benchmarks evaluate AI's performance on tasks involving sequential data over time, such as forecasting, anomaly detection, and temporal pattern recognition. These benchmarks are particularly relevant in finance, weather prediction, and IoT.

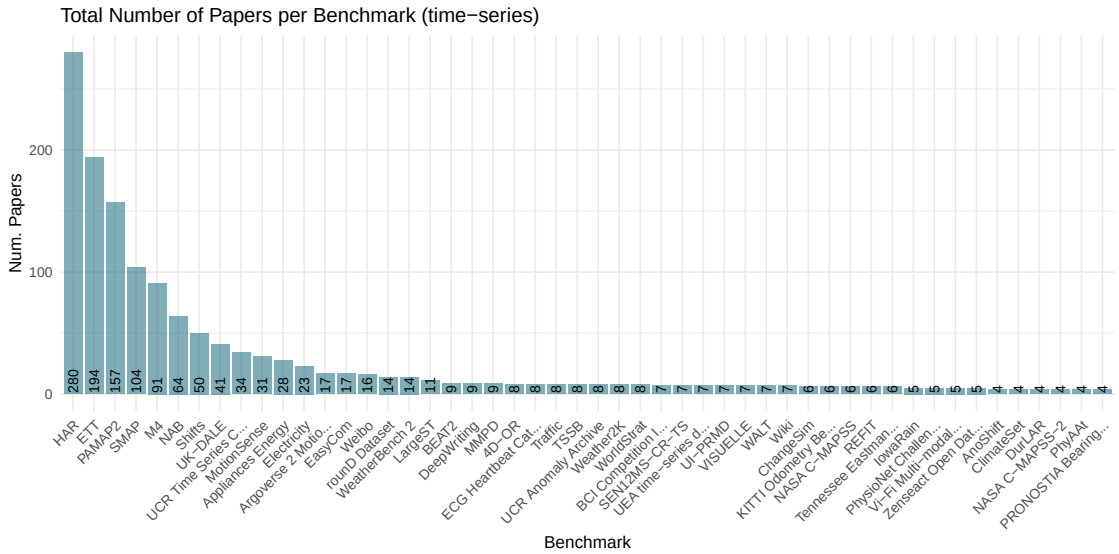

Figure 12: Histogram showing the top-50 benchmarks in the Time Series modality per number of papers addressing them.

Tabular

Tabular benchmarks involve tasks that use structured tabular data, such as classification, regression, and clustering. These benchmarks test AI's ability to analyse datasets commonly used in business, economics, and healthcare.

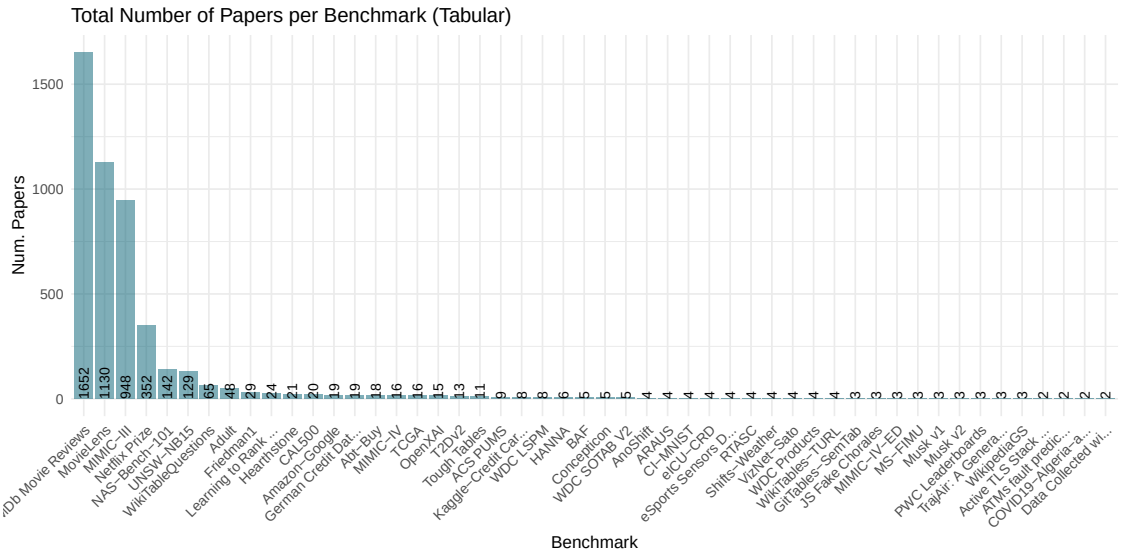

Figure 13: Histogram showing the top-50 benchmarks in the Tabular modality per number of papers addressing them.

Speech

Speech benchmarks are focused on tasks like automatic speech recognition (ASR), speaker identification, and speech synthesis. They evaluate how well AI can process and understand spoken language.

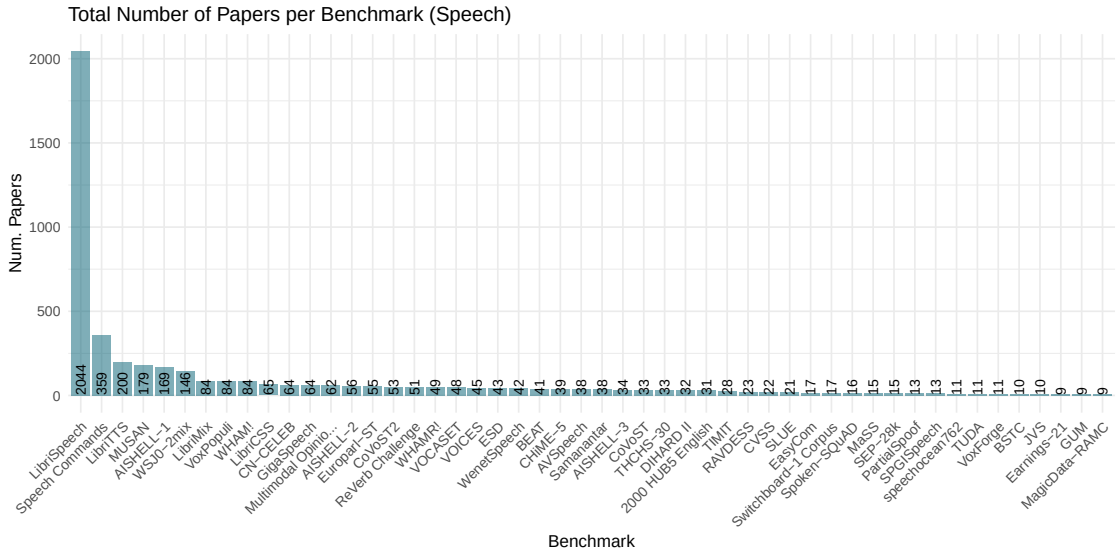

Figure 14: Histogram showing the top-50 benchmarks in the Speech modality per number of papers addressing them.

RGB-D

RGB-D benchmarks assess AI's performance in tasks using both colour (RGB) and depth (D) data, such as scene understanding and object detection. These benchmarks are used in applications requiring depth information, including robotics and AR/VR systems.

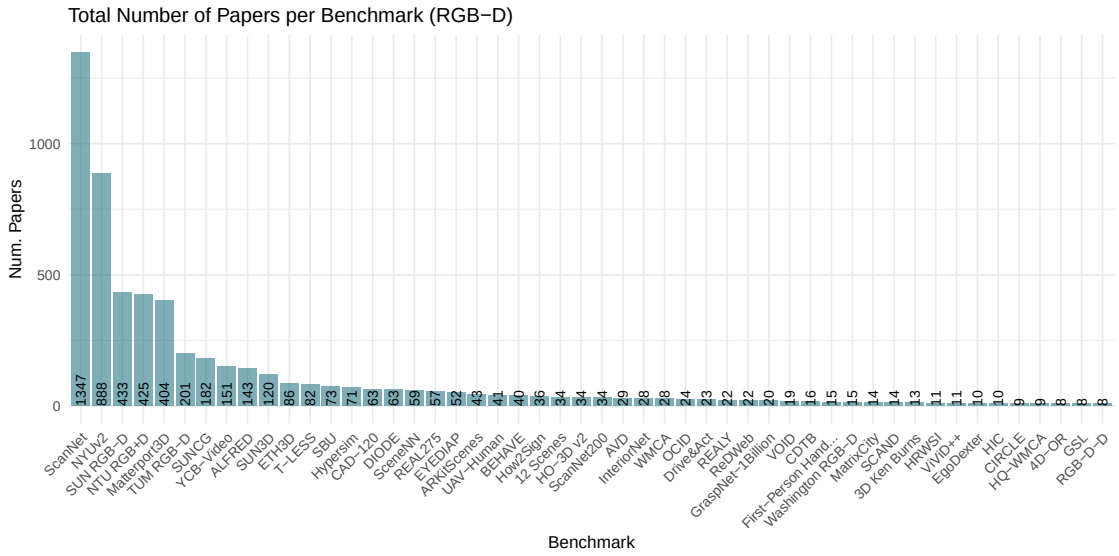

Figure 15: Histogram showing the top-50 benchmarks in the RGB-D modality per number of papers addressing them.

