LMMs-Lab

Building Multimodal Intelligence

Ziwei Liu

Nanyang Technological University, Singapore

https://lmms-lab.github.io/

LMMs-Lab

Building multimodal intelligence

- ❑ LMMs-Eval
- ❑ LLaVA-NeXT
- ❑ LLaVA-NeXT Video

LMMs-Eval

LMMs-Eval Framework Provides:

- ❑ Holistic, Standardized and Reproducible Evaluation.
- ❑ A Unified Interfaces for Multimodal Tasks and Models.

Holistic, Standardized and Reproducible Evaluation Guides AI

Holistic evaluation is necessary

- *More evaluations from different dimensions could better reflect model's overall performance.*

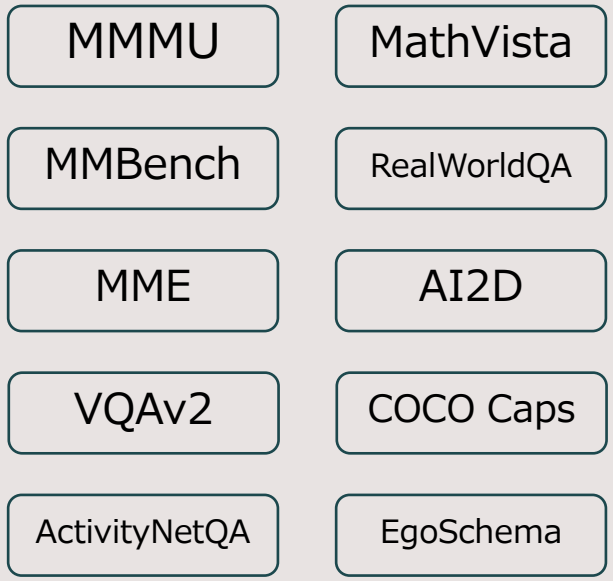
Method	FT	Shot	OKVQA (I)	VQAv2 (I)	COCO (I)	MSVDQA (V)	VATEX (V)	VizWiz (I)	Flick30K (I)	MSRVTTQA (V)	iVQA (V)	YouCook2 (V)	STAR (V)	VisDial (I)	TextVQA (I)	NextQA (I)	HatefulMemes (I)	RareAct (V)
Zero/Few shot SOTA	✗	(X)	[34] 43.3 (16)	[114] 38.2 (4)	[124] 32.2 (0)	[58] 35.2 (0)	-	-	-	[58] 19.2 (0)	[135] 12.2 (0)	-	[143] 39.4 (0)	[79] 11.6 (0)	-	-	[85] 66.1 (0)	[85] 40.7 (0)
Flamingo-3B	✗	0	41.2	49.2	73.0	27.5	40.1	28.9	60.6	11.0	32.7	55.8	39.6	46.1	30.1	21.3	53.7	58.4
	✗	4	43.3	53.2	85.0	33.0	50.0	34.0	72.0	14.9	35.7	64.6	41.3	47.3	32.7	22.4	53.6	-
	✗	32	45.9	57.1	99.0	42.6	59.2	45.5	71.2	25.6	37.7	76.7	41.6	47.3	30.6	26.1	56.3	-
Flamingo-9B	✗	0	44.7	51.8	79.4	30.2	39.5	28.8	61.5	13.7	35.2	55.0	41.8	48.0	31.8	23.0	57.0	57.9
	✗	4	49.3	56.3	93.1	36.2	51.7	34.9	72.6	18.2	37.7	70.8	**42.8**	50.4	33.6	24.7	62.7	-
	✗	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	50.4	32.6	28.4	63.5	-
Flamingo	✗	0	50.6	56.3	84.3	35.6	46.7	31.6	67.2	17.4	40.7	60.1	39.7	52.0	35.0	26.7	46.4	**60.8**
	✗	4	57.4	63.1	103.2	41.7	56.0	39.6	75.1	23.9	44.1	74.5	42.4	**55.6**	36.5	30.8	68.6	-
	✗	32	**57.8**	**67.6**	**113.8**	**52.3**	**65.1**	**49.8**	**75.4**	**31.0**	**45.3**	**86.8**	42.2	**55.6**	**37.9**	**33.5**	**70.0**	-
Pretrained FT SOTA	✔	(X)	54.4 [34] (10K)	80.2 [140] (444K)	143.3 [124] (500K)	47.9 [28] (27K)	76.3 [153] (500K)	57.2 [65] (20K)	67.4 [150] (30K)	46.8 [51] (130K)	35.4 [135] (6K)	138.7 [132] (10K)	36.7 [128] (46K)	75.2 [79] (123K)	54.7 [137] (20K)	25.2 [129] (38K)	79.1 [62] (9K)	-

Table 1: **Comparison to the state of the art.** A *single* Flamingo model reaches the state of the art on a wide array of image **(I)** and video **(V)** understanding tasks with few-shot learning, significantly outperforming previous best zero- and few-shot methods with as few as four examples. More importantly, using only 32 examples and without adapting any model weights, Flamingo *outperforms* the current best methods – fine-tuned on thousands of annotated examples – on seven tasks. Best few-shot numbers are in **bold**, best numbers overall are underlined.

Flamingo model was (at 2022) a state-of-the-art multimodal model on multiple datasets across image and video modalities.

Holistic, Standardized and Reproducible Evaluation Guides AI

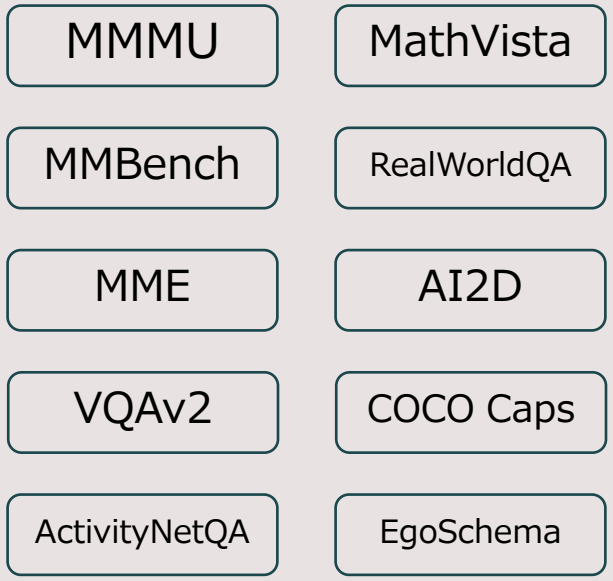
Standardized and reproducible evaluation is necessary



Previously, developers collect and evaluate on each task.
Non-standardized tasks make reproducibility and comparison with prior work expensive (and sometimes impossible)

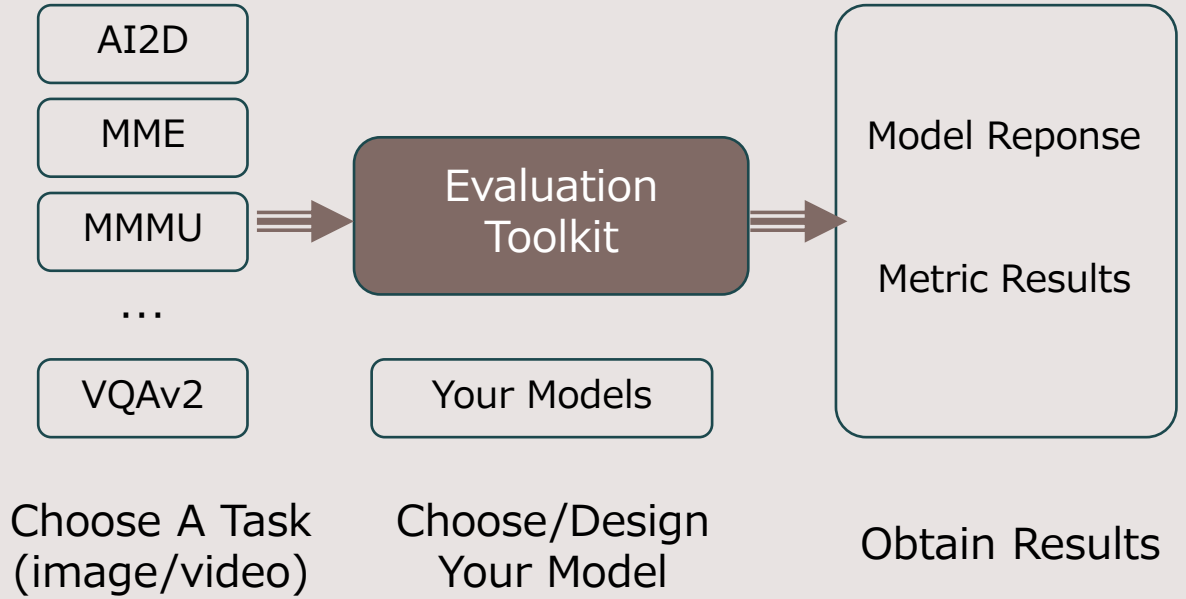
Holistic, Standardized and Reproducible Evaluation Guides AI

Benchmarks emerge fast. Everyday, we have new benchmarks.



