FM Leaderboard Collection Overview

Our leaderboard collection was conducted in four distinct rounds:

First Round (Early November – Late December 2023)

We began by investigating approximately 8,248 ML leaderboards listed in popular "awesome" lists and Papers With Code (PWC). During this phase, challenges such as ambiguous documentation, mismatched entities, and missing information became evident. While reporting these issues to leaderboard operators, our primary focus was on understanding leaderboard operations rather than identifying specific "smells." However, the lack of transparency in operations prompted us to delve deeper into the underlying issues associated with LBOps.

Second Round (Mid-February – Early March 2024)

With the emergence of impactful FM leaderboards, such as CompassRank [2] and SuperCLUE [4], we initiated a second round of data collection. Expanding our sources to include literature from Google Scholar, we increased the total number of leaderboards to 9,574. The inclusion of new sources introduced a broader variety of leaderboards, resulting in an extended process for resolving issues. Feedback from leaderboard operators during this phase highlighted additional operational challenges, expanding our understanding of leaderboard-specific concerns.

Third Round (Late April – Early June 2024)

In late April, PWC officials made a significant adjustment to their leaderboard data by removing all leaderboards collected through the Hugging Face (HF) model card method from their website and archives—this represented approximately 3,300 leaderboards based on our initial scraping. To stay consistent with the updated PWC data, we also excluded this subset from our collection. Meanwhile, the emergence of new FM leaderboards, such as the Domain LLM Leaderboard [3], led us to propose the P_3 workflow pattern. We recollected ML leaderboards using the same multivocal literature review (MLR) approach, resulting in a finalized collection of 6,849 ML leaderboards by early June. During this phase, we actively communicated with leaderboard operators via issue trackers to gather additional insights into operational "smells," which informed the preparation of our draft.

Fourth Round (Late September – Mid-October 2024)

In this final round, we performed additional leaderboard retrievals using two methods. Firstly, we collect FM leaderboards via blogs (via Google), tweets (via X), and videos (via YouTube) using the keywords "foundation model leaderboard," "large language model leaderboard," "large multimodal model leaderboard," and "large action model leaderboard." Considering all sources, we identified 142 FM leaderboards in total, and 20 of them are not present in the third round collection. Besides, we retrieve FM leaderboards via GitHub, PWC, and HF Spaces using the keyword "leaderboard". This round yielded: 1,681 leaderboard mentions from GitHub, 426 leaderboards from HF Spaces, and 7,538 leaderboards from PWC. After filtering, we identified up to 1,045 FM leaderboards, visualized in the outermost transparent circle of Figure 1, which highlights differences in leaderboard counts across different collection methods. As we see, our latest collection method in the current draft is the most comprehensive to date, as it captures the largest number of FM leaderboards while ensuring that none of the FM leaderboards identified through previous methods are overlooked.

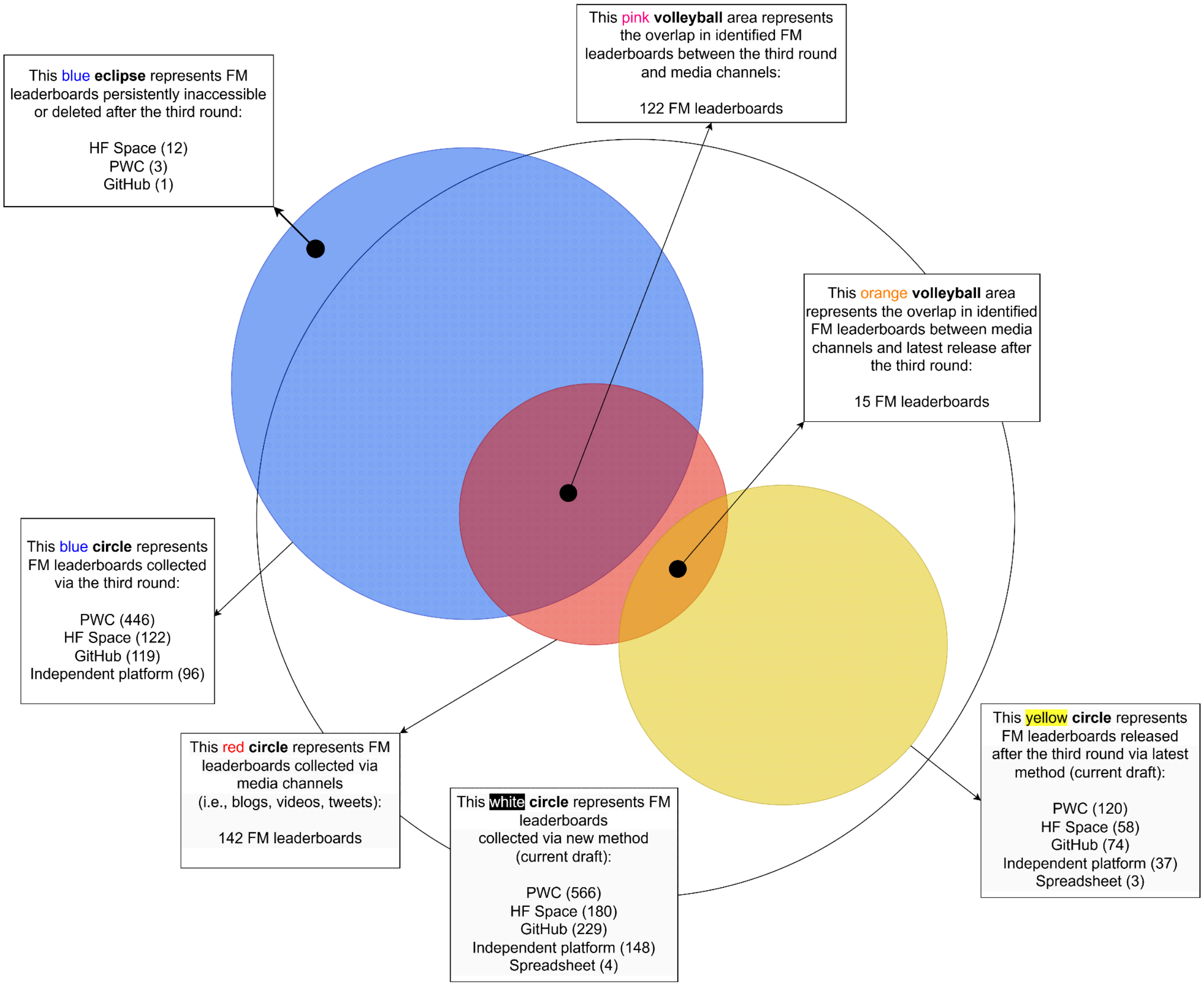
This pink **volleyball** area represents the overlap in identified FM leaderboards between the third round and media channels:
122 FM leaderboards
This blue **eclipse** represents FM leaderboards persistently inaccessible or deleted after the third round:
HF Space (12)
PWC (3)
GitHub (1)
This orange **volleyball** area represents the overlap in identified FM leaderboards between media channels and latest release after the third round:
15 FM leaderboards
This blue **circle** represents FM leaderboards collected via the third round:
PWC (446)
HF Space (122)
GitHub (119)
Independent platform (96)
This red **circle** represents FM leaderboards collected via media channels (i.e., blogs, videos, tweets):
142 FM leaderboards
This white **circle** represents FM leaderboards collected via new method (current draft):
PWC (566)
HF Space (180)
GitHub (229)
Independent platform (148)
Spreadsheet (4)
This yellow **circle** represents FM leaderboards released after the third round via latest method (current draft):
PWC (120)
HF Space (58)
GitHub (74)
Independent platform (37)
Spreadsheet (3)