Environment

Environment benchmarks involve tasks related to understanding and interacting with physical or simulated environments, such as navigation, simulation, and reinforcement learning. They test AI's ability to operate effectively in varied environments.

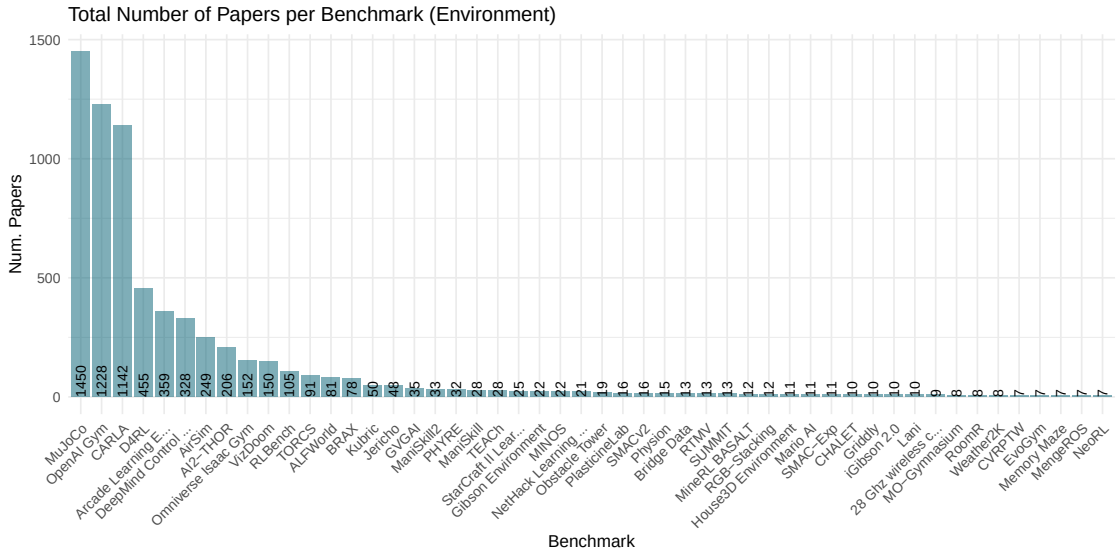

Figure 16: Histogram showing the top-50 benchmarks in the Environment modality per number of papers addressing them.

Point Cloud

Point cloud benchmarks focus on AI's ability to process and analyse data represented as sets of points in 3D space, used in tasks like 3D object detection and segmentation. They are crucial in areas like autonomous driving and robotics.

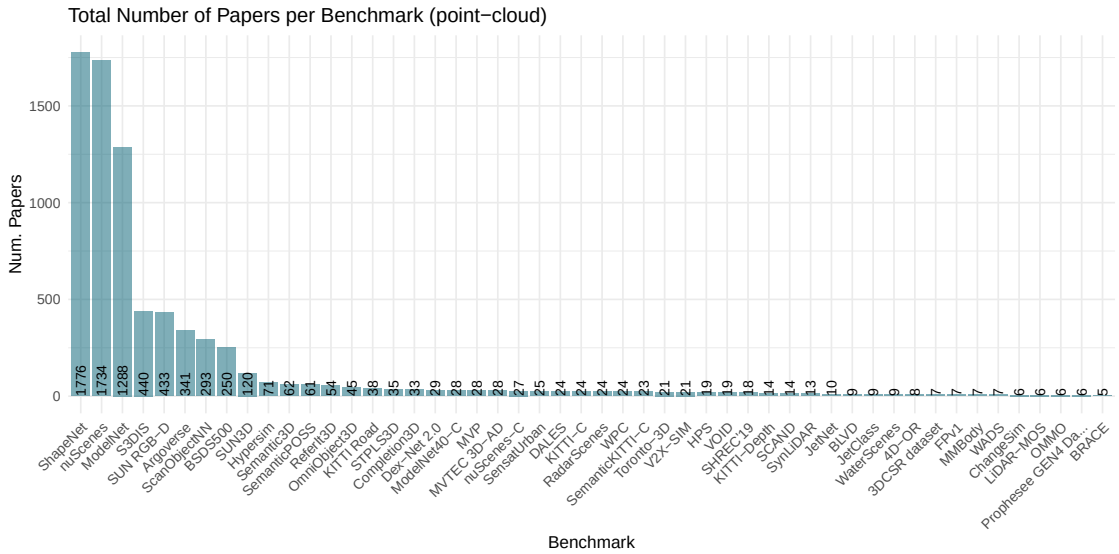

Figure 17: Histogram showing the top-50 benchmarks in the Point Cloud modality per number of papers addressing them.

Biomedical

Biomedical benchmarks are used to assess AI systems on tasks like molecular property prediction, protein structure prediction, and genomics. These benchmarks are pivotal for advancements in drug discovery and personalised medicine.

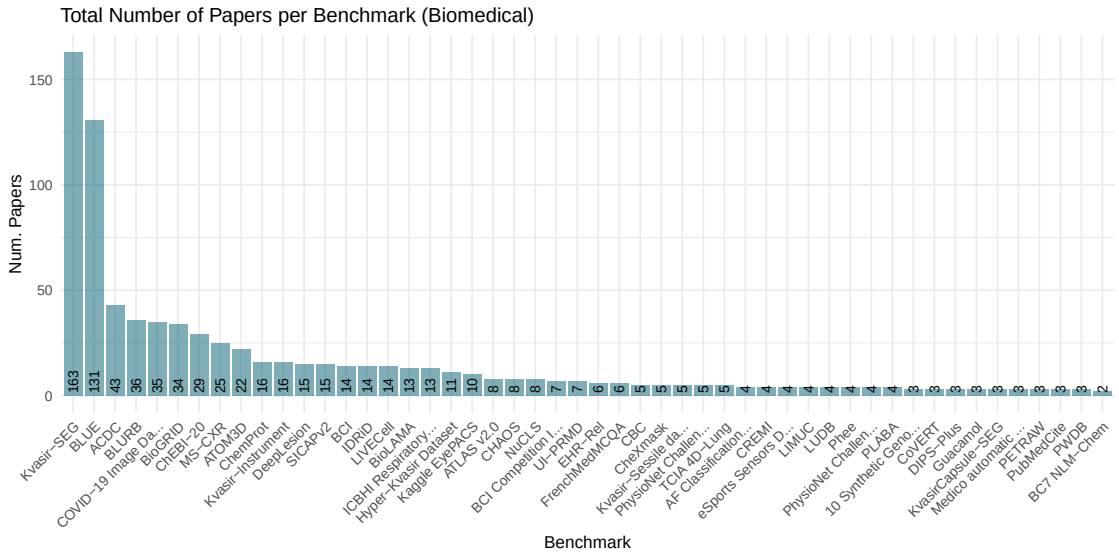

Figure 18: Histogram showing the top-50 benchmarks in the Biomedical modality per number of papers addressing them.

LiDAR

LiDAR benchmarks evaluate AI's capability to process Light Detection and Ranging data, often used in 3D mapping and autonomous vehicles. These benchmarks test the accuracy of 3D reconstructions and object detections from LiDAR data.

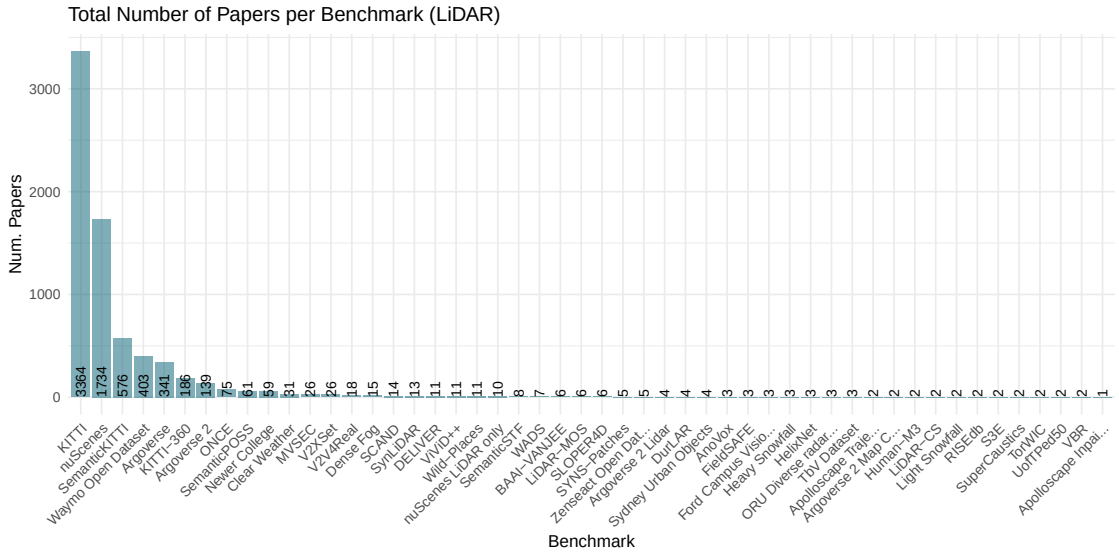

Figure 19: Histogram showing the top-50 benchmarks in the LiDAR modality per number of papers addressing them.

