

FLUFFY - The Comprehensive Benchmark for Multimodal Large Language Models

Clara Wicherz
Potsdam University
clara.wicherz@gmail.com



Abstract

This paper presents a comprehensive study on the capabilities and performance of multimodal large language models (MLLMs). We introduce a benchmark encompassing various task categories and assess the performance of seven leading MLLMs, including LLaVA, BLIP-2, InstructBlip, OpenFlamingo, IDEFICS, Otter, and Fuyu, using both quantitative and qualitative methods.

InstructBlip ranked first on our benchmark, closely followed by LLaVA. Our findings reveal that while MLLMs show promise, they also exhibit limitations, such as instruction following, object hallucination and poor performance on optical character recognition (OCR). Quantitative and qualitative findings suggest the following promising strategies: chain of thought generation, lightweight transformers with cross-attention as modality bridging modules, LLaMA-2 Chat as text encoder and instruction-tuning with larger datasets.

1 Introduction

The world presents itself multimodal and over 80 % of human perception, learning and cognition occurs via multimodal communication. MLLMs are being developed to empower LLMs to navigate real world problems and become more user-friendly.

This however also poses the challenge of tasks becoming much more diverse and MLLMs are required to show a larger ability to generalize [Li et al., 2023c]. As of November 2023, MLLMs are still an only recently emerged type of model with limited capabilities, but fast development. Our contribution to the field of MLLMs is to create a benchmark with a variety of task categories, assess the performance of seven Vision and Language MLLMs quantitatively and conduct a qualitative analysis, obtain complete information on the architectural building blocks and training data and derive hypotheses about the driving factors in MLLM

performance. We hope to spark further research with the following insights into LLaVA, BLIP-2, InstructBlip, OpenFlamingo, IDEFICS, Otter and Fuyu.

2 Related Work

To put our benchmarking in perspective we will first give a gentle introduction to MLLMs. Subsequently, we present other benchmarks, their insights and their blind.

2.1 MLLMs

To implement multimodality in LLMs different strategies have been developed regarding their architecture and training data that we will present in the following sections.

2.1.1 Architecture

There are two approaches towards integrating multimodality: the end-to-end tuning approach and the LLM-tool-using systems approach. The first and more common approach takes foundation models that each take care of one modality, namely a pre-trained vision and a pre-trained language encoder. An interface such as a linear layer or transformer is trained to bridge these models [Yu et al., 2023, Yin et al., 2023]. For the LLM-tool-using systems, vision models process images and output natural language which can be fed to an LLM without further interface [Yu et al., 2023]. This allows for better modularization, but depending on the LLM can lead to high API query costs [Li et al., 2023a]. Another disadvantage is the information loss in the process of translating visual information into natural language [Yin et al., 2023], especially when being unaware of the information of interest, delivered in the textual prompt. Examples for models of this category are ReAct and Transformer Agent [Yin et al., 2023]. In our benchmark we will only further look at end-to-end tuned models.

2.1.2 Learning Strategies

In analogy to unimodal LLM, multimodal LLMs can be pre-trained, instruction-tuned and can learn via in-context samples.

Pre-Training. In the LLM-tool-using systems approach all the pre-training has already taken place before integrating the natural language speaking expert models with LLMs. For the end-to-end tuning approach pre-training is necessary to bridge the foundation modules in a meaningful way. To do so, MLLMs get trained on large multimodal data. The pre-training objective is often to predict the next token given the textual and visual context. Often, but not always, the foundation models get frozen and only the bridging module gets updated. After the training the output of the vision encoder gets efficiently translated to the language model without compressing it into natural language as in the LLM-tool-using systems [Yin et al., 2023].

Instruction-Tuning. An instruction describes a task and how to solve it. An example can be “Describe the sentiment of the following tweet”. Instruction-formatted data contains triplets of instruction, further input such as the tweet in our example and a ground truth. Fine-tuning models on this type of data has improved the generalization capability of unimodal LLMs such as ChatGPT or FLAN [Yin et al., 2023, Liu et al., 2023, Li et al., 2023c]. This strategy is now also adopted by new MLLMs such as InstructBlip, the successor of the not-instruction-tuned BLIP-2.