Figure 1: Venn diagram about the number of FM leaderboards collected in the initial draft during the third round, media channels (blogs, tweets, videos), and latest methods (latest draft) during the fourth round.

The leaderboard collections from the first and second rounds are outdated and no longer serve as strong references for our current study. Since *Section IV.B* already provides a detailed account of our current collection methodology, we would now focus on elaborating the FM leaderboard collection processes conducted during the third and fourth rounds (media channels).

Third Round FM Leaderboard Collection

In the third round, we conducted a multivocal literature review using two key sources: **Google Scholar** and **GitHub Awesome Lists**.

1. **Rationale for Source Selection**
    - **Google Scholar**: Chosen for its broad coverage of academic literature, providing in-depth exploration of the relatively novel concept of FM evaluation.
    - **GitHub Awesome Lists**: Selected for their extensive collection of community-driven, domain-specific resources, cataloging widely recognized benchmarks and leaderboards across various fields.
2. **Search Query Preparation**
    - **Keyword Compilation**:
      We began by collaboratively brainstorming a list of FM-related keywords, guided by the frequent occurrence of these terms in prior literature and their alignment with our research objectives. These keywords, listed in Table 1 along with the number of retrieved grey literature, were validated through preliminary searches on Google and GitHub.
        - **Foundation Models (FM)**: Keywords such as *"foundation model," "large language model,"* and *"large multimodal model"* were used to encompass various types of FMs known to the authors.
        - **Evaluation Focus**: To reflect the study's emphasis on benchmark evaluations, we included terms like *"benchmark"* and *"leaderboard,"* representing tools used to rank and compare model performance.
        - **Search Platform Optimization**:
            - On Google Scholar, we appended core keywords with terms such as *"review," "survey,"* and *"study"* to target literature summarizing the field's state and evaluation practices, following established methods in prior tertiary studies [6].
            - On GitHub, we prefixed keywords with *"awesome"* to locate curated repositories highlighting influential benchmarks and leaderboards.

Table 1: The number of grey literature retrieved by each primary keyword for each source.

Primary Keyword	Number of GitHub Awesome Lists	Number of Google Scholar Literature
benchmark	104	~4,780,000
dataset	650	~7,300,000
leaderboard	24	~52,400
foundation model	62	~8,110,000
large action model	2	~4,740,000

large AI model	37	~7,980,000
large audio model	4	~4,730,000
large audio language model	2	~3,370,000
large behavior model	2	~8,230,000
large language model	197	~5,750,000
large multimodal model	26	~1,710,000
large speech model	1	~3,230,000
large time series model	3	~7,790,000
large trajectory model	1	~5,070,000
large vision model	14	~6,010,000
small language model	7	~5,730,000

3. **Search Execution**
    - **Google Scholar**:
      For each search query, we reviewed the first 20 pages of results to ensure a thorough yet manageable analysis. Each retrieved paper was skimmed for relevance, focusing on sections dedicated to ML evaluation.
    - **GitHub Awesome Lists**:
      Searches were conducted using optimized keywords to identify curated repositories that catalog domain-specific benchmarks and leaderboards.
4. **Validation of Search Terms**
   To evaluate the adequacy and comprehensiveness of the selected keywords, we compiled a "golden set" of widely recognized leaderboards: Open LLM Leaderboard [7], Chatbot Arena [8], SuperCLUE Leaderboard [9], CompassRank Leaderboard [10], HELM Leaderboard [11].
5. These leaderboards served as benchmarks to validate our approach. Our chosen keywords successfully retrieved these leaderboards in searches, demonstrating their relevance to foundational research and practical evaluation contexts.
6. **Final Selection**
   After manual filtering for relevance/ensuring that the results are ML-related and align with our focus on FMs and benchmarking), we identified 3,200 initial results and ultimately selected 30 relevant articles on FM evaluation for our literature review, as presented in Table 2. This comprehensive dataset formed the foundation for our subsequent analysis of FM evaluation practices.

Table 2: Finalized multivocal literature on FM evaluation retrieved by each primary keyword for each source.

Primary Keyword	Finalized Awesome GitHub Lists	Finalized Google Scholar Literature
benchmark	[15], [26], [27], [28], [29]	
dataset	[14], [15], [21], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68]	
leaderboard		
foundation model	[14], [24], [25]	[69], [70], [71], [72]
large action model		[73], [74], [75], [76]
large AI model		[77]
large audio model		
large audio language model		
large behavior model		
large language model	[12], [13], [14], [15], [16], [17], [18], [19], [20], [22], [23]	[71], [73], [74], [75], [76], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92], [93], [94], [95], [96], [97], [98], [99]
large multimodal model	[13], [14]	[74], [79]
large speech model		
large time series model		
large trajectory model		
large vision model	[12]	[74], [78]
small language model		

Fourth Round FM Leaderboard Collection across Media Channels

In the fourth round, we collected FM leaderboards from various media sources including blogs (via Google), tweets (via X), and videos (via YouTube) using the keywords "foundation model leaderboard," "large language model leaderboard," "large multimodal model leaderboard," and "large action model leaderboard." Table 3 presents the identified FM leaderboards along with their provenance references, categorized by their respective channels, based on targeted keyword searches.

Table 3: Metadata of FM leaderboards retrieved using primary keywords for each source.