RGB Video

RGB video benchmarks assess AI's ability to interpret and analyse standard colour videos, focusing on tasks like action recognition and video summarisation. These benchmarks are essential for applications in security, entertainment, and surveillance.

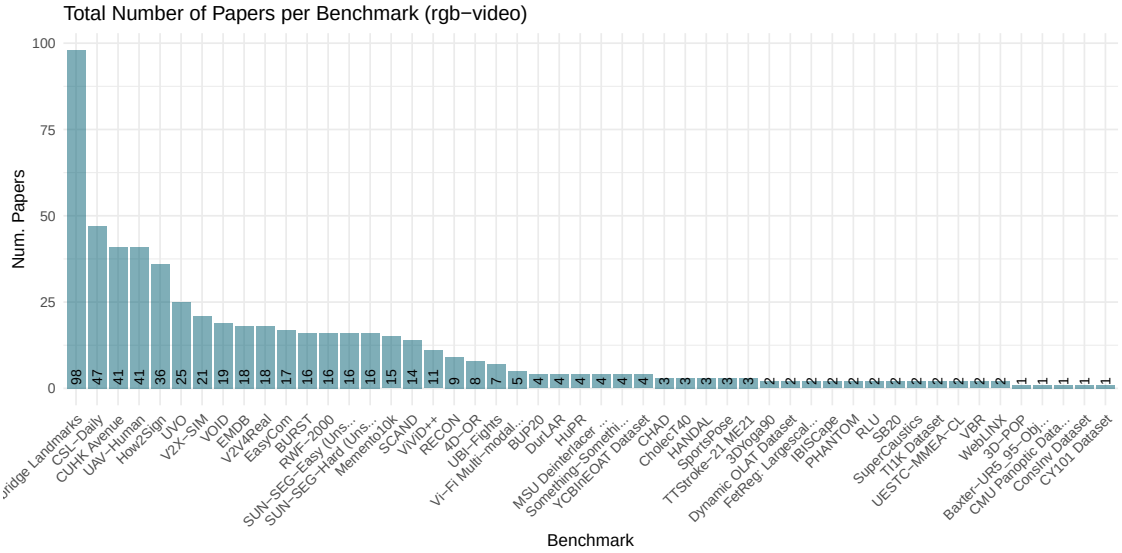

Figure 20: Histogram showing the top-50 benchmarks in the RGB Video modality per number of papers addressing them.

Tracking

Tracking benchmarks test AI's ability to follow objects or points over time in video sequences. Tasks include multi-object tracking and visual object tracking, relevant in robotics and video analysis.

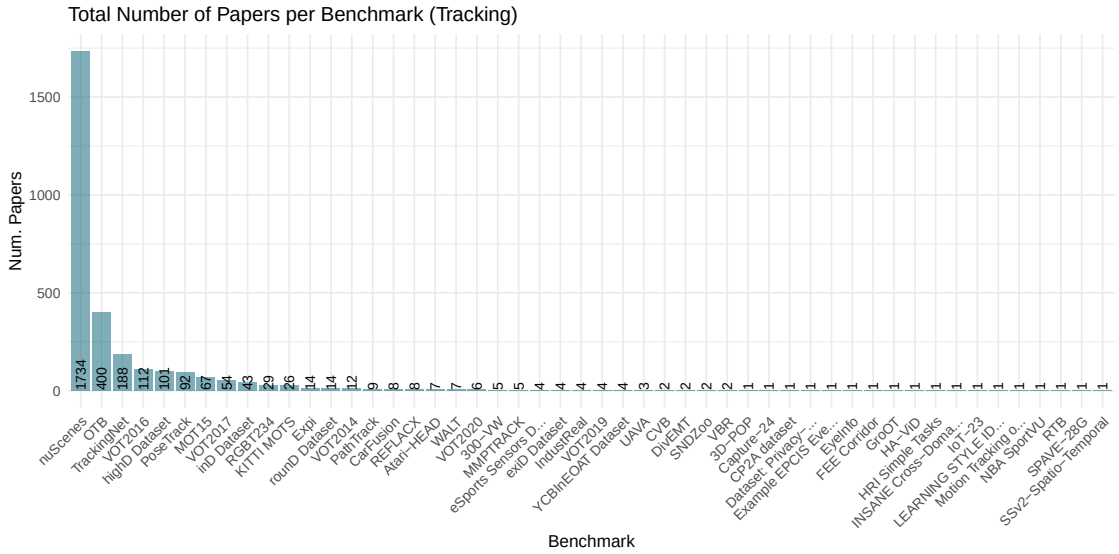

Figure 21: Histogram showing the top-50 benchmarks in the Tracking modality per number of papers addressing them.

6D

6D benchmarks involve tasks that require understanding six degrees of freedom (three spatial and three rotational), such as pose estimation and robotic manipulation. They test AI's capacity for precise spatial awareness and control.

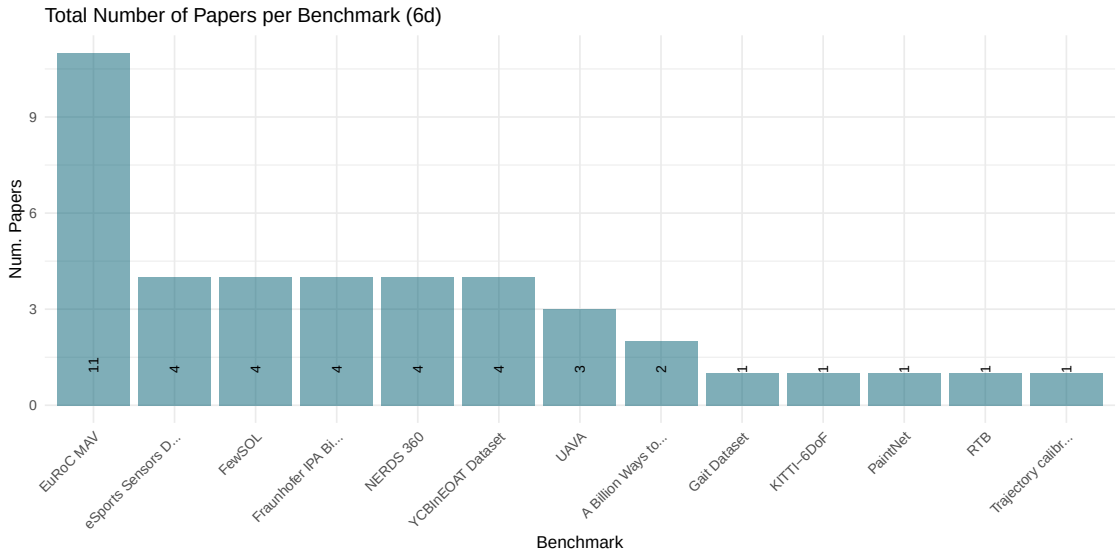

Figure 22: Histogram showing the top-50 benchmarks in the 6D modality per number of papers addressing them.

Biology

Biology benchmarks assess AI's performance on biological data tasks, including gene expression analysis and protein function prediction. These benchmarks are crucial for research in bioinformatics and computational biology.

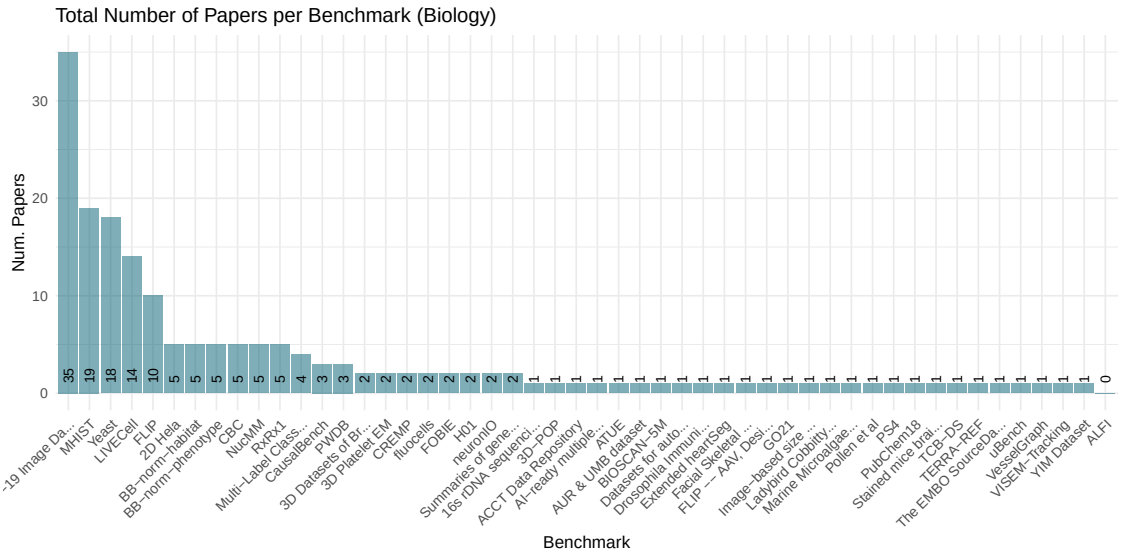

Figure 23: Histogram showing the top-50 benchmarks in the Biology modality per number of papers addressing them.