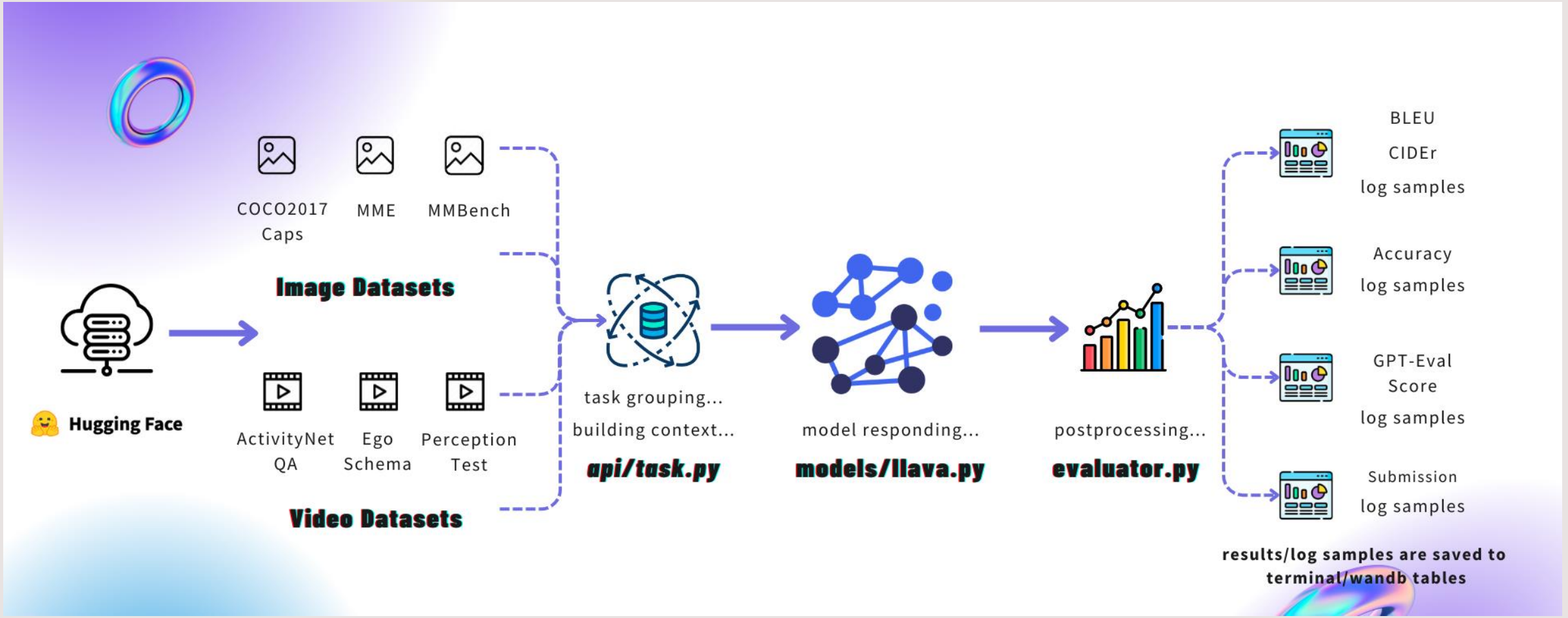
Previously, developers collect and evaluate on each task.

Non-standardized tasks make reproducibility and comparison with prior work expensive (and sometimes impossible)

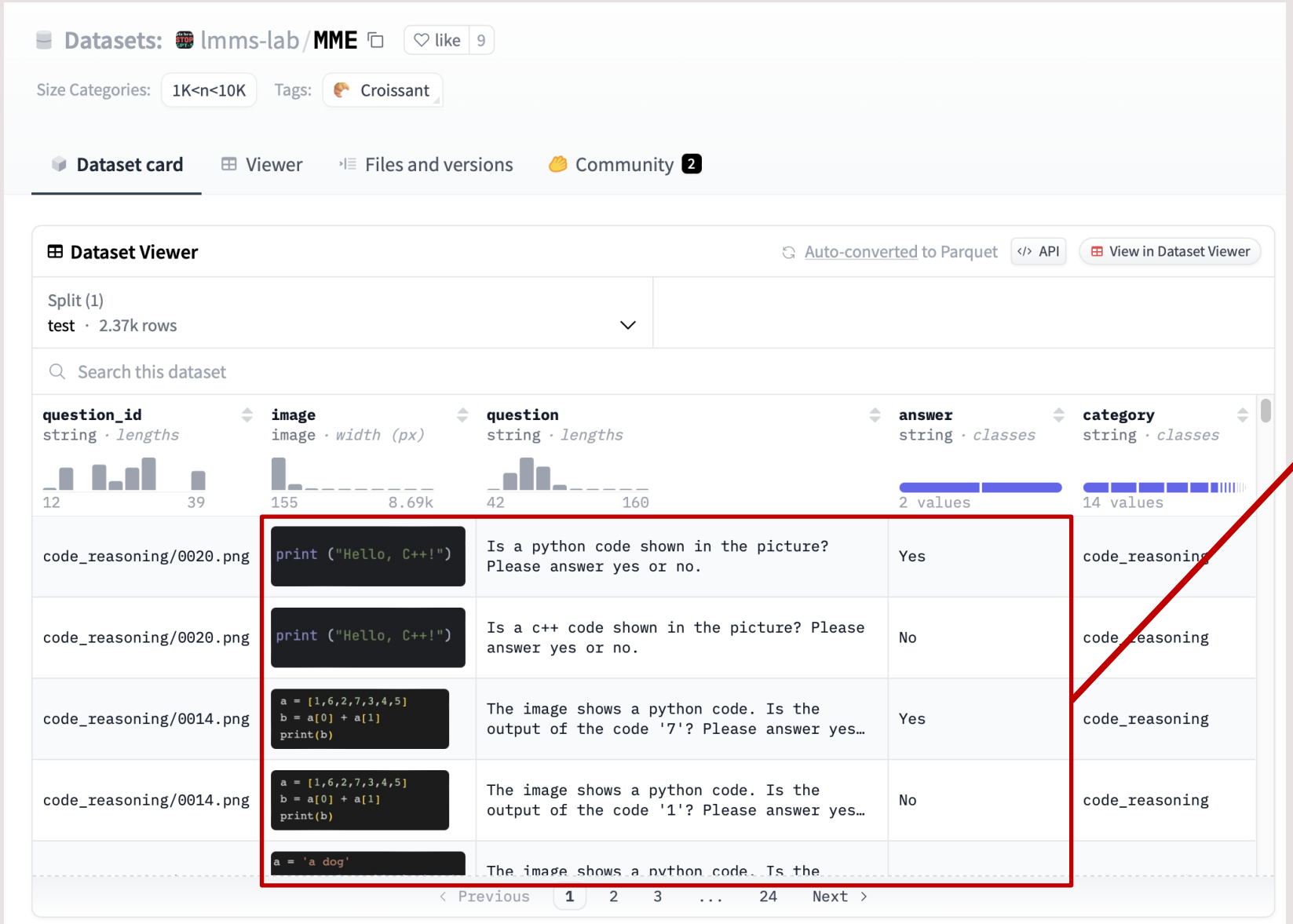


We wish to have a unified evaluation toolkit that defines the interface for multimodal tasks and models.

LMMs-Eval is an **efficient, standardized and reproducible** evaluation framework for accelerating the development for large-scale multimodal models



LMMs-Eval: Unified interface for multimodal tasks and models.



LMMs-Eval defines the evaluation dataset format.

It requires the necessary items fully visible on Huggingface, enabling developers to check and have the direct sense on evaluation dataset.

LMMs-Eval: Unified interface for multimodal tasks and models.

```yaml
dataset_path: lmms-lab/MME
dataset_kwargs:
  token: True
task: "mme"
test_split: test
output_type: generate_until
doc_to_visual: !function utils.mme_doc_to_visual
doc_to_text: !function utils.mme_doc_to_text
doc_to_target: "answer"
generation_kwargs:
  max_new_tokens: 16
  temperature: 0
  top_p: 1.0
  num_beams: 1
  do_sample: false
# The return value of process_results will be used by metrics
process_results: !function utils.mme_process_results
# Note that the metric name can be either a registed metric function (such
metric_list:
  - metric: mme_percetion_score
    aggregation: !function utils.mme_aggregate_results
    higher_is_better: true
  - metric: mme_cognition_score
    aggregation: !function utils.mme_aggregate_results
    higher_is_better: true
model_specific_prompt_kwargs:
  default:
    pre_prompt: ""
    post_prompt: "\nAnswer the question using a single word or phrase."
  gpt4v:
    pre_prompt: ""
    post_prompt: "\nAnswer the question with Yes or No."
  qwen_vl:
    pre_prompt: ""
    post_prompt: " Answer:"
```

LMMs-Eval defines the task interfaces via

- dataset metainfo
- generation_kwargs *(some tasks require shorter answer)*
- multiple evaluation metrics
- model specific prompts *(some models need specific prompt to better match the answer)*

LMMs-Eval: Unified interface for multimodal tasks and models.

The lmms class enforces a common interface via which we can extract responses from a model:

```
class MyCustomLM(lmms):
    #...
    def loglikelihood(self, requests: list[Instance]) -> list[tuple[float, bool]]:
        #...

    def generate_until(self, requests: list[Instance]) -> list[str]:
        #...
    #...
```

Where `Instance` is a dataclass defined in `lmms_eval.api.instance` with property `args` of request-dependent type signature described below.

We support three types of requests, consisting of different interactions / measurements with an autoregressive LM.