Primary Keyword	Finalized FM Leaderboards via Google	Finalized FM Leaderboards via Blogs (Google)	Finalized FM Leaderboards via Videos (YouTube)	Finalized FM Leaderboards via Tweets (via X)
"foundation model leaderboard"	Italian Open LLM Leaderboard [100], HELM AIR-Bench [101], HELM Classic [101], HELM CLEVA[101], HELM HEIM [101], HELM Image2Struct [101], HELM Instruct [101], HELM Lite [101], HELM MMLU [101], HELM ThaiExam [101], HELM VHELM [101], VMLU [102], Artificial Analysis LLM API Providers Leaderboard [103], Artificial Analysis Models Leaderboard [103], Artificial Analysis Speech to Text AI Model & Provider Leaderboard [103], Artificial Analysis Text to	Italian Open LLM Leaderboard [100], MMMU [105], Predibase Fine-Tuning Leaderboard [106], AlpacaEval [108], Julia LLM Leaderboard [111], ML.ENERGY Leaderboard [112], Chatbot Arena [115]	LLM Rankings [178], Aider LLM Leaderboards [178], Berkeley Function-Calling Leaderboard [178], LiveBench [178], Chatbot Arena [179]	Open LLM Leaderboard [146], Chatbot Arena [147], CLEM Leaderboard [175]

	Image AI Model & Provider Leaderboard [103], Artificial Analysis Text to Speech AI Model & Provider Leaderboard [103], C-Eval [104], C-Eval Hard [104], MMMU [105], Predibase Fine-Tuning Leaderboard[106], SecEval [107], AlpacaEval [108], FanOutQA [109], Vellum LLM Leaderboard [110], Julia LLM Leaderboard [111], ML.ENERGY Leaderboard [112], EQ-Bench [113], LLM Rankings [114], Chatbot Arena [115], Toloka LLM Leaderboard [116], OpenCompass Large Language Model Leaderboard [117], OpenCompass Multi-modal Leaderboard [117], SEAL LLM Leaderboards[118] , WildVision Arena [119], OpenVLM Leaderboard [119], AI2 A-OKVQA [120], AI2 ARC [120], AI2			

	CosmosQA [120], AI2 CSQA2 [120], AI2 GENIE - aNLG [120], AI2 GENIE - ARC-DA [120], AI2 GENIE - Summarization XSUM [120], AI2 MC-TACO [120], AI2 MOCHA [120], AI2 Natural Instructions [120], AI2 NYCC [120], AI2 OpenbookQA [120], AI2 PIQA [120], AI2 ProtoQA [120], AI2 QASC [120], AI2 Sherlock [120], AI2 SIQA [120], AI2 StrategyQA [120], AI2 WinoGrande[120], AI2 αNLI [120]			
"large language model leaderboard"	Open LLM Leaderboard [121], Vellum LLM Leaderboard [110], LLM Leaderboard [122], Chatbot Arena [123], Trustbit LLM Leaderboards [123], EQ-Bench [123], Berkeley Function-Calling Leaderboard [123], SEAL LLM Leaderboards [123], OpenCompass Large Language Model	Chatbot Arena [123], Trustbit LLM Leaderboards [123], EQ-Bench [123], Berkeley Function-Calling Leaderboard [123], SEAL LLM Leaderboards [123], OpenCompass Large Language Model Leaderboard [123], OpenCompass Multi-modal Leaderboard [123], CanAiCode	LLM Rankings [178], Aider LLM Leaderboards [178], Berkeley Function-Calling Leaderboard [178], LiveBench [178], WildVision Arena [180], OpenVLM Leaderboard [180], Chatbot Arena [179], Open LLM Leaderboard [181], AgentBench [107], Open Medical-LLM Leaderboard [182], Open CoT Leaderboard [183], Open Arabic LLM Leaderboard [185], TTS Arena [186],	HELM Classic [148], HELM CLEVA [148], HELM HEIM [148], HELM Image2Struct [148], HELM Instruct [148], HELM Lite [148], HELM MMLU [148], HELM ThaiExam [148], HELM VHELM [148], Chatbot Arena [149], Open LLM Leaderboard [150], BenBench [151], Open Medical-LLM Leaderboard

	Leaderboard [123], OpenCompass Multi-modal Leaderboard [123], CanAiCode Leaderboard [123], Open Multilingual LLM Evaluation Leaderboard [123], MTEB [123], MTEB Arena [123], AlpacaEval [123], UGI Leaderboard [123], La Leaderboard [124], Open Ko-LLM Leaderboard [125], Toloka LLM Leaderboard [116], Low-bit Quantized Open LLM Leaderboard [126], Aider LLM Leaderboards[127] , LLM Observatory Leaderboard [128], European LLM Leaderboard [129], Hughes Hallucination Evaluation Model leaderboard [130], SuperGLUE [131], Code Lingua [132], BigCodeBench [133], Artificial Analysis LLM API Providers Leaderboard [133], Artificial	Leaderboard [123], Open Multilingual LLM Evaluation Leaderboard [123], MTEB [123], MTEB Arena [123], AlpacaEval [123], UGI Leaderboard [123], La Leaderboard [124], Low-bit Quantized Open LLM Leaderboard [126], European LLM Leaderboard [129], Hughes Hallucination Evaluation Model leaderboard [130], BigCodeBench [133], Artificial Analysis LLM API Providers Leaderboard [133], Artificial Analysis Models Leaderboard, Artificial Analysis Speech to Text AI Model & Provider Leaderboard [133], Artificial Analysis Text to Image AI Model & Provider Leaderboard [133], Artificial Analysis Text to Speech AI Model & Provider Leaderboard [133], Open Medical-LLM Leaderboard	MTEB [187], Self-Improving Leaderboard [188], all HF Space leaderboards [235]	[152], ZebraLogic [153], TrustLLM [154], JailbreakBench [155], Open Ko-LLM Leaderboard [156], Powered-by-Intel LLM Leaderboard [157], Indic LLM Leaderboard [158], CyberSecEval [159], AlpacaEval [160], Hughes Hallucination Evaluation Model leaderboard [161], LLM-Perf Leaderboard [162], C-Eval [163], C-Eval Hard [163], SEAL LLM Leaderboards [164], BenCzechMark [165], AIR-Bench Leaderboard [166], MedS-Bench [167], BigCodeBench [168], LiveCodeBench [169], MMStar [170], FlagEval Large Language Model Evaluation Capability Leaderboard [171], FlagEval Large Model K12 Subject Test Leaderboard

