Lifelong Benchmarking

Adhiraj Ghosh, Sebastian Dziadzio, Ameya Prabhu, Vishaal Udandarao,
Matthias Bethge, Samuel Albanie

18.06.2024

@sbdzdz

Motivation

Benchmarking then

Measure generalization beyond the training set

Training, validation, and test set

Task specific (classification, segmentation, retrieval)

Driving progress

Benchmarking now

Measure zero-shot capabilities

Test set only (models train on vast web datasets)

Arbitrary tasks (question answering, image captioning, classification)

Driving progress

Leaderboards and arenas

📋 Leaderboards

- Aggregate performance across a set of datasets
- Curated and static
- Prone to hill-climbing and data leaks

⚔️ Arenas

- Crowd-source data instances and ratings
- General and dynamic
- Prone to evaluator bias and low quality data

Lifelong benchmarks

	Leaderboards	Arenas	Lifelong Benchmarks
Curated	✅	❌	✅
Open	✅	❌	✅
Dynamic	❌	✅	✅
Easy update	❌	✅	✅
Granular	❌	❌	✅
Measurements	Cardinal	Ordinal	Both

Framework

Lifelong benchmarks: components

Models

Data instances

- Atomic unit of data
- Image, prompt, question, references
- Metadata: tags, source

Measurements

- Binary, numerical, comparative
- Link one data instance with one or more models
- Metadata: metric, procedure

Ideal world

	Instance ID: cd8a0s Tags: maths Metric: accuracy	Instance ID: a1e892 Tags: captioning Metric: accuracy	Instance ID: c5880a Tags: physics Metric: accuracy	Instance ID: b43a83 Tags: law Metric: accuracy
GPT-4	1	1	0	0
Llama 3	0	0	0	1
Claude 3	0	1	1	1
Mistral 7B	1	0	0	0

Our world: heterogeneous

	Instance ID: cd8a0s Tags: maths Metric: accuracy	Instance ID: a1e892 Tags: captioning Metric: BLEU	Instance ID: c5880a Tags: physics Metric: accuracy	Instance ID: b43a83 Tags: law Metric: comparison
GPT-4	1	0.372	0	GPT-4 < Llama 3, GPT-4 > Mistral 7B
Llama 3	0	0.101	0	Llama 3 < GPT-4 Llama 3 > Claude 3
Claude 3	0	0.215	1	Claude 3 < Llama 3 Claude 3 < Mistral 7B
Mistral 7B	1	0.854	0	Mistral 7B > Claude 3 Mistral 7B < GPT-4

Our world: incomplete

	Instance ID: cd8a0s Tags: maths Metric: accuracy	Instance ID: a1e892 Tags: captioning Metric: BLEU	Instance ID: c5880a Tags: physics Metric: accuracy	Instance ID: b43a83 Tags: law Metric: comparison
GPT-4	1	0.372	N/A	GPT-4 > Mistral 7B
Llama 3	N/A	0.101	0	N/A
Claude 3	0	N/A	1	N/A
Mistral 7B	1	0.854	N/A	Mistral 7B < GPT-4

Solution: ordinal measurements

	Instance ID: cd8a0s Tags: maths Metric: accuracy	Instance ID: a1e892 Tags: captioning Metric: BLEU	Instance ID: c5880a Tags: physics Metric: accuracy	Instance ID: b43a83 Tags: law Metric: comparison
GPT-4	GPT-4 > Claude 3	GPT-4 > Llama 3 GPT-4 < Mistral 7B	N/A	GPT-4 > Mistral 7B
Llama 3	N/A	Llama 3 < GPT-4 Llama 3 < Mistral 7B	Llama 3 < Claude 3	N/A
Claude 3	Claude 3 < GPT-4 Claude 3 < Mistral 7B	N/A	Claude 3 > Llama 3	N/A
Mistral 7B	Mistral 7B > Claude 3	Mistral 7B > Llama 3 Mistral 7B > GPT-4	N/A	Mistral 7B < GPT-4

Implementation

Measurements

Lifelong LLM Benchmark

Benchmark	Datasets
HELM (5K)	NarrativeQA (F1) NaturalQuestions (F1) OpenbookQA (EM) MMLU (EM) GSM8K (QEM) MATH (QEM) LegalBench (QEM) MedQA (QEM) WMT 2014 (BLEU)
Open LLM Leaderboard (28K)	ARC (EM) HellaSwag (EM) MMLU (EM) TruthfulQA (EM) Winogrande (EM) GSM8K (EM)
Chatbot Arena (51K)	Chatbot Arena (Ordinal)