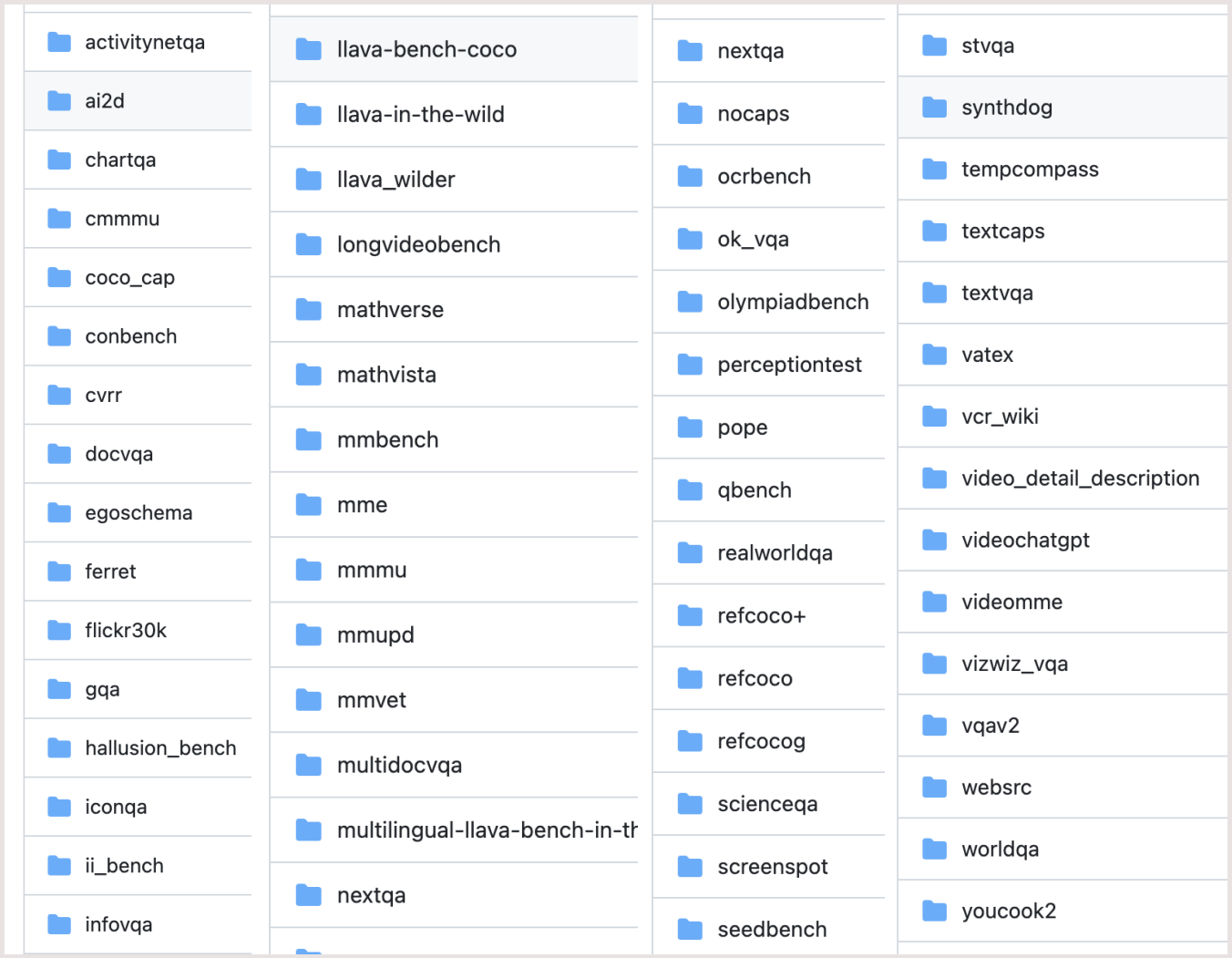
All three request types take as input `requests` of type `list[Instance]` that have a matching `Instance.request_type` to the method name. Overall, you can check the construct_requests to see how the arguments are being constructed for different types of output type requests.

- `generate_until`
  - Each request contains `Instance.args : Tuple[str, dict]` containing 1. an input string to the LM and 2. a dictionary of keyword arguments used to control generation parameters.
  - In each `Instance.args` there will be 6 elements which are `contexts, all_gen_kwargs, doc_to_visual, doc_id, task, split`. `contexts` refers to the formatted question and is the text input for the LMM. Sometimes it might contains image token and need to address differently for different models. `all_gen_kwargs` refers to the dict that contains all the generation configuration for the model. We use `doc_id`, `task`, and `split` to access the dataset and then you can use `doc_to_visual` which is a function reference to process the image. When you implement your own model, you should use these to write your own generate_util function.
  - Using this input and these generation parameters, text will be sampled from the language model (typically until a maximum output length or specific stopping string sequences--for example, `{"until": ["\n\n", "."], "max_gen_toks": 128}`).
  - The generated input+output text from the model will then be returned.

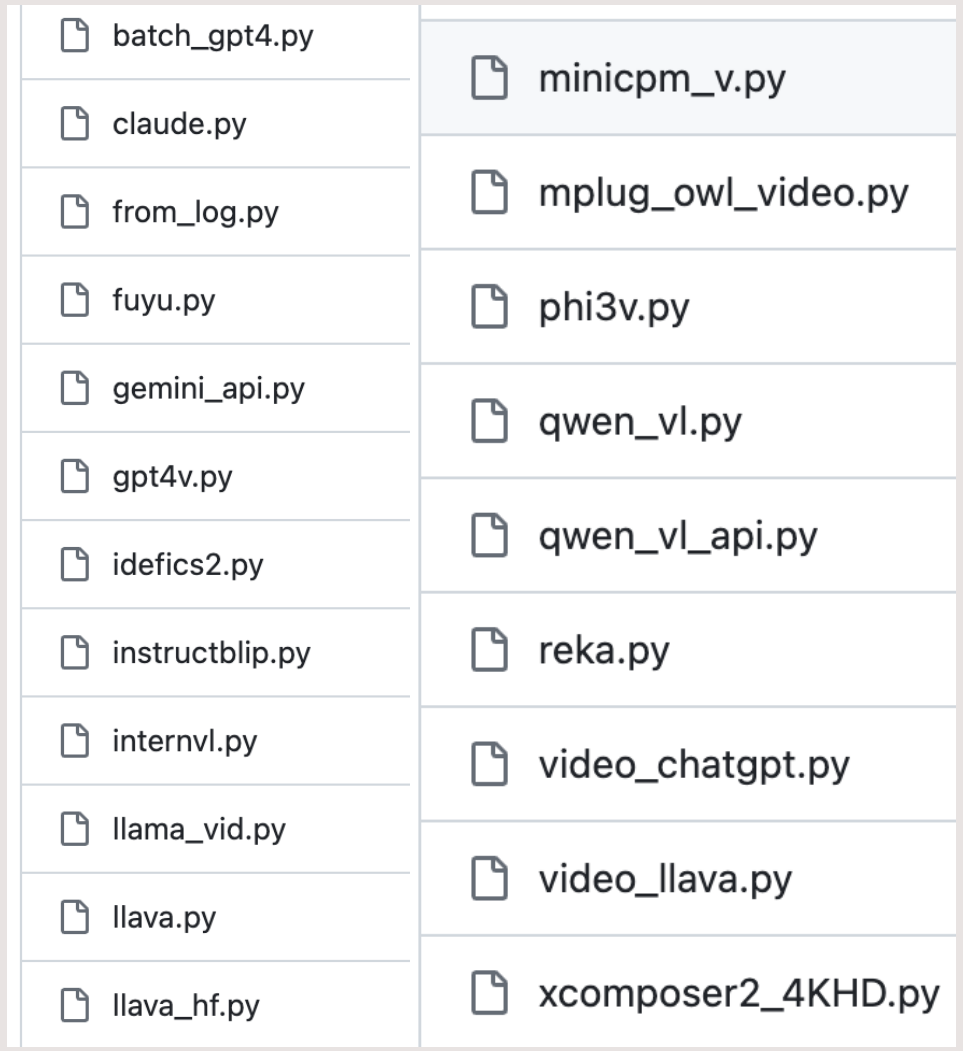
We abstract the models and allows:

- integrated implementation
- multi-gpu evaluation
- tensor parallelism evaluation
  (for 72b/110b models, or 34b models on limited-memory GPUs)

LMMs-Eval: Unified interface for multimodal tasks and models.



(a) Tasks



(b) Models

LMMs-Eval integrates **50+ image tasks**, **10+ video tasks** and **~10 state-of-the-art level LMM models.**

LMMs-Eval: Unified interface for multimodal tasks and models.

Evaluation of LLaVA on multiple datasets

```
python3 -m accelerate.commands.launch \
    --num_processes=8 \
    -m lmms_eval \
    --model llava \
    --model_args pretrained="liuhaotian/llava-v1.5-7b" \
    --tasks mme,mmbench_en \
    --batch_size 1 \
    --log_samples \
    --log_samples_suffix llava_v1.5_mme_mmbenchen \
    --output_path ./logs/
```

For other variants llava. Please change the `conv_template` in the `model_args`

> `conv_template` is an arg of the init function of llava in `lmms_eval/models/llava.py`, you could find the corresponding value at LLaVA's code, probably in a dict variable `conv_templates` in `llava/conversations.py`

```
python3 -m accelerate.commands.launch \
    --num_processes=8 \
    -m lmms_eval \
    --model llava \
    --model_args pretrained="liuhaotian/llava-v1.6-mistral-7b,conv_template=mistral_instruct"
    --tasks mme,mmbench_en \
    --batch_size 1 \
    --log_samples \
    --log_samples_suffix llava_v1.5_mme_mmbenchen \
    --output_path ./logs/
```

LMMs-Eval is

- *One-line installation.*
- *One-line evaluation.*

LMMs-Eval: Unified interface for multimodal tasks and models.