	Analysis Models Leaderboard, Artificial Analysis Speech to Text AI Model & Provider Leaderboard [133], Artificial Analysis Text to Image AI Model & Provider Leaderboard [133], Artificial Analysis Text to Speech AI Model & Provider Leaderboard [133], Open Medical-LLM Leaderboard [133], EvalPlus [133], HELM AIR-Bench [101], HELM Classic [101], HELM CLEVA [101], HELM HEIM [101], HELM Image2Struct [101], HELM Instruct [101], HELM Lite [101], HELM MMLU[101], HELM ThaiExam [101], HELM VHELM [101], LLM-Perf Leaderboard [134], VMLU [102], AI2 A-OKVQA [120], AI2 ARC [120], AI2 CosmosQA [120], AI2 CSQA2 [120], AI2 GENIE - aNLG [120], AI2	[133], EvalPlus [133], LLM-Perf Leaderboard [134], Thai LLM Leaderboard [135], ML.ENERGY Leaderboard [112], Open Arabic LLM Leaderboard [138], Icelandic LLM leaderboard [139], Provider Leaderboard [140], Indic LLM Leaderboard [142], Julia LLM Leaderboard [143]		[171], FlagEval Multimodal Model Evaluation Leaderboard [171], KoLA [172], Thai LLM Leaderboard [173], Accubits Open-source Large Language Models Leaderboard [174], Accubits InstructEval Leaderboard [174], Accubits Business Friendly LLMs Leaderboard [174], Accubits Text to Image Models Leaderboard [174], Accubits Program Synthesis Models Leaderboard [174], CLEM Leaderboard [175], Turkish LLM Leaderboard [176], Toloka LLM Leaderboard [177]

	GENIE - ARC-DA [120], AI2 GENIE - Summarization XSUM [120], AI2 MC-TACO [120], AI2 MOCHA [120], AI2 Natural Instructions [120], AI2 NYCC [120], AI2 OpenbookQA [120], AI2 PIQA [120], AI2 ProtoQA [120], AI2 QASC [120], AI2 Sherlock [120], AI2 SIQA [120], AI2 StrategyQA [120], AI2 WinoGrande [120], AI2 αNLI [120], Thai LLM Leaderboard [135], ML.ENERGY Leaderboard [112], Papers With Code [136], LLM-Leaderboard [137], Open Arabic LLM Leaderboard [138], Icelandic LLM leaderboard [139], Provider Leaderboard [140], Predibase Fine-Tuning Leaderboard [106], Accubits Open-source Large Language Models Leaderboard [141], Accubits InstructEval Leaderboard [141], Accubits			

	Business Friendly LLMs Leaderboard [141], Accubits Text to Image Models Leaderboard [141], Accubits Program Synthesis Models Leaderboard [141], Indic LLM Leaderboard [142], Julia LLM Leaderboard [143], LLM Rankings[114], LiveBench [145], all PWC leaderboards [136]			
"large multimodal model leaderboard"	Chatbot Arena [227], ConTextual [228], MULTI [229], Open LLM Leaderboard [230], MMMU [231], WildVision Arena [232], Open VLM Leaderboard [232], ConTextual [233], Open LLM Leaderboard [234], MTEB [234], Big Code Models Leaderboard [234], SEAL Leaderboards [234], Berkeley Function-Calling Leaderboard [234], Occiglot Euro LLM Leaderboard [234], Artificial	Chatbot Arena [227], ConTextual [228], Open LLM Leaderboard [234], MTEB [234], Big Code Models Leaderboard [234], SEAL Leaderboards [234], Berkeley Function-Calling Leaderboard [234], Occiglot Euro LLM Leaderboard [234], Artificial Analysis LLM API Providers Leaderboard [234], Artificial Analysis Models Leaderboard [234], Artificial Analysis Speech to	WildVision Arena [180], OpenVLM Leaderboard [180], Open LLM Leaderboard [181]	La Leaderboard [219], ConTextual [220], Open LLM Leaderboard [221], GAIA Leaderboard [222], MME [223], OpenCompass Large Language Model Leaderboard [223], OpenCompass Multi-modal Leaderboard [223], SEED-Bench [223], Multi-Modality Arena [223], Chatbot Arena [224], WildVision Arena [225], AgentBoard [226]

	Analysis LLM API Providers Leaderboard [234], Artificial Analysis Models Leaderboard [234], Artificial Analysis Speech to Text AI Model & Provider Leaderboard [234], Artificial Analysis Text to Image AI Model & Provider Leaderboard [234], Artificial Analysis Text to Speech AI Model & Provider Leaderboard [234], Open Medical LLM Leaderboard [234], Hughes Hallucination Evaluation Model Leaderboard [234], LLM-Perf Leaderboard [234], OpenCompass Large Language Model Leaderboard [117], OpenCompass Multi-modal Leaderboard [117], HELM AIR-Bench [101], HELM Classic [101], HELM CLEVA [101], HELM HEIM [101], HELM Image2Struct	Text AI Model & Provider Leaderboard [234], Artificial Analysis Text to Image AI Model & Provider Leaderboard [234], Artificial Analysis Text to Speech AI Model & Provider Leaderboard [234], Open Medical LLM Leaderboard [234], Hughes Hallucination Evaluation Model Leaderboard [234], LLM-Perf Leaderboard [234], MULTI [210], AlpacaEval [108]		

	[101], HELM Instruct [101], HELM Lite [101], HELM MMLU [101], HELM ThaiExam [101], HELM VHELM [101], Vellum LLM Leaderboard [110], MULTI [210], Aider LLM Leaderboards [127], AlpacaEval [108], LLM Leaderboard [122], all PWC leaderboards [136]			
"large action model leaderboard"	Berkeley Function Calling Leaderboard [203], Toolbench [203], AgentBoard [203], Open LLM Leaderboard [204], Chatbot Arena [204], CanAiCode Leaderboard [204], C-Eval [204], MTEB [204], Hallucinations Leaderboard [204], Big Code Models Leaderboard [204], EvalPlus [204], Enterprise Scenarios leaderboard [204], NPHardEval [204], ProLLM Benchmarks [204], OpenCompass Large Language	Berkeley Function Calling Leaderboard [203], Toolbench [203], AgentBoard [203], Open LLM Leaderboard [204], Chatbot Arena [204], CanAiCode Leaderboard [204], C-Eval [204], MTEB [204], Hallucinations Leaderboard [204], Big Code Models Leaderboard [204], EvalPlus [204], Enterprise Scenarios leaderboard [204], NPHardEval [204], ProLLM Benchmarks [204], OpenCompass Large Language	Open LLM Leaderboard [214], Chatbot Arena [215], AgentBench [216], Open Medical LLM Leaderboard [217], SEAL Leaderboards [218]	Berkeley Function-Calling Leaderboard [211], SEAL Leaderboards [212], Chatbot Arena [213]