Actions

Action benchmarks evaluate AI's capacity to recognise and analyze human actions in images or videos. These tasks are key in areas like human-computer interaction, sports analytics, and surveillance.

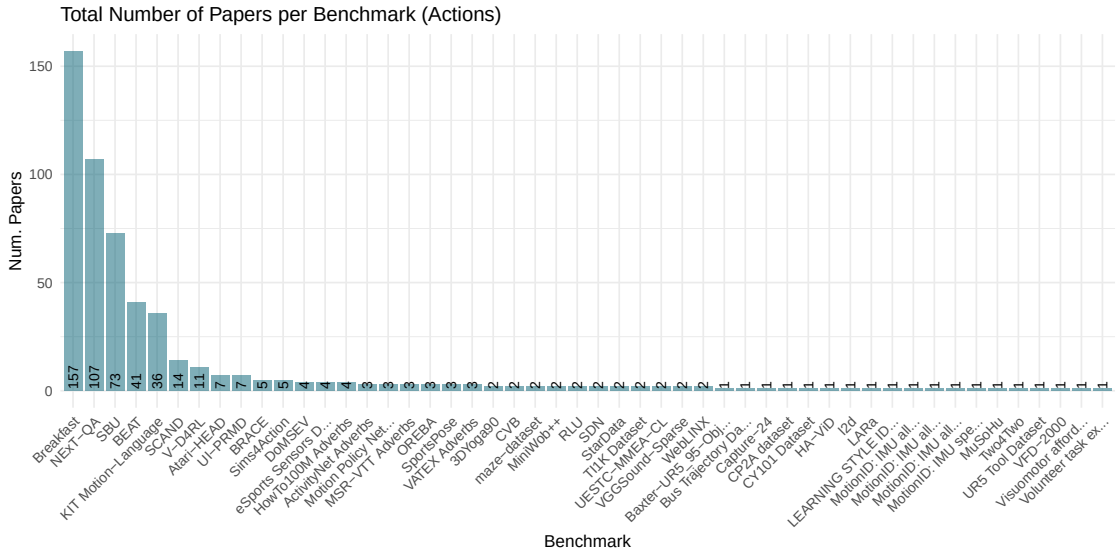

Figure 24: Histogram showing the top-50 benchmarks in the Actions modality per number of papers addressing them.

Stereo

Stereo benchmarks involve tasks using stereo vision data, such as depth estimation and 3D reconstruction from stereo images. These benchmarks are important for applications in robotics, AR/VR, and 3D modelling.

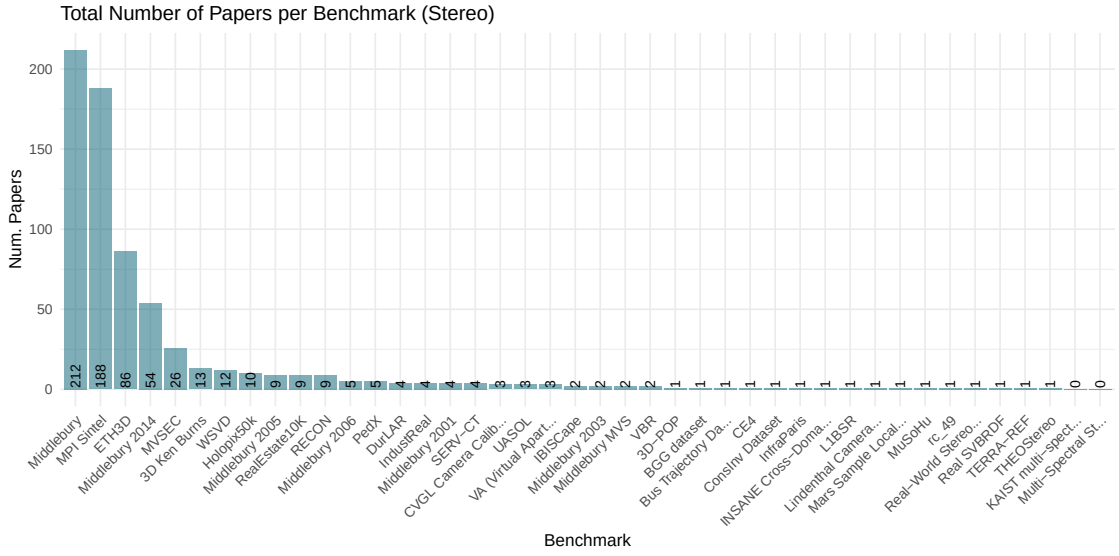

Figure 25: Histogram showing the top-50 benchmarks in the Stereo modality per number of papers addressing them.

EEG

EEG benchmarks focus on tasks involving electroencephalogram data, including brain-computer interfaces and neurological disorder detection. They test AI's ability to process and interpret brainwave data.

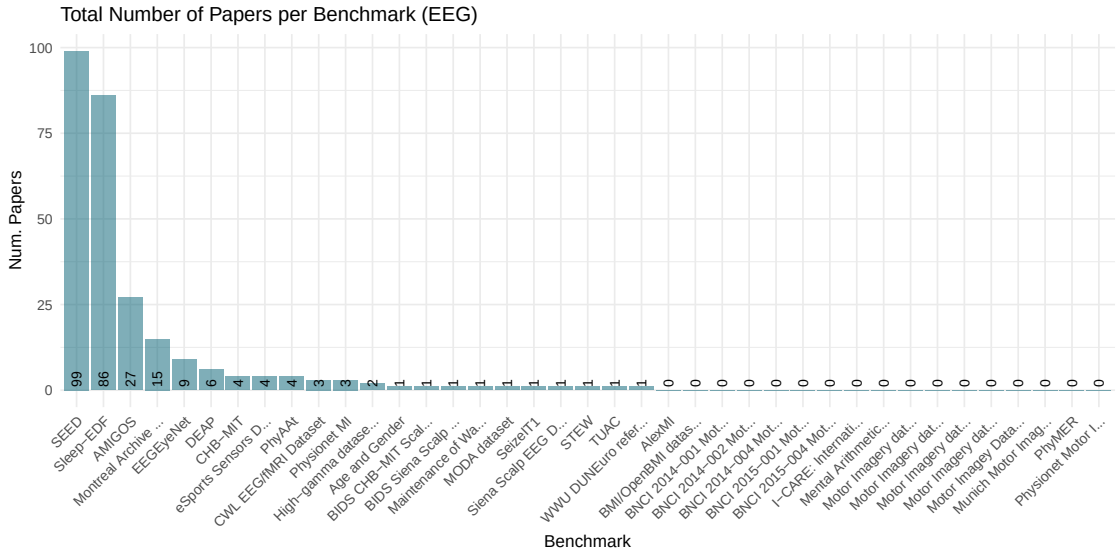

Figure 26: Histogram showing the top-50 benchmarks in the EEG modality per number of papers addressing them.

Tables

Tables benchmarks evaluate the performance of AI systems in processing and understanding tabular data representations, relevant for information extraction, data summarisation, and relational database tasks.

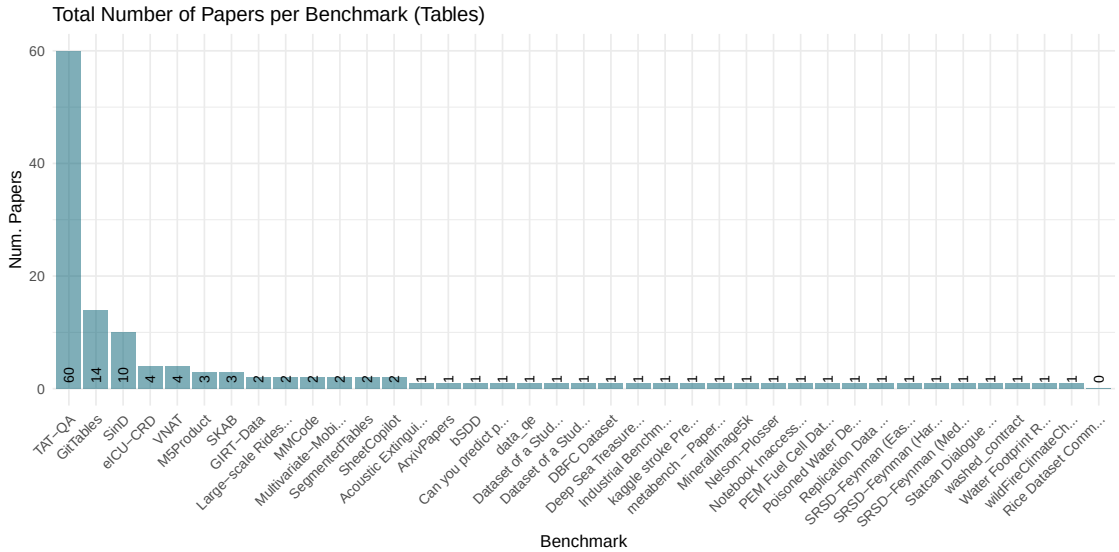

Figure 27: Histogram showing the top-50 benchmarks in the Tables modality per number of papers addressing them.