	A	B	C	D	E	F	G	H	I	J	K	L
1	LMMs-Eval				LLaVA-1.5				LLaVA-1.6			
2	Datasets	Meta Info			1.5-7B (report)	1.5-7B (lmms-eval)	1.5-13B (report)	1.5-13B (lmms-eval)	1.6-7B (lmms-eval)	1.6-7B (lmms-eval)	1.6-13B (lmms-eval)	1.6-34B (lmms-eval)
3		Split	Metric	#Num	liuhaotian/llava-v1.5-7b	liuhaotian/llava-v1.5-7b	liuhaotian/llava-v1.5-13b	liuhaotian/llava-v1.5-13b	liuhaotian/llava-v1.6-mistral-7b	liuhaotian/llava-v1.6-vicuna-7b	liuhaotian/llava-v1.6-vicuna-13b	liuhaotian/llava-v1.6-34b
4	AI2D	test	Acc	3,088	-	54.79	-	59.49	60.75	66.58	70.04	74.94
5	ChartQA	test	RelaxedAcc	2,500	-	18.24	-	18.20	38.76	54.84	62.2	68.72
6	CMMMU	val	Acc	900	-	21.80	-	26.30	22.7	24	23,2	39.9
7	COCO-Cap	cococap_val_2014	CIDEr	40,504	-	108.66	-	113.88	107.66	96.98	99.45	103.16
8	COCO-Cap	cococap_val_2017	CIDEr	5,000	-	110.38	-	115.61	109.22	99.93	101.99	105.89
9	DocVQA	val	ANLS	5,349	-	28.08	-	30.29	72.16	74.35	77.45	83.98
10	Flickr	-	CIDEr	31,784	-	74.93	-	79.59	73.14	68.44	66.7	68.48
11	GQA	gqa_eval	Acc	12,578	62.00	61.97	63.30	63.24	54.98	64.23	65.36	67.08
12	Hallusion-Bench	test	All Acc.	951		44.90		42.27	41.74	41.53	44.47	
13	InfoVQA	val	ANLS	2,801	-	25.81	-	29.35	43.77	37.09	41.34	51.45
14	LLaVA-W	test	GPT-Eval-Avg	60	63.40	65.3 (0314) 59.6 (0613)	-	72.8 (0314) 66.1 (0613)	71.7 (0613)	72.3 (0613)	72.3 (0613)	
15	MathVista	testmini	Acc	1,000	27.40	26.70	27.60	26.40	37.4	34.4	35.1	
16	MMBench	dev	Acc	4377 (dev)\	64.30	64.80	67.70	68.73				
17	MMBench-Ch	dev	Acc	4329 (dev)	58.30	57.62	63.60	62.54				
18	MME-Cognition	test	total score	2,374	-	348.21	-	295.35	323.92	322.5	316.78	397.14
19	MME-Perception	test	total score	2,374	1510.70	1510.75	-	1522.59	1500.85	1519.29	1575.07	1633.24
20	MMMU	val	Acc	900	-	35.30	36.40	34.80	33.4	35.1	35.9	46.7
21	MMVet	test	GPT-Eval-Avg	218	30.50	30.55	-	35.25	47.75	44.08	49.12	
22	MultidocVQA	val	Anls/acc	5,187		16.65/7.21		18.25/8.02	41.4/27.89	44.42/31.32	46.28/32.56	50.16/34.93
23	NoCaps	nocaps_eval	CIDEr	4,500	-	105.54	-	109.28	96.14	88.29	88.27	91.94
24	OKVQA	val	Acc	5,046	-	53.44	-	58.22	54.77	44.25	46.27	46.84
25	POPE	test	F1 Score	9,000	85.90	85.87	-	85.92	86.79	86.4	86.26	87.77
26	ScienceQA	scienceqa-full	Acc.	4,114	-	70.41	-	74.96	0.23	73.21	75.85	85.81
27	ScienceQA	scienceqa-img	Acc	2,017	66.80	70.43	71.60	72.88	0	70.15	73.57	
28	SEED-Bench	Seed-1	Image-Acc	17,990	total: 58.6	total: 60.49	image: 66.92	image: 67.06	65.97	64.74	65.64	69.55

We aligned the LLaVA series models results on LMMs-Eval with original reported values.

During our review to public PRs, we require the authors to report an aligned results and then accept the PRs.

Post-evaluation Analysis


```
{
    "doc_id": 0,
    "target": "answer",
    "doc": {
        "question_id": 0,
        "question": "Please describe the content of the image in detail, describe attributes of objects, scenes and background are correct.",
        "image": "<PIL.PngImagePlugin.PngImageFile image mode=RGB size=576x384 at 0x7F30B5A0DE20>",
        "category": "conv"
    },
    "arguments": [
        [
            "Please describe the content of the image in detail, describe attributes of objects, scenes and background are correct.",
            0,
            "dc100_en",
            "test"
        ]
    ],
    "resps": [
        [
            "The image shows a table setting with various items arranged on it. The main focus is a white rectangular sign that reads \"F\" an
        ]
    ],
    "filtered_resps": [
        "The image shows a table setting with various items arranged on it. The main focus is a white rectangular sign that reads \"F\" and \"
    ],
    "gpt_eval_info": {
        "question_id": 0,
        "question": "Please describe the content of the image in detail, describe attributes of objects, scenes and background are correct.",
        "model_caption": "The image shows a table setting with various items arranged on it. The main focus is a white rectangular sign that r
        "explanation": "The text caption provides a mostly accurate and comprehensive description of the image. It correctly identifies the wh
        "eval_model": "gpt-4-vision-preview",
        "score": 85,
        "prompt": "Text Caption: The image shows a table setting with various items arranged on it. The main focus is a white rectangular sign
    },
    "gpt_eval_avg_score": {
        "score": 85
    }
},
```


We provide the detail logs including:

- *Model Input*
- *Model Response*

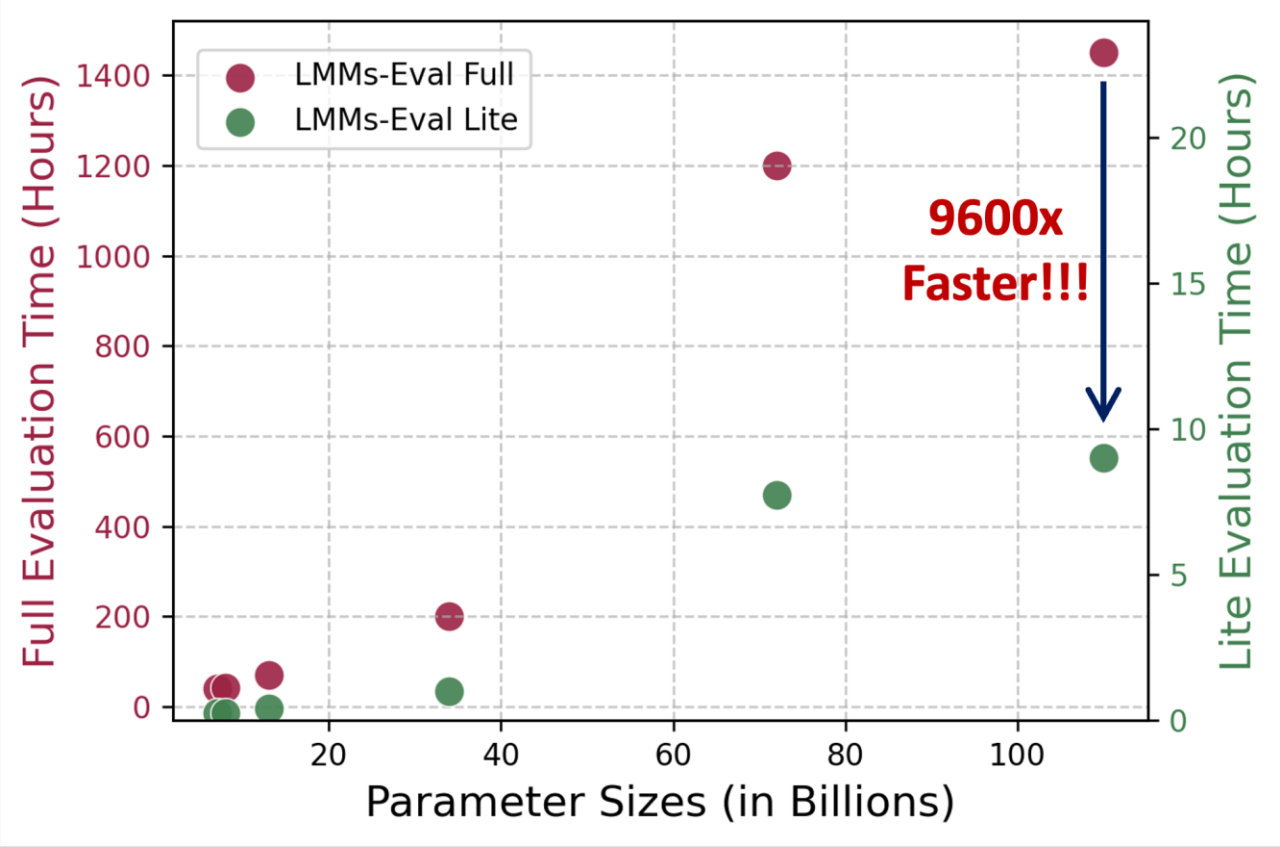
For tasks require GPT as judge, we also record