	Model Leaderboard [204], OpenCompass Multi-modal Leaderboard [204], SEAL Leaderboards [204], AIR-bench [204], Vision-Arena [204], Aider LLM Leaderboard [204], RepoQA [204], BigCodeBench [204], Vellum LLM Leaderboard [204], EQBench [204], oobabooga [204], LiveBench [204], LiveCodeBench [204], SWE-bench [204], MMLU-Pro [204], LLM-Leaderboard [204], AlpacaEval [204], HELM AIR-Bench [204], HELM Classic [204], HELM CLEVA [204], HELM HEIM [204], HELM Image2Struct [204], HELM Instruct [204], HELM Lite [204], HELM MMLU [204], HELM ThaiExam [204], HELM VHELM [204], Chain-of-Thought Hub [204],	Model Leaderboard [204], OpenCompass Multi-modal Leaderboard [204], SEAL Leaderboards [204], AIR-bench [204], Vision-Arena [204], Aider LLM Leaderboard [204], RepoQA [204], BigCodeBench [204], Vellum LLM Leaderboard [204], EQBench [204], oobabooga [204], LiveBench [204], LiveCodeBench [204], SWE-bench [204], MMLU-Pro [204], LLM-Leaderboard [204], AlpacaEval [204], HELM AIR-Bench [204], HELM Classic [204], HELM CLEVA [204], HELM HEIM [204], HELM Image2Struct [204], HELM Instruct [204], HELM Lite [204], HELM MMLU [204], HELM ThaiExam [204], HELM VHELM [204], Chain-of-Thought Hub [204],		

	Artificial Analysis LLM API Providers Leaderboard [205], Artificial Analysis Models Leaderboard [205], Artificial Analysis Speech to Text AI Model & Provider Leaderboard [205], Artificial Analysis Text to Image AI Model & Provider Leaderboard [205], Artificial Analysis Text to Speech AI Model & Provider Leaderboard [205], Open Medical LLM Leaderboard [206], GAIA Leaderboard [207], Low-Bit Quantized Open LLM Leaderboard [208], BIRD [209], all PWC leaderboards [136]	Artificial Analysis LLM API Providers Leaderboard [205], Artificial Analysis Models Leaderboard [205], Artificial Analysis Speech to Text AI Model & Provider Leaderboard [205], Artificial Analysis Text to Image AI Model & Provider Leaderboard [205], Artificial Analysis Text to Speech AI Model & Provider Leaderboard [205], Open Medical LLM Leaderboard [206], GAIA Leaderboard [207], Low-Bit Quantized Open LLM Leaderboard [208]		

Accordingly, Table 4 provides the aggregated statistics showing the number of results retrieved and the number of FM leaderboards identified after applying our inclusion/exclusion criteria (*Section IV.B.2*). For example, a search on Google using the keyword "foundation model leaderboard" retrieved 138 webpages. After applying our inclusion/exclusion criteria, we identified 57 relevant FM leaderboards based on the webpage content. Similarly, the keyword "large language model leaderboard" returned 175 webpages, from which 82 FM leaderboards were identified. Similar interpretations apply to keywords "large action model leaderboard" and "large multimodal model leaderboard" from different sources, respectively. In particular, we did not count in any leaderboards hosted in PWC or HF Space, as these were already thoroughly collected in Phase 1 (Page 5, right column, line 18-21) of our initial draft.

Table 4: FM leaderboard statistics according to media type based on primary keyword.

Primary Keyword	**Number of FM Leaderboards via Google**	**Number of FM Leaderboards via Blogs (Google)**	**Number of FM Leaderboards via Videos (YouTube)**	**Number of FM Leaderboards via Tweets (X)**
"foundation model leaderboard"	138 webpages 57 leaderboards	7 blogs 7 leaderboards	589 videos 5 leaderboards	155 posts 3 leaderboards
"large language model leaderboard"	175 webpages 82 leaderboards (non-PWC)	14 blogs 32 leaderboards	609 videos 15 leaderboards (non-HF Space)	259 posts 45 leaderboards
"large multimodal model leaderboard"	283 webpages 40 leaderboards (non-PWC)	5 blogs 18 leaderboards	599 videos 3 leaderboards	110 posts 12 leaderboards
"large action model leaderboard"	166 webpages 53 leaderboards (non-PWC)	7 blogs 51 leaderboards	690 videos 5 leaderboards	99 posts 3 leaderboards

References

[1]
https://github.com/paperswithcode/paperswithcode-client/issues/24#issuecomment-2071784295
[2] https://rank.opencompass.org.cn
[3] https://huggingface.co/spaces/NexaAIDev/domain_llm_leaderboard
[4] https://www.superclueai.com
[5] https://huggingface.co/docs/hub/en/model-cards
[6] Kotti, Zoe, Rafaila Galanopoulou, and Diomidis Spinellis. "Machine learning for software engineering: A tertiary study." ACM Computing Surveys 55.12 (2023): 1-39.
[7] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
[8] https://lmarena.ai/?leaderboard
[9] https://www.superclueai.com
[10] https://rank.opencompass.org.cn
[11] https://crfm.stanford.edu/helm
[12] https://github.com/coderonion/awesome-llm-and-aigc
[13] https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models
[14] https://github.com/Atomic-man007/Awesome_Multimodel_LLM
[15] https://github.com/onejune2018/Awesome-LLM-Eval
[16] https://github.com/WLiK/LLM4Rec-Awesome-Papers
[17] https://github.com/CHIANGEL/Awesome-LLM-for-RecSys
[18] https://github.com/ActiveVisionLab/Awesome-LLM-3D
[19] https://github.com/yingpengma/Awesome-Story-Generation
[20] https://github.com/quqxui/Awesome-LLM4IE-Papers
[21] https://github.com/alelopes/awesome-rgbd-datasets
[22] https://github.com/tjunlp-lab/Awesome-LLMs-Evaluation-Papers
[23] https://github.com/horseee/Awesome-Efficient-LLM
[24] https://github.com/Jack-bo1220/Awesome-Remote-Sensing-Foundation-Models
[25] https://github.com/reasoning-survey/Awesome-Reasoning-Foundation-Models
[26] https://github.com/xephonhq/awesome-time-series-database
[27] https://github.com/alexklwong/awesome-state-of-depth-completion
[28] https://github.com/j-andrews7/awesome-bioinformatics-benchmarks
[29] https://github.com/Seyed-Ali-Ahmadi/Awesome_Satellite_Benchmark_Datasets
[30] https://github.com/awesomedata/awesome-public-datasets
[31] https://github.com/chrieke/awesome-satellite-imagery-datasets
[32] https://github.com/jdorfman/awesome-json-datasets
[33] https://github.com/minar09/awesome-virtual-try-on
[34] https://github.com/youngguncho/awesome-slam-datasets
[35] https://github.com/wq2012/awesome-diarization
[36] https://github.com/shramos/Awesome-Cybersecurity-Datasets
[37] https://github.com/jianzhnie/awesome-instruction-datasets
[38] https://github.com/LIANGKE23/Awesome-Knowledge-Graph-Reasoning
[39] https://github.com/shaypal5/awesome-twitter-data
[40] https://github.com/xiaobai1217/Awesome-Video-Datasets
[41] https://github.com/NEU-Gou/awesome-reid-dataset
[42] https://github.com/szenergy/awesome-lidar