Hyperspectral Images

Hyperspectral image benchmarks assess AI's capability to process images capturing data across multiple spectral bands. They are used in applications like agriculture, mineralogy, and environmental monitoring.

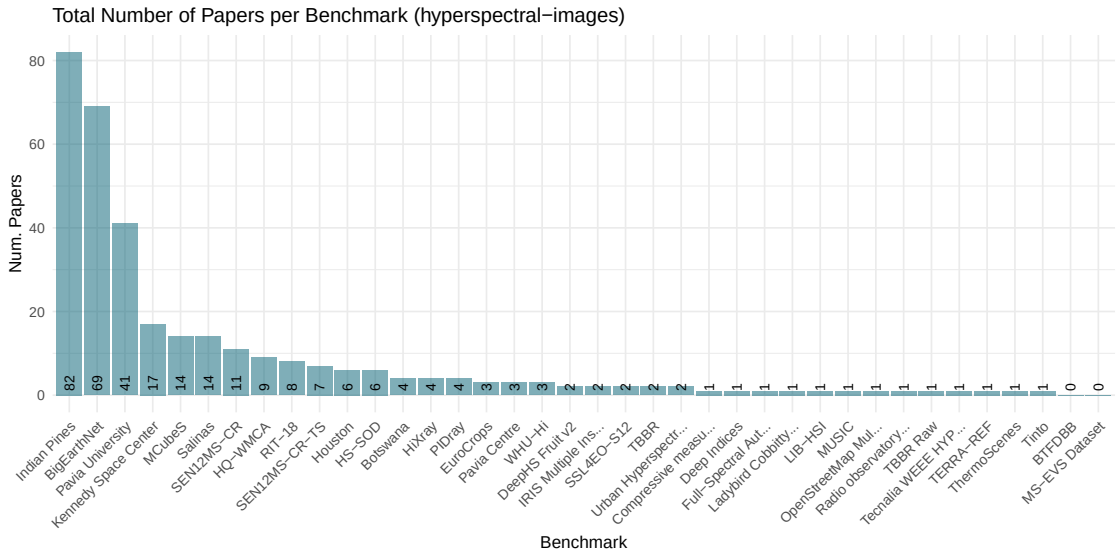

Figure 28: Histogram showing the top-50 benchmarks in the Hyperspectral Images modality per number of papers addressing them.

Music

Music benchmarks test AI's ability in music-related tasks, such as genre classification, music generation, and audio feature extraction. These benchmarks are crucial for innovations in music technology and entertainment.

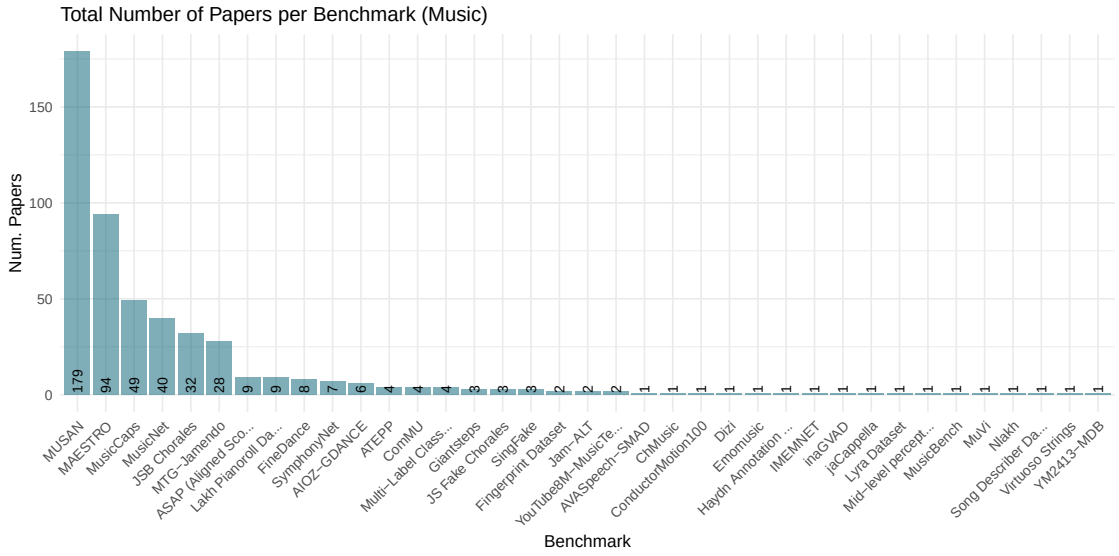

Figure 29: Histogram showing the top-50 benchmarks in the Music modality per number of papers addressing them.

MRI

MRI benchmarks assess AI's performance in analyzing Magnetic Resonance Imaging data used in medical diagnostics, focusing on tasks like image segmentation, disease detection, and treatment planning.

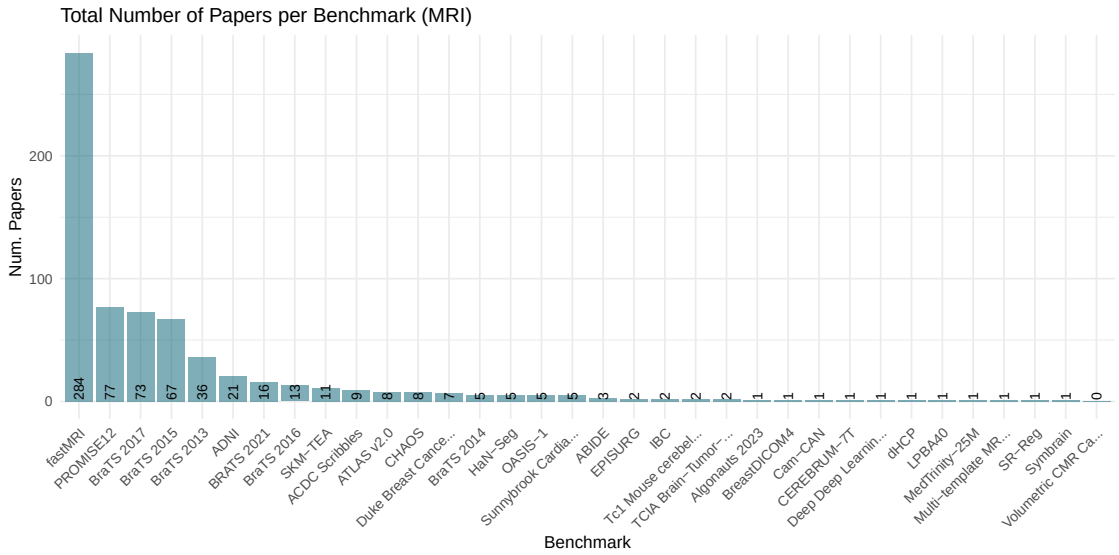

Figure 30: Histogram showing the top-50 benchmarks in the MRI modality per number of papers addressing them.

Physics

Physics benchmarks involve tasks requiring AI to understand and simulate physical phenomena, such as particle simulations and quantum mechanics. They are vital for research in computational physics and engineering.

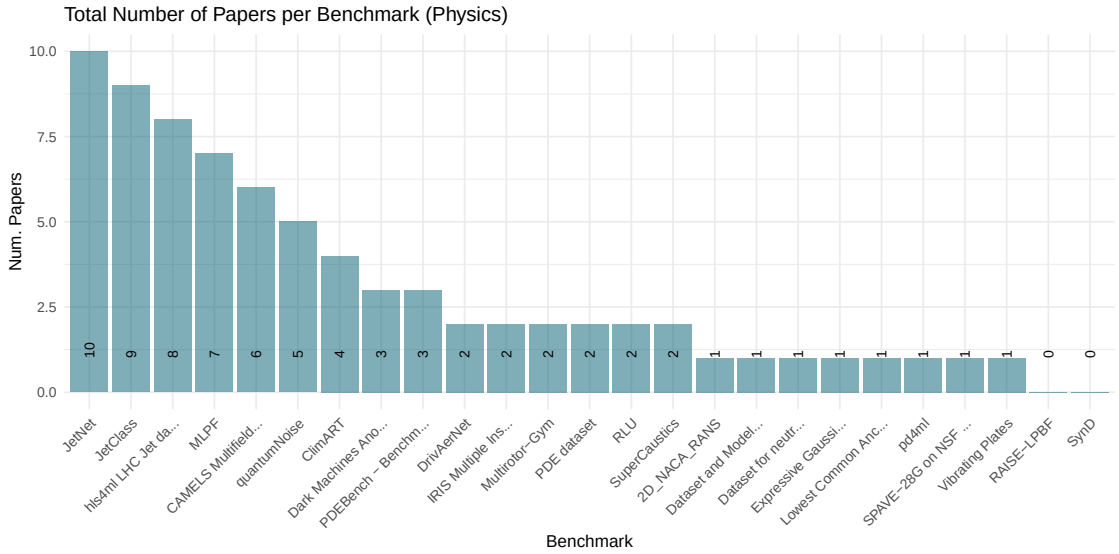

Figure 31: Histogram showing the top-50 benchmarks in the Physics modality per number of papers addressing them.