Lifelong LMM Benchmark

Benchmark	Datasets
VHELM (30K)	A-OKVQA Bingo (ROUGE) COCO (ROUGE) Crossmodal-3600 (ROUGE) Flickr30k (ROUGE) GQA (QEM) Hateful Memes(QEM) MathVista (QEM) Mementos (GPT4) MME (QEM) MMMU (QEM) MultipanelVQA (QEM) OODCV-VQA (QEM) PAIRS (QEM) POPE(QEM) SEED-Bench(QEM) Sketchy-VQA (QEM) VizWiz (QEM) VQAv2 (QEM)
LMMs-Eval (540K)	AI2D (QEM) ChartQA (QEM) CMMMU (QEM) COCO (ROUGE) DocVQA (ANLS) Flickr30k (ROUGE) GQA (QEM) IconQA (ANLS) LLaVA-Wild (GPT) MathVista (GPT) MMBench (GPT) MME (C,P) MMMU (QEM) MMVET (GPT) MP-DocVQA (QEM) NoCaps (ROUGE) OK-VQA (ANLS) POPE (EM) RefCOCO (ROUGE) ScienceQA (EM) SEED-Bench (QEM) SEED-Bench-2 (QEM) TextCaps (ROUGE) TextVQA (EM) VizWiz (EM) VQAv2 (EM)
WildVision Arena (9K)	Vision Arena (Ordinal)

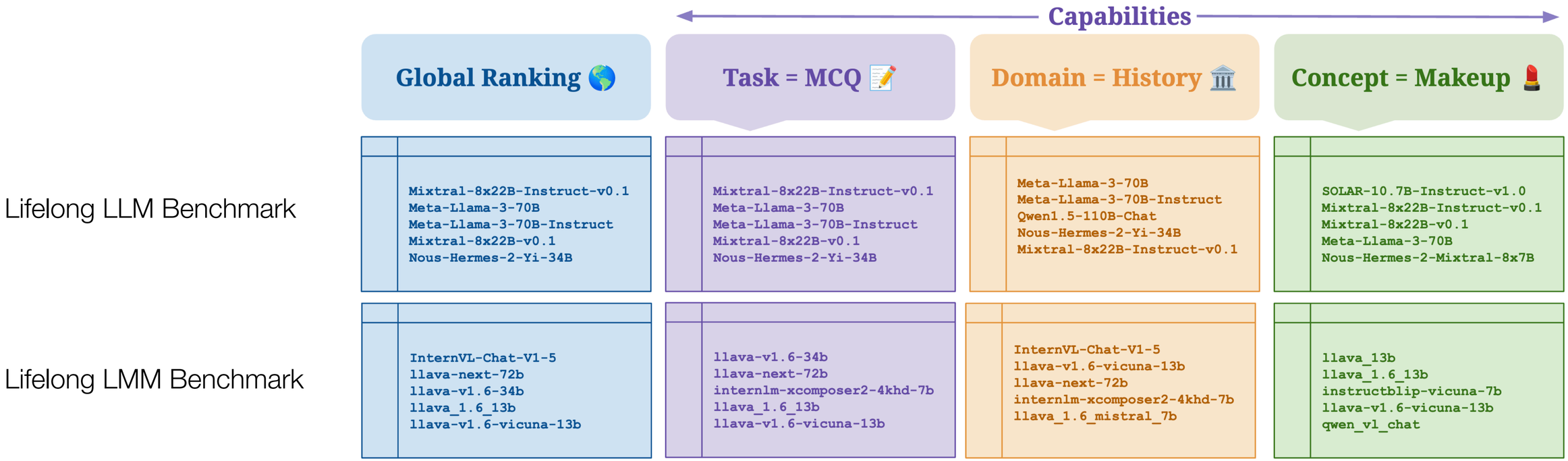
Data instances


```
"00152723d9a14a15891d90b2001b35d3": {
    "metadata": {
        "source": "wmt_2014",
        "subject": "news",
        "task": "translation"
    },
    "question": {
        "text": "Bundesberufungsgericht blockiert Entscheidung einer Ric
    },
    "references": [
        {
            "output": {
                "text": "A federal appeals court blocks a judge's ruling
            },
            "tags": [
                "correct"
            ]
        }
    ]
},
```