[43] https://github.com/samapriya/awesome-gee-community-datasets
[44] https://github.com/modenaxe/awesome-biomechanics
[45] https://github.com/yueliu1999/Awesome-Deep-Graph-Clustering
[46] https://github.com/voidful/awesome-chatgpt-dataset
[47] https://github.com/leomaurodesenv/game-datasets
[48] https://github.com/pengxiao-song/awesome-chinese-legal-resources
[49] https://github.com/Daisy-Zhang/Awesome-Deepfakes-Detection
[50] https://github.com/lartpang/awesome-segmentation-saliency-dataset
[51] https://github.com/xahidbuffon/Awesome_Underwater_Datasets
[52] https://github.com/layumi/Vehicle_reID-Collection
[53] https://github.com/dspinellis/awesome-msr
[54] https://github.com/makinarocks/awesome-industrial-machine-datasets
[55] https://github.com/bytewax/awesome-public-real-time-datasets
[56] https://github.com/mint-lab/awesome-robotics-datasets
[57] https://github.com/ColinEberhardt/awesome-public-streaming-datasets
[58] https://github.com/XiaoxiaoMa-MQ/Awesome-Deep-Graph-Anomaly-Detection
[59] https://github.com/Psychic-DL/Awesome-Traffic-Agent-Trajectory-Prediction
[60] https://github.com/colour-science/awesome-colour
[61] https://github.com/zhilizju/Awesome-instruction-tuning
[62] https://github.com/OpenGene/awesome-bio-datasets
[63] https://github.com/oroszgy/awesome-hungarian-nlp
[64] https://github.com/glgh/awesome-llm-human-preference-datasets
[65] https://github.com/openfootball/awesome-football
[66] https://github.com/kjappelbaum/awesome-chemistry-datasets
[67] https://github.com/minwoo0611/Awesome-3D-LiDAR-Datasets
[68] https://github.com/wenhwu/awesome-remote-sensing-change-detection
[69] Sun, Jiankai, et al. "A survey of reasoning with foundation models." arXiv preprint arXiv:2312.11562 (2023).
[70] Zhou, Ce, et al. "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt." International Journal of Machine Learning and Cybernetics (2024): 1-65.
[71] Liu, Jiawei, et al. "Towards graph foundation models: A survey and beyond." arXiv preprint arXiv:2310.11829 (2023).
[72] Rawte, Vipula, Amit Sheth, and Amitava Das. "A survey of hallucination in large foundation models." arXiv preprint arXiv:2309.05922 (2023).
[73] Wang, Lei, et al. "A survey on large language model based autonomous agents." Frontiers of Computer Science 18.6 (2024): 186345.
[74] Cui, Can, et al. "A survey on multimodal large language models for autonomous driving." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024.
[75] Zhao, Wayne Xin, et al. "A survey of large language models." arXiv preprint arXiv:2303.18223 (2023).
[76] Xi, Zhiheng, et al. "The rise and potential of large language model based agents: A survey." arXiv preprint arXiv:2309.07864 (2023).
[77] Qiu, Jianing, et al. "Large ai models in health informatics: Applications, challenges, and the future." IEEE Journal of Biomedical and Health Informatics (2023).
[78] Lee, Ho Hin, et al. "Foundation models for biomedical image segmentation: A survey." arXiv preprint arXiv:2401.07654 (2024).

[79] Wu, Jiayang, et al. "Multimodal large language models: A survey." 2023 IEEE International Conference on Big Data (BigData). IEEE, 2023.

[80] Wang, Yiqi, et al. "Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning." arXiv preprint arXiv:2401.06805 (2024).

[81] Chang, Yupeng, et al. "A survey on evaluation of large language models." ACM Transactions on Intelligent Systems and Technology 15.3 (2024): 1-45.

[81] Qiao, Shuofei, et al. "Reasoning with language model prompting: A survey." arXiv preprint arXiv:2212.09597 (2022).

[82] Zhang, Shengyu, et al. "Instruction tuning for large language models: A survey." arXiv preprint arXiv:2308.10792 (2023).

[83] Shen, Tianhao, et al. "Large language model alignment: A survey." arXiv preprint arXiv:2309.15025 (2023).

[84] Kalyan, Katikapalli Subramanyam. "A survey of GPT-3 family large language models including ChatGPT and GPT-4." Natural Language Processing Journal 6 (2024): 100048.

[85] Zhu, Xunyu, et al. "A survey on model compression for large language models." Transactions of the Association for Computational Linguistics 12 (2024): 1556-1577.

[86] Zhang, Yue, et al. "Siren's song in the AI ocean: a survey on hallucination in large language models." arXiv preprint arXiv:2309.01219 (2023).

[87] Zhao, Pengyu, Zijian Jin, and Ning Cheng. "An in-depth survey of large language model-based artificial intelligence agents." arXiv preprint arXiv:2309.14365 (2023).

[88] Guo, Zishan, et al. "Evaluating large language models: A comprehensive survey." arXiv preprint arXiv:2310.19736 (2023).

[89] Gallegos, Isabel O., et al. "Bias and fairness in large language models: A survey." Computational Linguistics (2024): 1-79.