Dialog

Dialog benchmarks evaluate AI's capability to engage in conversations and understand dialogue contexts. Tasks include question answering, chatbots, and conversational agents, relevant in customer service and virtual assistants.

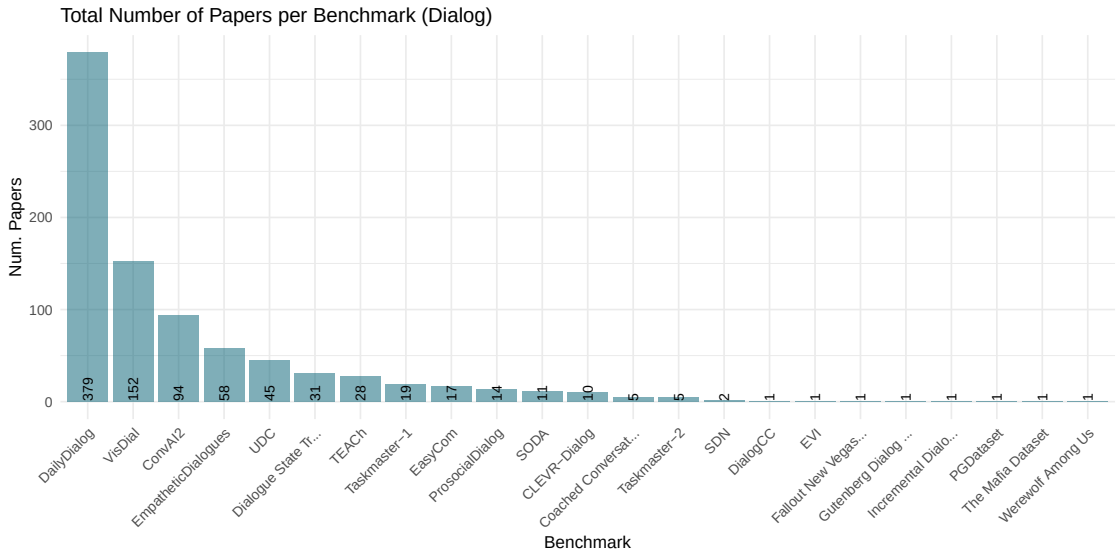

Figure 32: Histogram showing the top-50 benchmarks in the Dialog modality per number of papers addressing them.

Midi

Midi benchmarks assess AI's ability to process and generate music in MIDI format, focusing on tasks like music composition and performance modelling. They are critical for advancements in digital music creation.

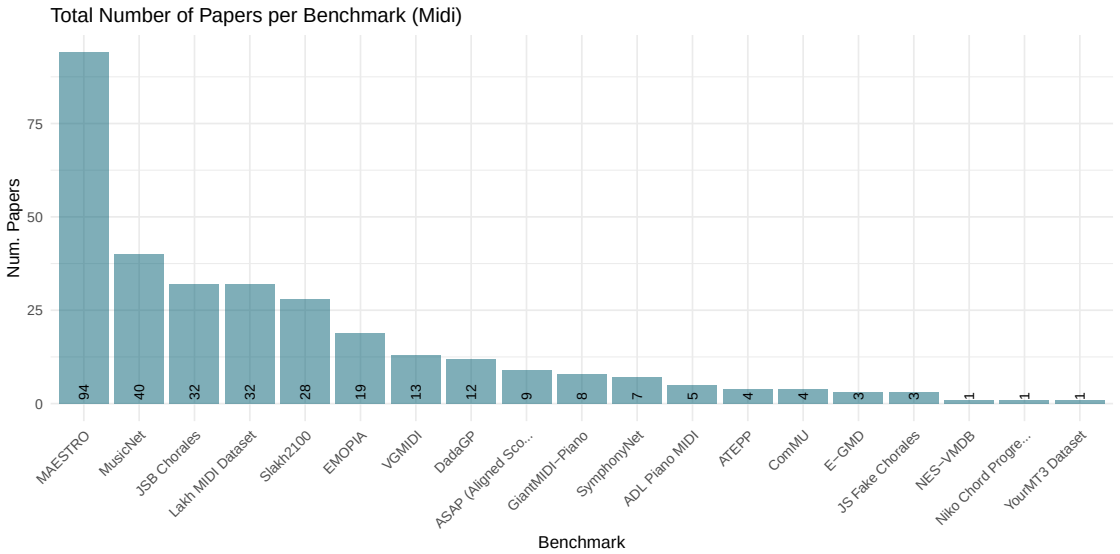

Figure 33: Histogram showing the top-50 benchmarks in the Midi modality per number of papers addressing them.

Ranking

Ranking benchmarks involve tasks that require ordering or prioritising items, such as search engine result ranking and recommendation systems. They test AI's ability to deliver relevant and personalised results.

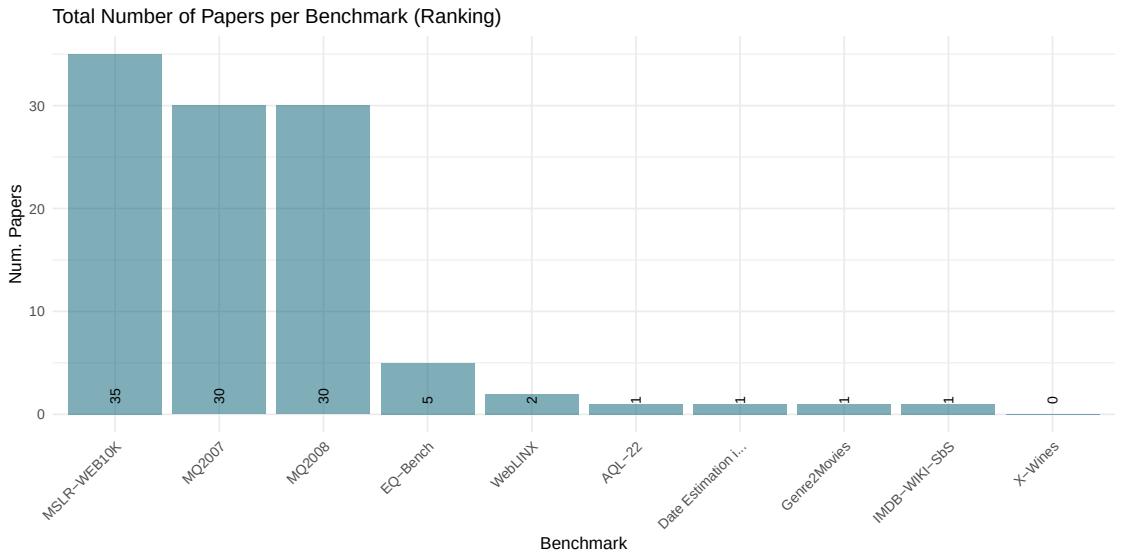

Figure 34: Histogram showing the top-50 benchmarks in the Ranking modality per number of papers addressing them.

Replay Data

Replay data benchmarks assess AI's performance using stored historical data for tasks like reinforcement learning and anomaly detection. These benchmarks are important for learning from past experiences and improving future decisions.

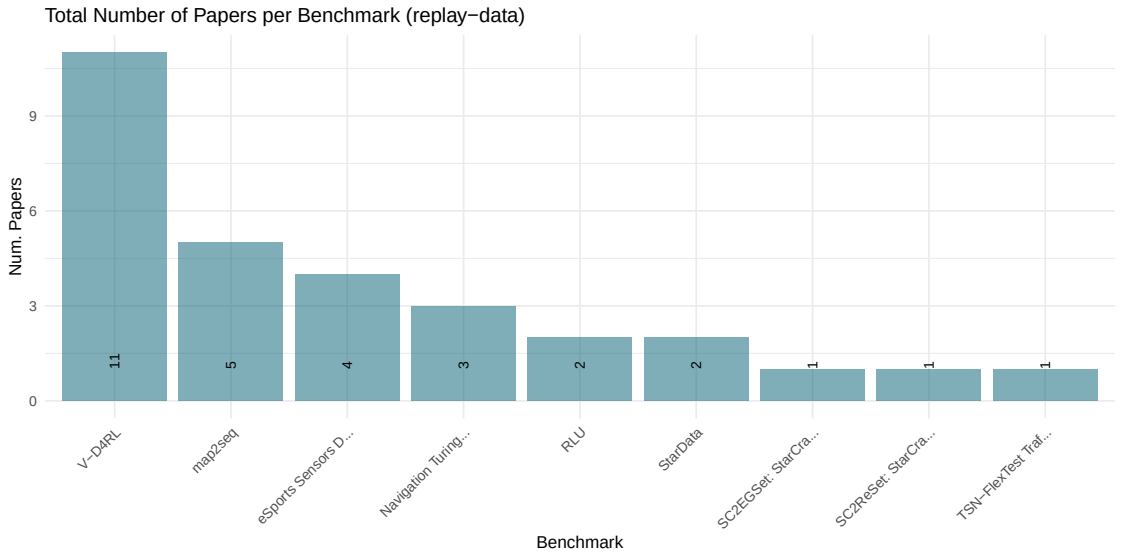

Figure 35: Histogram showing the top-50 benchmarks in the Replay Data modality per number of papers addressing them.

fMRI

fMRI benchmarks assess AI's capability to analyze functional Magnetic Resonance Imaging data, relevant for understanding brain activity and neurological research.