- *GPT Scores and Responses*

What’s More

- *Lite Tool*
- *Decontainmination Tool*
- *LiveBench*

Lite Tool



Development problems

- Many checkpoints
- Unable to perform large scale evaluation
- Can't decide which one is better without holistic evaluation

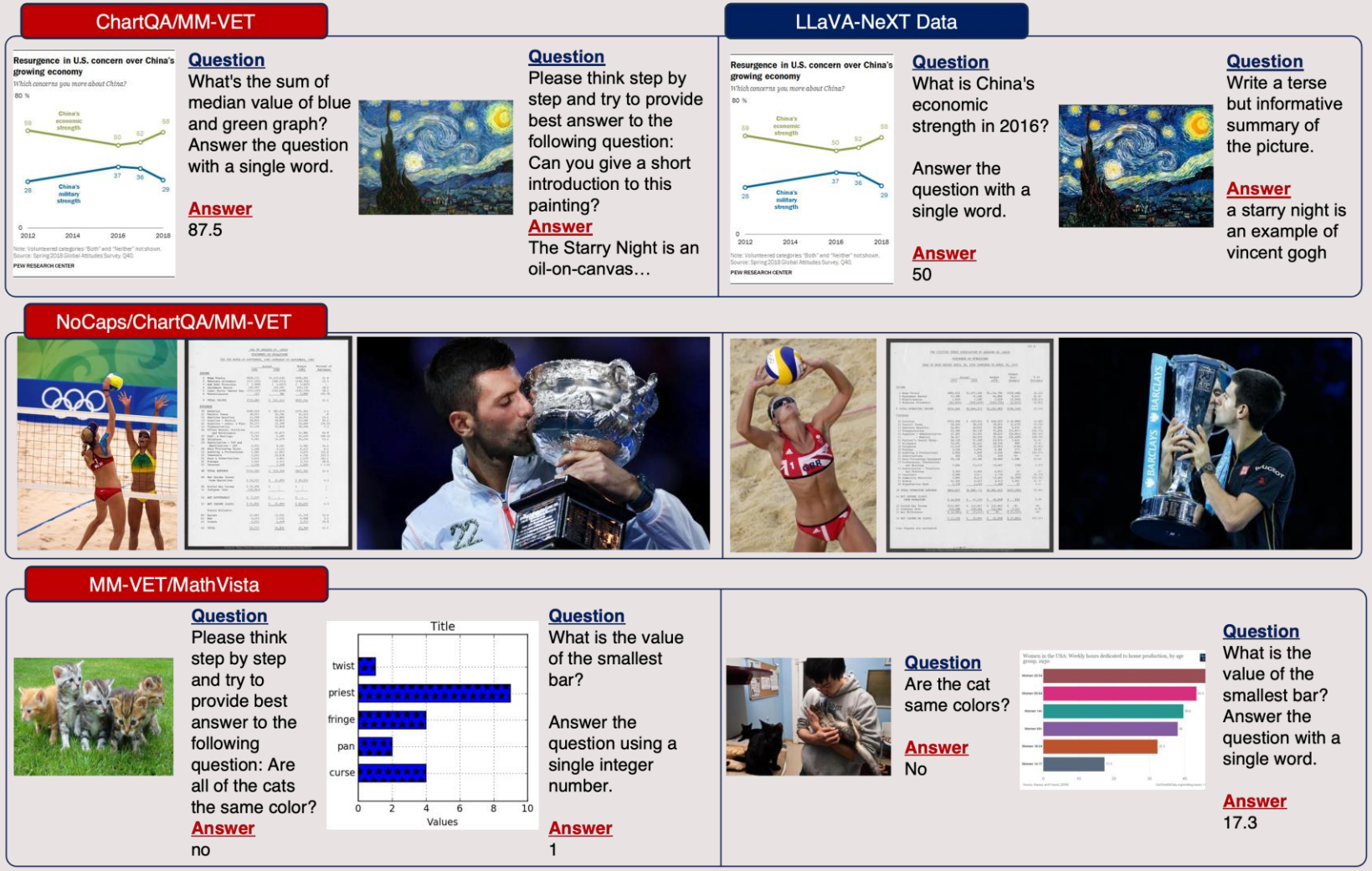
LMMs-Eval Lite

- Trade-off between evaluation accuracy and speed
- More diverse evaluation
- Lower cost

Task Domain	Dataset	Split	Full Size	Lite Size
Doc & Infographic Understanding	ChartQA	test	2500	400
	DocVQA	val	5349	400
	InfoVQA	val	2801	200
Image Understanding & Captioning	Flickr30k	val	31784	400
	NoCaps	val	4500	400
	TextCaps	val	3166	300
	RefCOCO	val	8811	500
Visual Question Answering	TextVQA	val	5000	300
Math & Science	MathVista	testmini	1000	1000
	AI2D	test	3088	300
Visual Dialogue	LLaVA-W	test	60	60
Multi-discipline	MME	cog. & percep.	2374	2374
	MMMU	val	900	900
	CMMMU	val	900	900
	Seed-Bench	test	17990	700
-	**Total**	-	**90223**	**9134**

- Current Lite dataset statistics
- Working on including more datasets!