[90] Huang, Lei, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." ACM Transactions on Information Systems (2023).

[91] Yang, Zhenjie, et al. "A survey of large language models for autonomous driving." arXiv preprint arXiv:2311.01043 (2023).

[92] Feng, Zhangyin, et al. "Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications." arXiv preprint arXiv:2311.05876 (2023).

[93] Wang, Song, et al. "Knowledge editing for large language models: A survey." ACM Computing Surveys 57.3 (2024): 1-37.

[94] Wei, Shaopeng, et al. "Graph learning and its advancements on large language models: A holistic survey." arXiv preprint arXiv:2212.08966 (2022).

[95] Dong, Qingxiu, et al. "A survey on in-context learning." arXiv preprint arXiv:2301.00234 (2022).

[96] Rawte, Vipula, Amit Sheth, and Amitava Das. "A survey of hallucination in large foundation models." arXiv preprint arXiv:2309.05922 (2023).

[97] Cao, Boxi, et al. "The life cycle of knowledge in big language models: A survey." Machine Intelligence Research 21.2 (2024): 217-238.

[98] Fan, Mingyuan, et al. "On the trustworthiness landscape of state-of-the-art generative models: A comprehensive survey." arXiv preprint arXiv:2307.16680 (2023).

[99] Naveed, Humza, et al. "A comprehensive overview of large language models." arXiv preprint arXiv:2307.06435 (2023).

[100] https://www.linkedin.com/posts/eleutherai_open-ita-llm-leaderboard-a-hugging-face-activity-7161749334894612483-mRxb
[101] https://crfm.stanford.edu/helm
[102] https://vmlu.ai/leaderboard
[103] https://artificialanalysis.ai/leaderboards
[104] https://cevalbenchmark.com/static/leaderboard.html
[105] https://www.e2enetworks.com/blog/a-study-of-mmmu-massive-multi-discipline-multimodal-understanding-and-reasoning-benchmark-for-agi
[106] https://predibase.com
[107] https://github.com/XuanwuAI/SecEval
[108] https://snorkel.ai/blog/how-snorkel-topped-the-alpacaeval-leaderboard-and-why-we-re-not-there-anymore
[109] https://fanoutqa.com/leaderboard
[110] https://www.vellum.ai/llm-leaderboard
[111] https://discourse.julialang.org/t/ann-julia-llm-leaderboard-help-us-make-it-more-relevant-for-every-day-problems/107319
[112] https://ml.energy/leaderboard
[113] https://eqbench.com
[114] https://openrouter.ai/rankings
[115] https://lmsys.org/blog/2023-05-10-leaderboard
[116] https://toloka.ai/llm-leaderboard
[117] https://rank.opencompass.org.cn
[118] https://scale.com/leaderboard
[119] https://www.youtube.com/watch?v=UJy-zvWMs4g
[120] https://leaderboard.allenai.org
[121] https://huggingface.co/spaces/open-llm-leaderboard/blog
[122] https://klu.ai/llm-leaderboard
[123] https://www.nebuly.com/blog/llm-leaderboards
[124] https://somosnlp.org/blog/en/la-leaderboard
[125] https://aclanthology.org/2024.acl-long.177
[126] https://www.intel.com/content/www/us/en/developer/articles/technical/low-bit-quantized-open-llm-leaderboard.html
[127] https://aider.chat/docs/leaderboards
[128] https://ai-sandbox.list.lu/llm-leaderboard
[129] https://multilingual.com/european-llm-leaderboard-a-new-move-in-multilingual-ai-development
[130] https://ediscoverytoday.com/2024/01/05/hughes-hallucination-evaluation-h2em-model-leaderboard-artificial-intelligence-trends
[131] https://super.gluebenchmark.com/leaderboard
[132] https://codetlingua.github.io/leaderboard.html

[133] https://www.shakudo.io/blog/demystifying-llm-leaderboards-what-you-need-to-know
[134] https://www.anyscale.com/blog/comparing-llm-performance-introducing-the-open-source-leaderboard-for-llm
[135] https://blog.opentyphoon.ai/introducing-the-thaillm-leaderboard-thaillm-evaluation-ecosystem-508e789d06bf
[136] https://paperswithcode.com/sota
[137] https://llm-leaderboard.streamlit.app
[138] https://www.tii.ae/news/introducing-open-arabic-llm-leaderboard-empowering-arabic-language-modeling-community
[139] https://miðeind.is/en/greinar/islensk-stigatafla-risamallikana
[140] https://docs.withmartian.com/provider-leaderboard
[141] https://accubits.com/generative-ai-models-leaderboard
[142] https://www.cognitivelab.in/blog/introducing-indic-llm-leaderboard
[143] https://discourse.julialang.org/t/ann-julia-llm-leaderboard-help-us-make-it-more-relevant-for-every-day-problems/107319
[145] https://livebench.ai
[146] https://x.com/abacusai/status/1758174896117420400
[147] https://x.com/elij/status/1785225079468568847
[148] https://x.com/SCB10X_OFFICIAL/status/1843876598451450322
[149] https://x.com/rohanpaul_ai/status/1797993688175411679
[150] https://x.com/Thom_Wolf/status/1660649857437126656
[151] https://x.com/ai_bites/status/1785254953994399814
[152] https://x.com/aadityaura/status/1780951563584115167
[153] https://x.com/Marktechpost/status/1814856417465282916
[154] https://x.com/QuanquanGu/status/1745893623068004498
[155] https://x.com/arankomatsuzaki/status/1775357937558360164
[156] https://x.com/ceobillionaire/status/1798437390782279865
[157] https://x.com/JohnJGentry/status/1822436550829089083
[158] https://x.com/Analyticsindiam/status/1761700126517506230
[159] https://x.com/davilagrau/status/1814330609474355373
[160] https://x.com/Saboo_Shubham_/status/1675146981998567424
[161] https://x.com/ohmyshambles/status/1725034891941695966
[162] https://x.com/plotlygraphs/status/1677394376333336596
[163] https://x.com/tphuang/status/1734052676684423368
[164] https://x.com/yukitaylor00/status/1796341838158368835
[165] https://x.com/martin_fajcik/status/1841145773305549069
[166] https://x.com/JinaAI_/status/1792991297889767538
[167] https://x.com/WeidiXie/status/1827239188766560527
[168] https://x.com/mpyllan/status/1803992238466208032
[169] https://x.com/rohanpaul_ai/status/1782848099721089409
[170] https://x.com/TheTuringPost/status/1775659599384994149
[171] https://x.com/BAAIBeijing/status/1691681674927784432