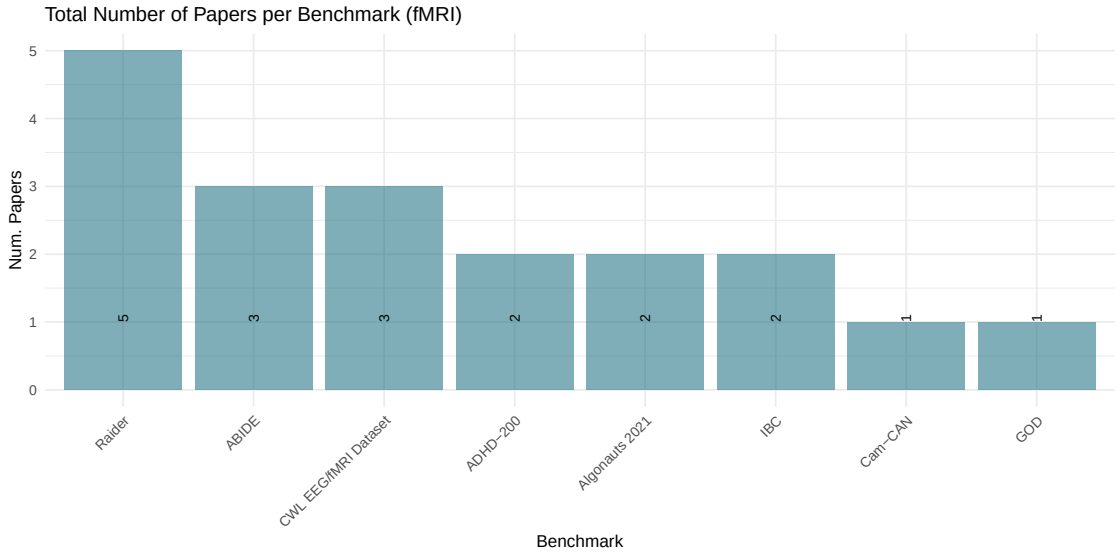

Figure 36: Histogram showing the top-50 benchmarks in the fMRI modality per number of papers addressing them.

Financial

Financial benchmarks evaluate AI's effectiveness in tasks related to financial data, including stock price prediction, fraud detection, and risk assessment. They are crucial for applications in finance and investment.

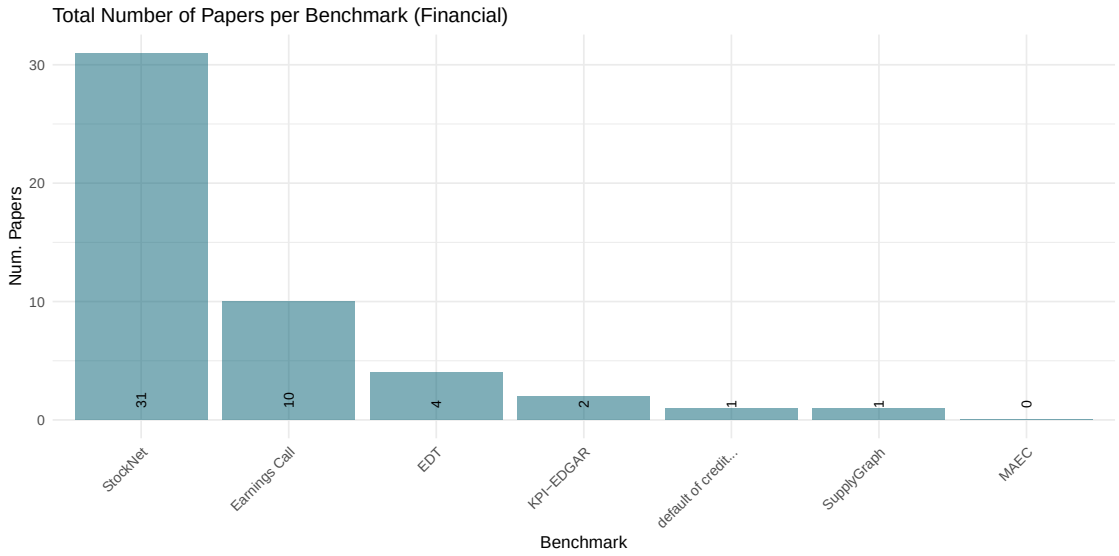

Figure 37: Histogram showing the top-50 benchmarks in the Financial modality per number of papers addressing them.

Cad

Cad benchmarks involve tasks related to computer-aided design, such as 3D model generation and design optimisation. They are essential for advancements in engineering, architecture, and manufacturing.

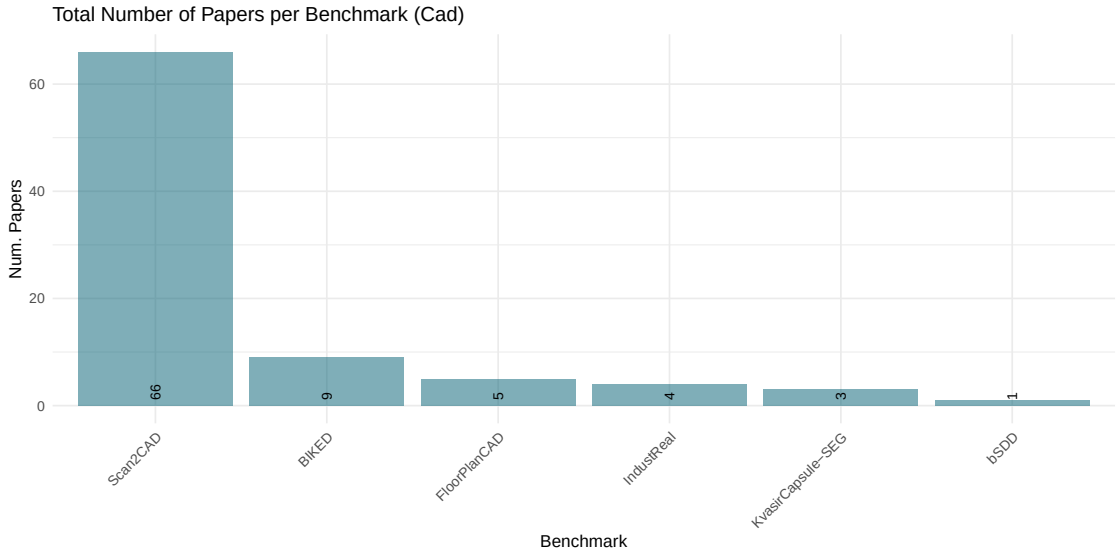

Figure 38: Histogram showing the top-50 benchmarks in the Cad modality per number of papers addressing them.

Parallel

Parallel benchmarks assess AI's efficiency and capability in handling parallel processing tasks and algorithms. These benchmarks are relevant for high-performance computing and scalable AI solutions.

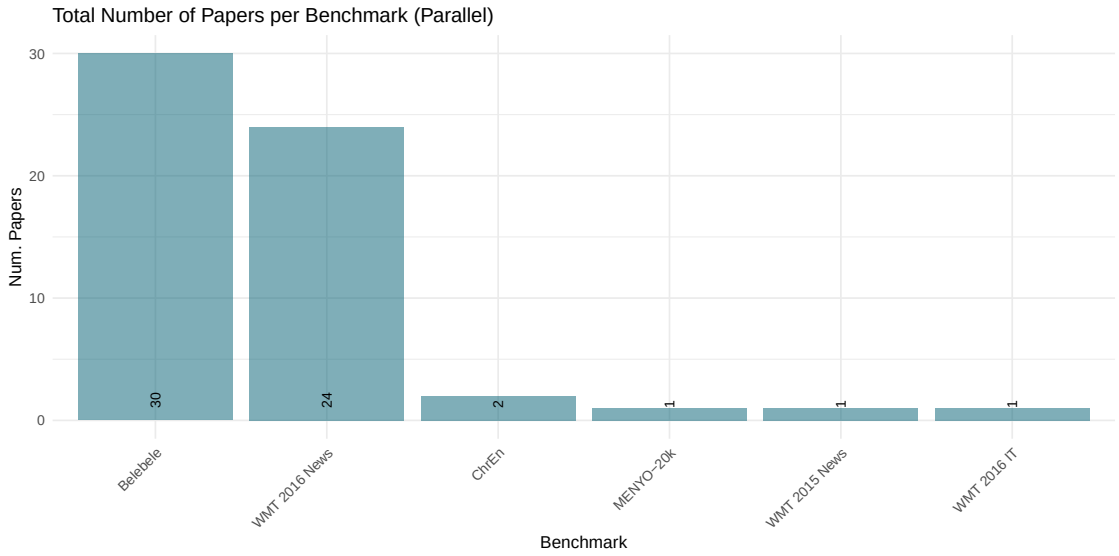

Figure 39: Histogram showing the top-50 benchmarks in the Parallel modality per number of papers addressing them.