Decontamination Analysis

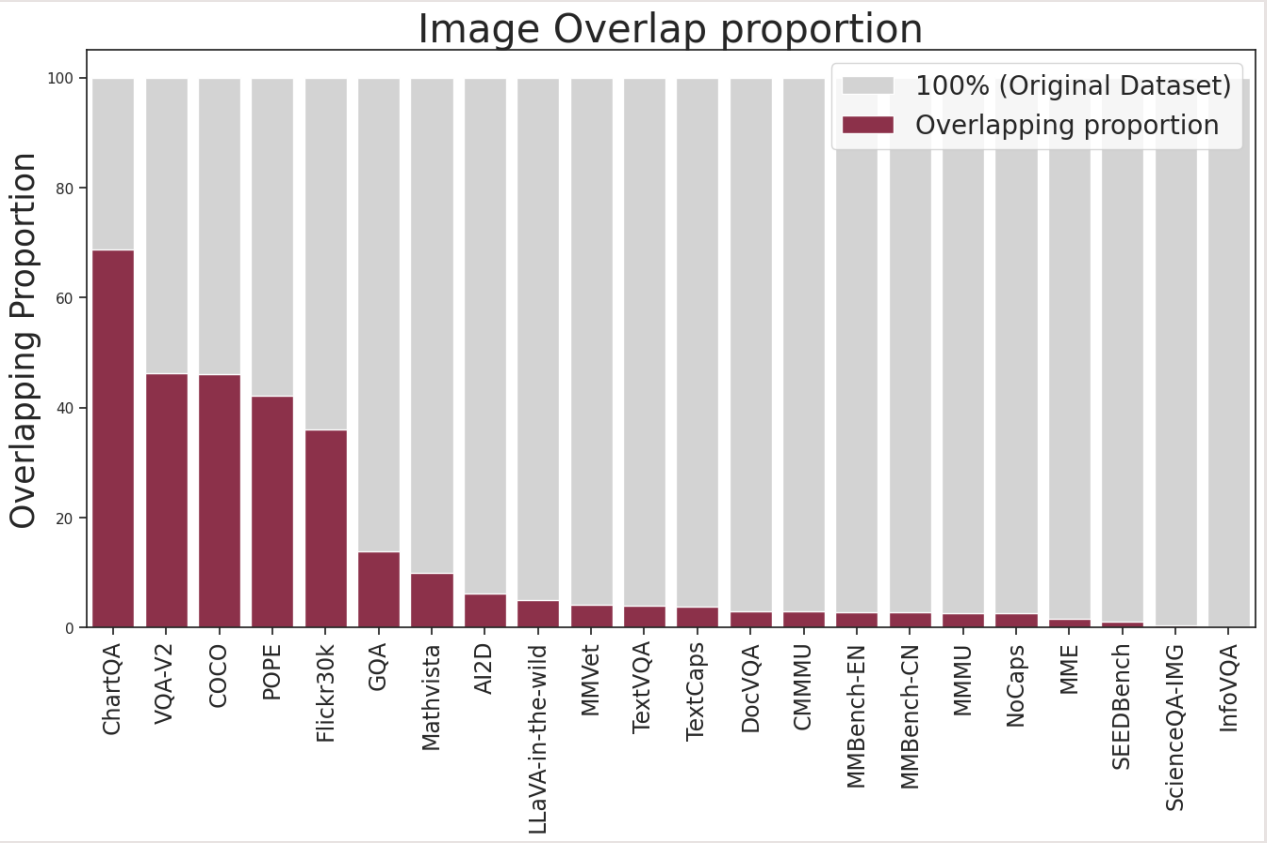


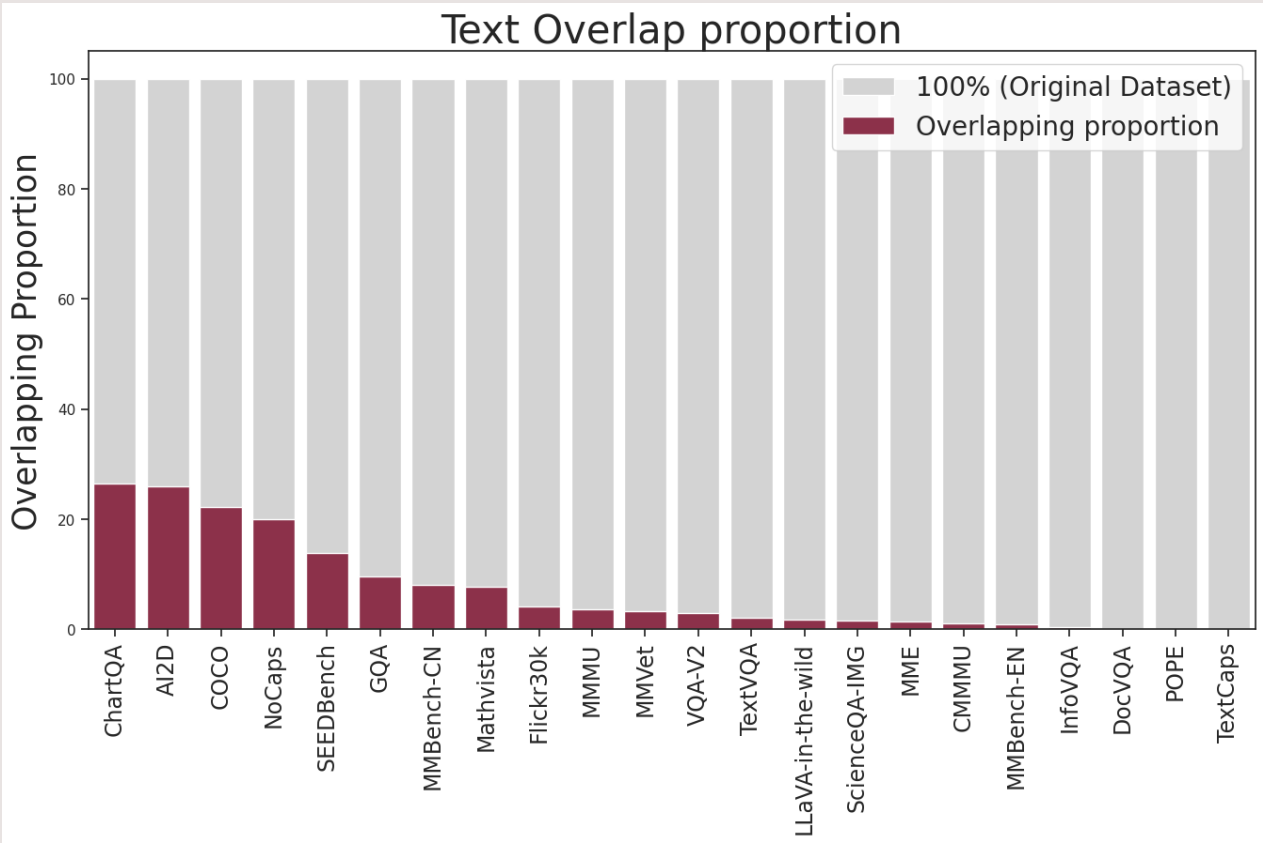
LMMs-Eval as a decontamination tool

Decontamination problem

- Duplicate Image
- Similar Image
- Similar Questions (In-domain)

Data Overlapping Statistics - (LLaVA 1.6 mix training data)



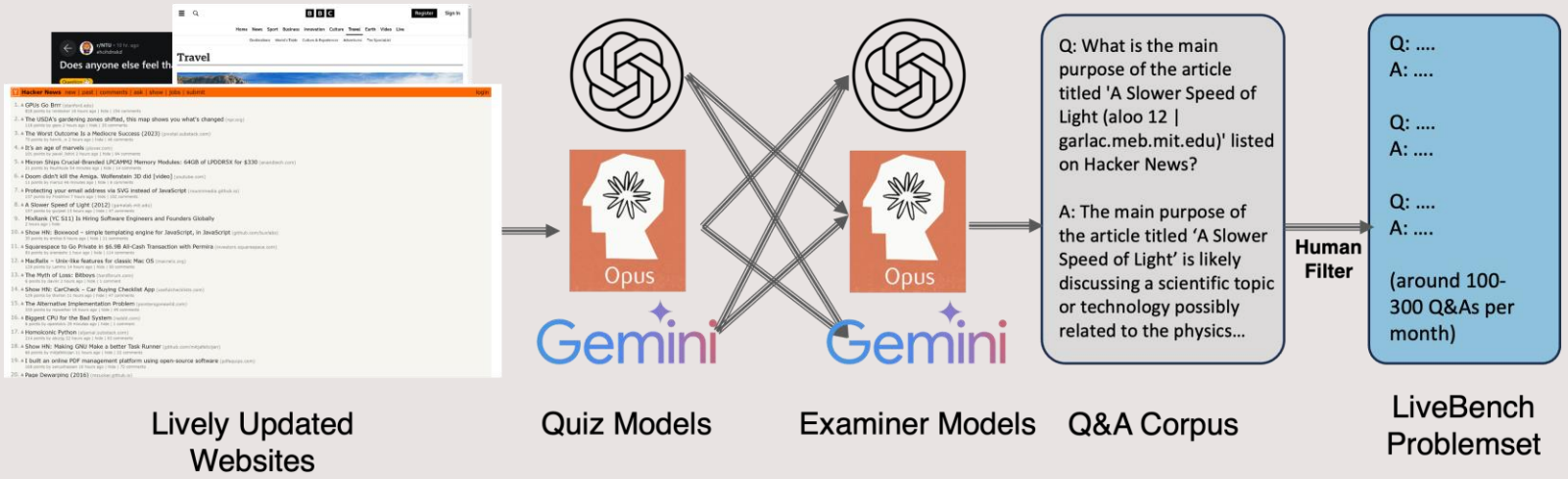


LiveBench

Multi-Modal LiveBench:

- ❑ Real-world scenarios
- ❑ Diversity
- ❑ Uncontaminated

Multi-Modal LiveBench



LiveBench Features:

- Automated updated
- Diverse Knowledge Base
- Real world use case
- Unable to contaminate

LiveBench-May Results

Models	Overall Acc. (%)	Basic	Contextual	Deeper Impl.	Broader Impl.	Insights
Idefics-2-8B	48.5	56.7	46.7	41.3	47.9	38.8
InstructBLIP-13B	53.0	26.0	72.5	69.4	78.6	49.4
LLaVA-1.5-7B	73.0	57.0	85.4	80.6	94.3	66.5
LLaVA-1.5-13B	78.9	64.2	91.5	77.5	93.8	83.0
XComposer-4KHD	79.3	80.5	82.5	75.6	82.3	72.9
Gemini-1.5-Flash	88.7	87.6	87.7	88.1	91.1	90.4
InternVL-1.5	85.1	86.4	90.6	86.5	81.0	77.5
LLaVA-NeXT-34B	86.9	81.2	89.2	93.8	91.3	85.6
LLaVA-NeXT-72B	88.1	85.5	91.0	93.8	88.6	83.8
Best Human	92.6	**97.4**	97.1	88.7	90.0	89.9
GPT-4-Vision-Preview	93.2	94.7	96.7	86.3	95.3	89.9
GPT-4(Omni)	96.3	93.3	97.6	95.6	**98.4**	**100.0**
GPT-4-Turbo	**96.9**	93.8	**99.5**	**96.9**	97.9	**100.0**

- More websites and questions will be included in June release

Community Support

During the v0.1 to v0.2, we thank the community support from pull requests (PRs). Details are in [lmms-eval/v0.2.0 release notes]

Datasets:
VCR: Visual Caption Restoration (officially from the authors, MILA)
ConBench (officially from the authors, PKU/Bytedance)
MathVerse (officially from the authors, CUHK)
MM-UPD (officially from the authors, University of Tokyo)
WebSRC (from Hunter Heiden)
ScreeSpot (from Hunter Heiden)
RealworldQA (from Fanyi Pu, NTU)
Multi-lingual LLaVA-W (from Gagan Bhatia, UBC)
II-Bench (from MAP Research Community)
LongVideoBench (from NTU, 01.AI)