[172] https://x.com/arxivsanitybot/status/1669693226243051521

[173] https://x.com/SCB10X_OFFICIAL/status/1834413201368498247

[174] https://x.com/accubits/status/1666460746086072320

[175] https://x.com/SherzodHakimov/status/1843979753113678106

[176] https://x.com/ccss_ku/status/1773418480894656991

[177] https://x.com/Al_Ramich_/status/1725127577763737876

[178] https://www.youtube.com/watch?v=Q4A_b9DP_3I

[179] https://youtu.be/tHHcgjYFtlo?si=1r3V3pER_AbZqGpw

[180] https://www.youtube.com/watch?v=UJy-zvWMs4g

[181] https://www.youtube.com/watch?v=aOjgPJ94-aM

[182] https://www.youtube.com/watch?v=Eb0Ga5igBuQ

[183] https://www.youtube.com/watch?v=-ZospyglS1Y

[184] https://www.youtube.com/watch?v=VYDbEQU__7I

[185] https://www.youtube.com/watch?v=h4uQF1WQRbI

[186] https://www.youtube.com/shorts/5tNiolZGP5Y

[187] https://www.youtube.com/watch?v=CzVBeEd9YV8

[188] https://www.youtube.com/watch?v=QqCLfVYYCDI

[189] https://www.youtube.com/watch?v=WbQh2Ur6H0w

[190] https://production-media.paperswithcode.com/about/evaluation-tables.json.gz

[191] https://pages.github.com

[192] https://github.com/sourcegraph/src-cli

[193] https://github.com/SAILResearch/awesome-foundation-model-leaderboards

[194] https://www.zdnet.com/article/what-google-chromes-incognito-mode-really-does-and-doesnt-do-for-you

[195] https://www.reddit.com/r/SEO/comments/y66ejl/google_displays_a_lot_less_results_than_shown_on

[196] https://redarena.ai/leaderboard

[197] https://lmarena.ai/?leaderboard

[198] https://rank.opencompass.org.cn

[199] https://www.superclueai.com

[200] Ott, Simon, et al. "Mapping global dynamics of benchmark creation and saturation in artificial intelligence." Nature Communications 13.1 (2022): 6793.

[201] Maslej, Nestor, et al. "Artificial intelligence index report 2023." arXiv preprint arXiv:2310.03715 (2023).

[202] Polo, Felipe Maia, et al. "tinyBenchmarks: evaluating LLMs with fewer examples." arXiv preprint arXiv:2402.14992 (2024).

[203] https://www.salesforce.com/blog/large-action-model-ai-agent

[204] https://www.reddit.com/r/LocalLLaMA/comments/144rg6a/all_model_leaderboards_that_i_know/?rdt=54355

[205] https://www.reddit.com/r/ClaudeAI/comments/1g79pea/what_are_best_ai_model_leaderboards_score_tables

[206] https://www.shakudo.io/blog/demystifying-llm-leaderboards-what-you-need-to-know

[207] https://www.redcellpartners.com/perspectives/trase-tops-gaia-leaderboard
[208] https://www.intel.com/content/www/us/en/developer/articles/technical/low-bit-quantized-open-llm-leaderboard.html
[209] https://bird-bench.github.io
[210] https://synthical.com/article/7f892719-c969-452e-870c-a28d494b5b59
[211] https://x.com/huan__wang/status/1835379458427252742
[212] https://x.com/alexandr_wang/status/1831406334518149220
[213] https://x.com/lmarena_ai/status/1778555678174663100
[214] https://www.youtube.com/watch?v=7Pj_CikLex4
[215] https://www.youtube.com/watch?v=ieoeyEfUfXc
[216] https://www.youtube.com/watch?v=EiFVJUFiRVQ
[217] https://www.youtube.com/watch?v=Eb0Ga5igBuQ
[218] https://www.youtube.com/watch?v=NlXpT_8ScYk
[219] https://x.com/SomosNLP_/status/1838906278187241546
[220] https://x.com/sooperset/status/1765200789310742969
[221] https://x.com/Fails_us/status/1830068251919000022
[222] https://x.com/JaynitMakwana/status/1859157734240993287
[223] https://x.com/violetto96/status/1744971246029406421
[224] https://x.com/dchaplot/status/1745139082886279398
[225] https://x.com/yujielu_10/status/1755819380880122248
[226] https://x.com/ma_chang_nlp/status/1750369056539218082
[227] https://lmsys.org/blog/2024-06-27-multimodal
[228] https://www.reddit.com/r/LocalLLaMA/comments/1b78w1z/new_leaderboard_on_hf_multimodal_reasoning
[229] https://github.com/OpenDFM/MULTI-Benchmark
[230] https://huggingface.co/open-llm-leaderboard
[231] https://mmmu-benchmark.github.io
[232] https://restack.io/p/computer-vision-answer-visual-language-model-leaderboard-cat-ai
[233] https://con-textual.github.io
[234] https://blog.monsterapi.ai/blogs/top-12-llm-leaderboards-to-help-you-choose-the-right-model
[235] https://www.youtube.com/watch?v=pXtvFEZ6drA
[236] https://huggingface.co/spaces/junkim100/self-improving-leaderboard
[237] https://huggingface.co/spaces/logikon/open_cot_leaderboard
[238] https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard
[239] https://somosnlp.org/blog/en/la-leaderboard
[240] https://ai-sandbox.list.lu/llm-leaderboard
[241] https://hkust-nlp.github.io/agentboard
[242] https://vmlu.ai/leaderboard
[243] https://huggingface.co/spaces/mideind/icelandic-llm-leaderboard
[244] https://docs.withmartian.com/provider-leaderboard/
[245] https://accubits.com/large-language-models-leaderboard
[246] https://accubits.com/instructeval-leaderboard
[247] https://accubits.com/business-friendly-llms-leaderboard

[248] https://accubits.com/text-to-image-models-leaderboard
[249] https://www.swebench.com
[250] https://huggingface.co/spaces/CZLC/BenCzechMark
[251] https://henrychur.github.io/MedS-Bench/
[252] https://jailbreakbench.github.io/#leaderboard
[253] https://huggingface.co/spaces/allenai/ZebraLogic
[254] https://prollm.toqan.ai/leaderboard
[255] https://www.vellum.ai/llm-leaderboard