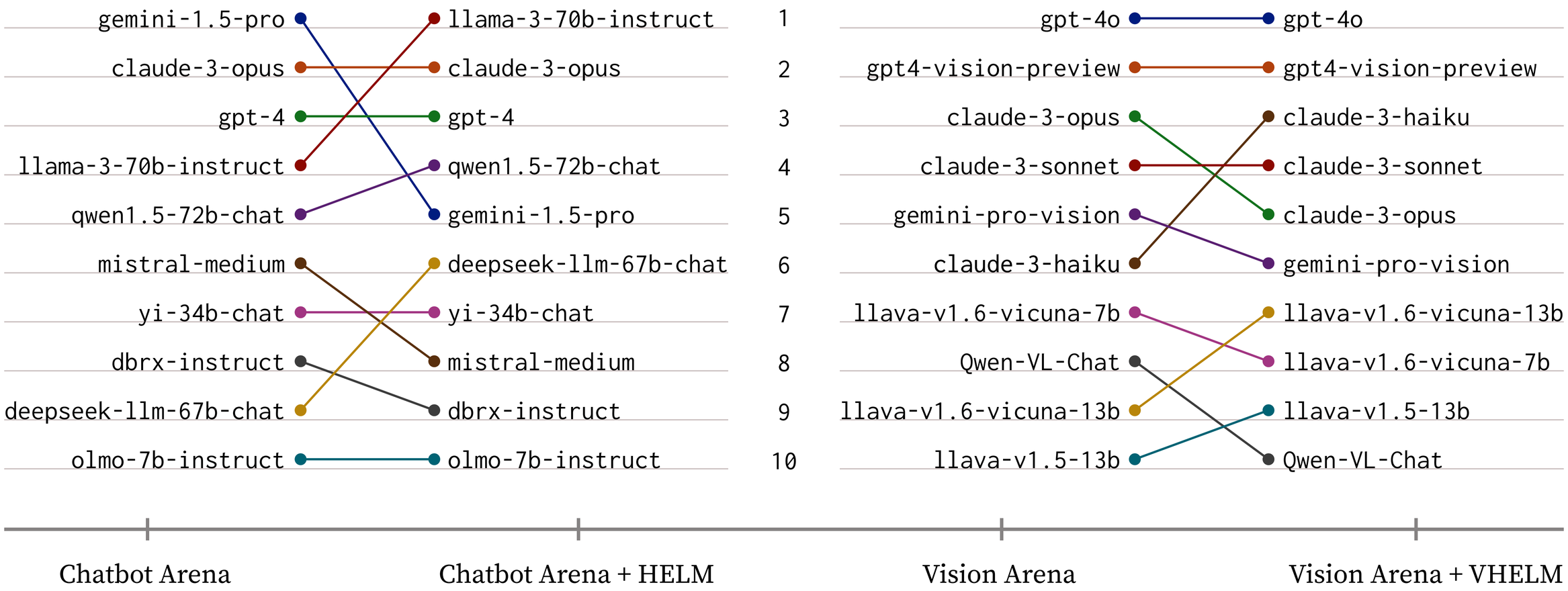
```
"00006f69e5db49099c968331eea57250": {
    "metadata": {
        "source": "natural_qa",
        "subject": "music",
        "task": "qa"
    },
    "question": {
        "text": "Who sings the theme song for living single?'
    },
    "references": [
        {
            "output": {
                "text": "Queen Latifah"
            },
            "tags": [
                "correct"
            ]
        }
    ]
},
```

```
"648fb03768ea4ce596ebca909a835494": {
    "metadata": {
        "source": "allenai/winogrande"
    },
    "question": {
        "text": "Terry tried to bake the eggplant in the toaster oven but the _ was
    },
    "references": [
        {
            "tags": [],
            "text": "eggplant"
        },
        {
            "tags": [],
            "text": "toaster"
        }
    ]
},
```

Capability probing



Ranking


gemini-1.5-pro
claude-3-opus
gpt-4
llama-3-70b-instruct
qwen1.5-72b-chat
mistral-medium
yi-34b-chat
dbrx-instruct
deepseek-llm-67b-chat
olmo-7b-instruct
llama-3-70b-instruct
claude-3-opus
gpt-4
qwen1.5-72b-chat
gemini-1.5-pro
deepseek-llm-67b-chat
yi-34b-chat
mistral-medium
dbrx-instruct
olmo-7b-instruct
1
2
3
4
5
6
7
8
9
10
gpt-4o
gpt4-vision-preview
claude-3-opus
claude-3-sonnet
gemini-pro-vision
claude-3-haiku
llava-v1.6-vicuna-7b
Qwen-VL-Chat
llava-v1.6-vicuna-13b
llava-v1.5-13b
gpt-4o
gpt4-vision-preview
claude-3-haiku
claude-3-sonnet
claude-3-opus
gemini-pro-vision
llava-v1.6-vicuna-13b
llava-v1.6-vicuna-7b
llava-v1.5-13b
Qwen-VL-Chat
Chatbot Arena
Chatbot Arena + HELM
Vision Arena
Vision Arena + VHELM

Future directions

Crowdsource data instances and measurements

Improve the querying system

Prune old, easy, or contaminated data instances

Automate the pipeline

Run the evaluation

Thank you!

@sbdzdz