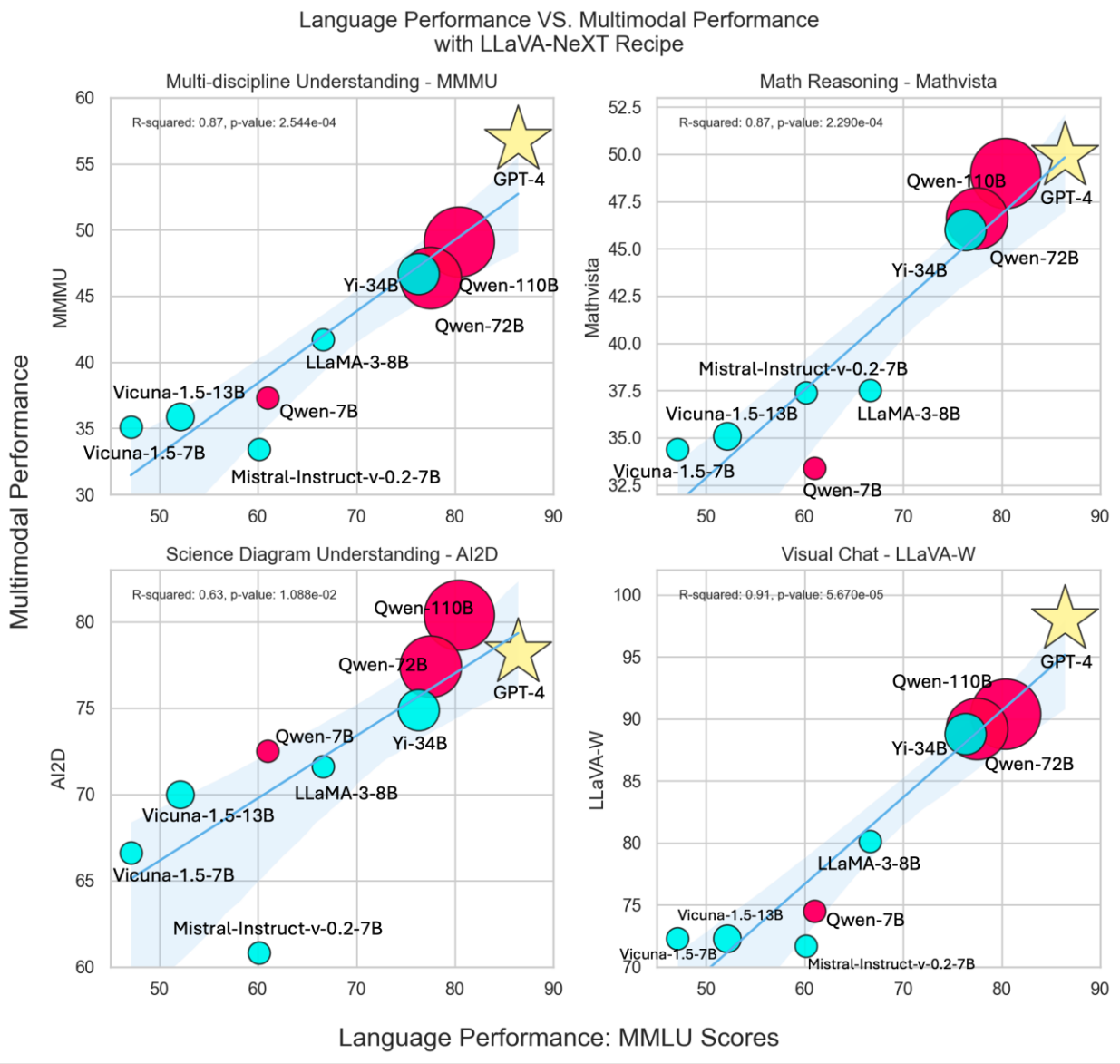
Models:
LLaVA-HF (officially from Huggingface)
Idefics-2 (from the lmms-lab team)
microsoft/Phi-3-Vision (officially from the authors, Microsoft)
LLaVA-SGlang (from the lmms-lab team)

More on LMMs-Lab

LLaVA-NeXT Project

LLaVA-NeXT Video Project

LLaVA-NeXT



Stronger LLMs Supercharge Multimodal Capabilities in the Wild

- We expand current LLaVA-NeXT to 3x model size, up to 72B and 110B model.
- We found that, under the same training strategy, stronger LLM naturally brings stronger multimodal performance.

LLaVA-NeXT

Benchmark Results

Results with LMMs-Eval				GPT4-V	LLaVA-NeXT (2024-05 Release)		
Datasets	Split	Metric	Instances		Qwen1.5-110B	Qwen1.5-72B	LLaMA3-8B
AI2D*	test	Acc.	3088	78.2	**80.4**	77.4	71.6
ChartQA*	test	RelaxedAcc.	2500	78.5	**79.7**	77.0	69.5
DocVQA*	val	ANLS	5349	-	**85.7**	84.4	78.2
MathVista	test	Acc.	1000	**49.9**	49.0	46.6	37.5
MMBench	dev	Acc.	4377	75.0	**80.5**	**80.5**	72.1
MME-Cognition	test	Total Score	2374	**517.1**	453.9	459.6	367.8
MME-Perception	test			1409.4	**1746.5**	1699.3	1603.7
MMMU	val	Acc.	900	**56.8**	50.1	49.9	41.7
RealWorldQA	test	Acc.	765	61.4	63.1	**65.4**	60.0
LLaVA-W**	test	GPT4-Eval	60	**98.0**	90.4	89.2	80.1
LLaVA-Bench (Wilder)	Small	GPT4V-Eval	120	**71.5**	70.5	71.2	62.5
	Medium	GPT4V-Eval	1020	**78.5**	72.5	73.4	63.1

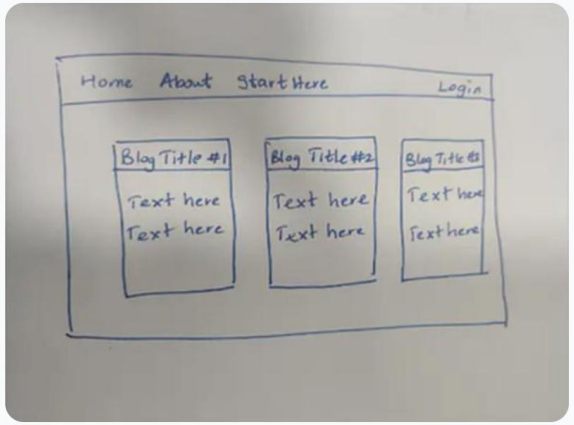
Train split observed during SFT stage.

***We report the evaluation results with GPT-4-0613 on LLaVA-W.*

- LLaVA-NeXT reaches to GPT4-V level multimodal performance.
- Our largest model, LLaVA-NeXT-110B model only trained on 128-H800 for 18 hours.

LLaVA-NeXT

Chatbot



Here's a design for blogging website. Provide the working source code for the website using HTML, CSS and JavaScript as required.

The image you've provided is a hand-drawn sketch of a website layout, which includes a navigation bar with links to "Home," "About," "Start Here," and "Login," as well as three blog post sections with titles and text content.

To create a website based on this design, you would typically use HTML for the structure, CSS for the styling, and JavaScript for any interactive elements. Here's a simplified example of how you might code this in HTML and CSS:

```
<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<title>Blog Website</title>
<style>
  body {
    font-family: Arial, sans-serif;
```

Enter message or upload file...

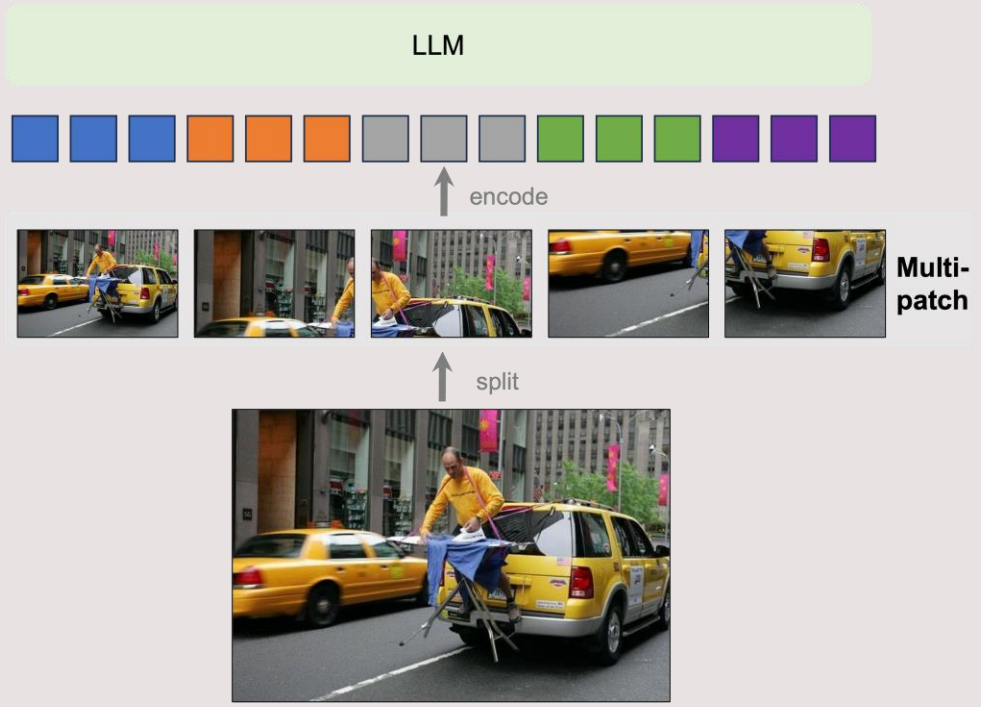
None | Up-Vote | Down-Vote | Regenerate | Send