Lyrics

Lyrics benchmarks test AI's ability to process and generate song lyrics, focusing on tasks like lyric generation and sentiment analysis. These benchmarks contribute to innovations in music and entertainment.

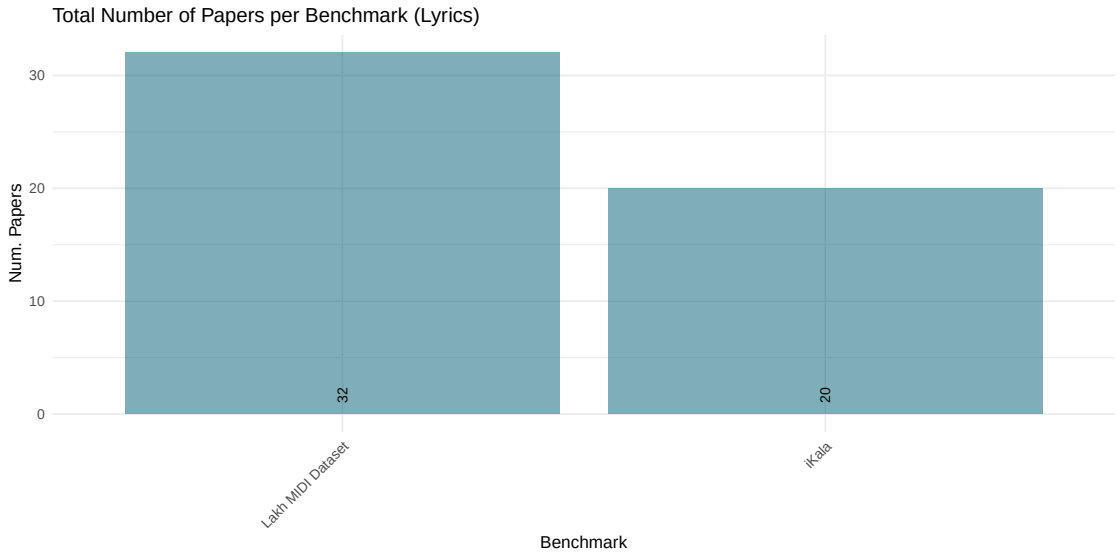

Figure 40: Histogram showing the top-50 benchmarks in the Lyrics modality per number of papers addressing them.

PSG

PSG benchmarks involve Polysomnography data, used in tasks like sleep stage classification and sleep disorder detection. They test AI's ability to interpret comprehensive sleep study data.

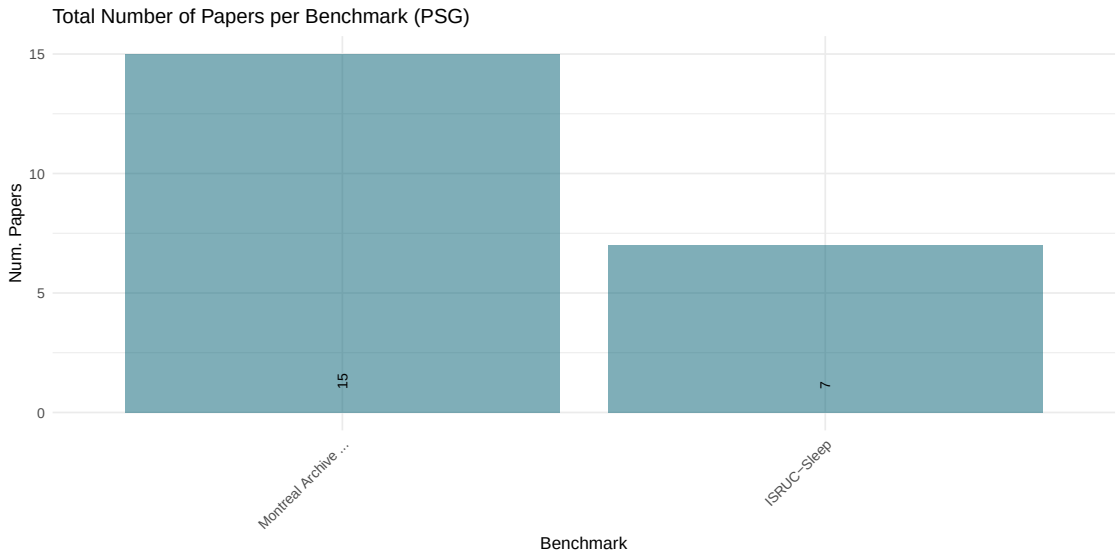

Figure 41: Histogram showing the top-50 benchmarks in the PSG modality per number of papers addressing them.

Conclusions

In this deliverable, we have presented a comprehensive catalogue of AI benchmarks, categorised by data modality and detailing the volume of research papers addressing each benchmark. This work lays the foundation for further analysis and a deeper understanding of the evolving capabilities of AI across domains. By extracting and organising benchmarks based on data modalities such as text, image, audio, and more we have simplified the complex landscape of AI evaluations, allowing for more focused and meaningful analysis. This effort also highlights the importance of structured benchmarking in tracking AI progress and guiding future innovation.

Looking ahead, our project will include detailed mapping and annotation of benchmarks to the cognitive skills they are designed to test. We plan to use both automated tools, such as GPT-4, and manual review to ensure the accuracy and completeness of this mapping. By systematically categorising benchmarks and identifying gaps in current AI capabilities, we aim to facilitate targeted research and development efforts. The analysis of advances in AI capabilities compared to benchmarks up to 2019 will also play a crucial role in understanding the trajectory of AI progress.

The methodology for measuring AI intensity will be also refined to reflect recent advances in LLMs, collecting evidence regarding the activity level around each especific benchmark in term of production (e.g., research publications). This will involve again the use of web-scraping techniques to collect publications, preprints and conference proceedings, as well as develop accurate metrics for assessing current research trends and investments in AI.

References

[1] Analyticsinsight.
Deep Dive into Data Modalities: Understanding Generative AI Inputs. Medium. Accessed: 2024-07-16. July 2024.
URL: https://medium.com/@analyticsinsightsubmissions/deep-dive-into-data-modalities-understanding-generative-ai-inputs-022a845c5e31.

[2] AtlasML. *Papers With Code*. https://www.paperswithcode.com/. 2019.

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

[4] Léon Bottou et al. "Comparison of Classifier Methods: A Case Study in Handwritten Digit Recognition". In: *International Conference on Pattern Recognition*. IEEE Computer Society Press. 1994, pp. 77–77.

[5] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[6] KR1442 Chowdhary and KR Chowdhary. "Natural language processing". In: *Fundamentals of artificial intelligence* (2020), pp. 603–649.

[7] 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].

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

[9] 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.

[10] Stefan Feuerriegel et al. "Generative ai". In: *Business & Information Systems Engineering* 66.1 (2024), pp. 111–126.

[11] V. Gewin. "Data sharing: An open mind on open data". In: *Nature* 529.7584 (2016), pp. 117–119.

[12] José Hernández-Orallo et al. "General intelligence disentangled via a generality metric for natural and artificial intelligence". In: *Scientific reports* 11.1 (2021), pp. 1–16.

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

[14] Percy Liang et al. "Holistic evaluation of language models". In: *arXiv preprint arXiv:2211.09110* (2022).

[15] J. S. Lowndes et al. "Our path to better science in less time using open data science tools". In: *Nature ecology & evolution* 1.6 (2017), p. 0160.

[16] Fernando Martínez-Plumed, Emilia Gómez, and José Hernández-Orallo. "Futures of artificial intelligence through technology readiness levels". In: *Telematics and Informatics* 58 (2021), p. 101525.

[17] Fernando Martínez-Plumed, Emilia Gómez, and José Hernández-Orallo. "Tracking the Evolution of AI: The AIcollaboratory". In: *Proceedings of the 1st International Workshop: Evaluating Progress in Artificial Intelligence (EPAI 2020)*. 2020.

[18] Fernando Martínez-Plumed and José Hernández-Orallo. "Dual Indicators to Analyze AI Benchmarks: Difficulty, Discrimination, Ability, and Generality". In: *IEEE Transactions on Games* 12.2 (2020), pp. 121–131. DOI: 10.1109/TG.2018.2883773.

[19] 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.

[20] 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.

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

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

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

[24] Aarohi Srivastava et al. "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models". In: *arXiv preprint arXiv:2206.04615* (2022).

[25] 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.

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

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