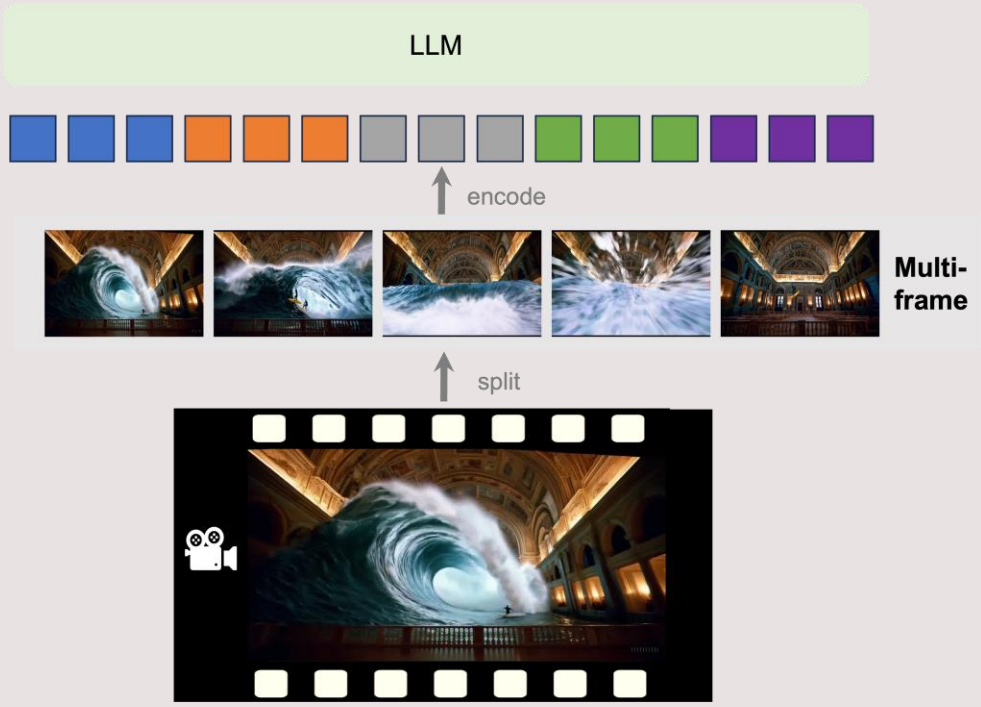
Please visit our demos to try more https://llava-next.lmms-lab.com/

Info for LLaVA-NeXT:
1. Technical report: https://llava-vl.github.io/blog
2. Model and code: https://github.com/LLaVA-VL/LLaVA-NeXT?tab=readme-ov-file

LLaVA-NeXT Video

Takeaway message 1: A unified view of images and videos is important for leveraging high-quality image data in video understanding.





LLaVA-NeXT Video digests a set of image patches/frames as a sequence of concatenated visual tokens, allowing unified view of image and video input.

LLaVA-NeXT Video

Takeaway message 2: Direct Preference Optimization (DPO) from AI Feedback Direct Preference Optimization from AI Feedback allows an LMM to reduce hallucination and follow the instructions better.

User: Please provide a detailed description of the video, focusing on the main subjects, their actions, the background scenes, and the temporal transitions.

LLaVA-NeXT-Video-DPO

The video depicts a dynamic and surreal scene where the action of surfing is juxtaposed with an indoor setting. In the foreground, there are two surfers riding waves, one slightly ahead of the other, maintaining their balance on surfboards. The surfer in the foreground is skillfully navigating the crest of a wave, while the one behind is positioned on the face of the wave, both displaying athletic postures typical of surfing. The waves are large and crashing, with the wave in the background creating a dramatic arc that suggests motion and power.

The background is an opulent interior space with high vaulted ceilings, ornate detailing, and classical architecture, including arches and what appears to be marble or stone work. The lighting is warm, with a golden hue that contrasts with the cool tones of the water. The architecture suggests a grand hall or a space designed for public gatherings or performances, possibly inspired by classical or neoclassical styles.

The temporal transition in the video is interesting as it captures a moment of action (the surfers riding the wave) within a stationary, non-moving environment (the interior space). This creates a surreal effect, blending the dynamic motion of the ocean with the static elegance of a building interior. The image is a blend of two disparate subjects—the exhilarating sport of surfing and the serene grandeur of architectural design—into a single, captivating composition.

LLaVA-NeXT-Video

The video depicts a large wave crashing inside a building, with a surfer riding the wave. The surfer is positioned in the center of the frame, skillfully navigating the wave's crest. The wave is massive, filling the majority of the space from the top to the bottom of the frame, with its white foam and water spray creating a dynamic and powerful visual effect. The background of the scene is a grand hall with high ceilings and ornate architecture, featuring arches and columns that add a sense of grandeur to the setting. The lighting is dim, with the focus on the surfer and the wave, casting shadows and highlighting the contours of the wave's movement. The camera captures the surfer from a low angle, emphasizing the height of the wave and the surfer's skill in maintaining balance. As the wave begins to break, the surfer is seen riding it towards the right side of the frame, creating a sense of motion and direction. The video transitions from a close-up of the surfer to a wider shot, showing the full scale of the wave and the grand hall. The sound of the crashing wave is prominent, adding to the immersive experience. The video concludes with a shot of the surfer disappearing from view as the wave subsides, leaving the grand hall empty and still.

We illustrate two examples to demonstrate the superiority of DPO. Texts of interest are highlighted in blue, while parts that might contain hallucinations are marked in red

LLaVA-NeXT Video

Takeaway message 3: Our LLaVA-NeXT-Video 34B model achieves SoTA performance on the recently proposed, most comprehensive diagnosis benchmark: Video-MME.

Models	LLM Params	Short (%)		Medium (%)		Long (%)		Overall (%)	
		w/o subs	w/ subs	w/o subs	w/ subs	w/o subs	w/ subs	w/o subs	w/ subs
Open & Closed-source Image MLLMs									
Qwen-VL-Chat [5]	7B	46.4	47.1	38.1	39.8	38.0	38.3	40.9	41.7
Qwen-VL-Max [5]	-	56.5	58.3	**49.9**	49.8	**49.0**	46.9	**51.8**	51.7
InternVL-Chat-V1.5 [9]	20B	**61.2**	**62.4**	47.3	**50.0**	46.0	**47.0**	51.5	**53.2**
Open-source Video MLLMs									
Video-LLaVA [28]	7B	45.9	47.1	38.1	40.2	37.3	39.6	40.4	42.3
VideoChat2 [24]	7B	38.2	41.6	33.2	34.3	29.7	31.9	33.7	35.9
ST-LLM [33]	7B	47.0	49.9	36.9	42.2	31.8	37.3	38.6	43.2
Chat-UniVi-V1.5 [19]	7B	46.3	51.4	40.3	45.2	36.9	42.3	41.2	46.3
LLaVA-NeXT-Video [68]	34B	**63.1**	**66.4**	**51.1**	**53.2**	**44.6**	**48.7**	**52.5**	**56.0**
Closed-source MLLMs									
GPT-4V [45]	-	71.4	74.5	56.5	59.3	54.2	57.2	60.7	63.7
GPT-4o [46]	-	77.1	77.5	62.1	63.0	59.2	56.7	66.2	65.8
Gemini 1.5 Pro [51]	-	**82.3**	**84.7**	**75.3**	**82.6**	**67.5**	**76.3**	**75.7**	**81.6**

Info for LLaVA-NeXT Video:
1. Technical report: https://llava-vl.github.io/blog/2024-04-30-llava-next-video/
2. Model and code: https://github.com/LLaVA-VL/LLaVA-NeXT?tab=readme-ov-file

LMMs-Lab Team and Information

Student Team

- Bo Li, PhD Student, NTU, Singapore
- Yuanhan Zhang, PhD Student, NTU, Singapore
- Peiyuan Zhang, Research Assistant, NTU -> PhD Student, UCSD
- Kaichen Zhang, Undergraduate Student, NTU, Singapore
- Fanyi Pu, Undergraduate Student, NTU, Singapore
- Kairui Hu, Undergraduate Student, NTU, Singapore
- Shuai Liu, MS Student, NTU, Singapore
- Jingkang Yang, PhD Student, NTU, Singapore

Advisor/Faulty Team

- Chunyuan Li, ByteDance
- Ziwei Liu, Assistant Professor, NTU, Singapore

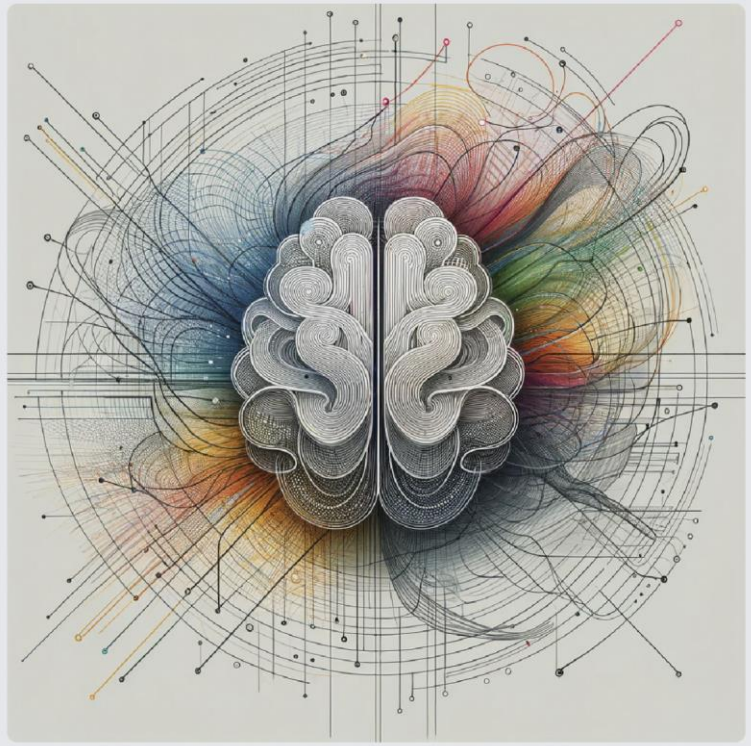
Github

Huggingface

LMMs-Lab Website

LMMs-Lab

Feeling and building multimodal intelligence



GitHub

HuggingFace

LMMs-Lab Website