In-Context Learning. Humans can learn via observation and so can LLMs. You can help humans and models understand a task and how to execute it by first showing it some examples. From these so-called in-context samples the model can perform learning by analogy, also called in-context learning. Including these in-context samples in the prompt is called few-shotting, excluding them is called zero-shotting [Yin et al., 2023].

Via fine-tuning, zero- or fewshotting one can stir the model to output not just an immediate answer to a question, but to produce the reasoning steps which should lead to that outcome. While these chains are still fragile, they can help a model give a better answer to tasks that require reasoning to get to the correct answer [Yin et al., 2023].

2.1.3 Evaluation Strategies

To assess the performance of MLLMs, several strategies have been developed. On datasets that require very short answers or only have predeter-

mined answer options models can be evaluated with accuracy or other dataset-adequate standard metrics. Are the questions more open, lengthy and require the model to chat with the user, narrow evaluation strategies such as accuracy cannot adequately assess the model’s output anymore. Instead, automatic scoring systems have been developed such as sending an instruction, input data and the tested MLLM’s output to ChatGPT/GPT-4. ChatGPT/GPT-4 is then prompted to produce its own answer to the task and compare the tested MLLM’s solution against its own. The result of this comparison is then treated as a measurement of the quality of the tested MLLM [Yin et al., 2023].

2.1.4 Challenges

MLLMs are still in a rudimentary stage with limited capabilities. Some of the challenges are limited visual perception which leads to wrong visual information acquisition, hence information loss. Furthermore, reasoning chains appear to be fragile and models suffer from object hallucination (seeing and reading things that are not there). Furthermore, the limited instruction-following capabilities pose a challenge on precise, objective evaluation [Yin et al., 2023, Yu et al., 2023].

2.2 Other Benchmarks

On our quest to investigate how good MLLMs are we investigated other relevant benchmarks as of November 2023. We identified four relevant MLLM benchmarks, their datasets, evaluation strategies, insights and limitations. We collected general information on the benchmarks in Table 1 and will now present the benchmarks individually.

2.2.1 MM-Vet

This mostly humanly annotated small scale benchmark tests 6 vision and language capabilities and how well models can integrate them in 16 different combinations. Those capabilities are recognition, OCR, knowledge, language generation, spatial awareness and performing mathematical operations. The open questions ask for varying answer styles that range from single words to essays. Therefore, the authors chose to perform the evaluation scoring by fewshotting ChatGPT/GPT-4 that returns a percent value with 0 % having no performance and 100 % the maximum performance quality. End-to-end tuned and the LLM-tool-using systems were tested. The LLM-tool-using system MM-ReAct-GPT-4 won in all capability categories

Benchmark	Datatype	Image/Video	Data Collection Questions	Ground Truth	Size (in K)	Eval
MM-Vet	{image, direct answer question, ground truth}	various online sources, VCR, ChestX-ray14	generated by humans, VCR	annotated by human, collected from the internet, VCR	0.218	automatic evaluation with fewshotting GPT-4
M3Exam	{image, multiple choice question, correct answer}	School Exam	School Exam	School Exam	12	accuracy based on answer ranking strategy; answer modeled as first produced character indicating the alphabetic index of the answer choice
SEED	{images/videos, multiple choice question, correct answer}	Images: Sample from CC3M Videos: Something-Something-v2, Epic-Kitchens 100, Breakfast	generated by fewshot prompting ChatGPT/GPT-4	generated by fewshot prompting ChatGPT/GPT-4 and validating the correct answer via a human	19	accuracy based on answer ranking strategy; 19 answer modeled as choice the model assigned the highest likelihood to
MME	{image, yes/no question, correct answer}	Sample from COCO, CTW1500	generated by human	generated by human	2.378	accuracy

Table 1: Relevant benchmarks as of November 2023 and the type of data used, information on data collection and size as well as evaluation strategy.

except recognition, in which it was outperformed by LLaVA-13B(LLaMA-2) and LLaVA-13B (V1.3, 336px). Averaged over all single capability test results, MM-ReAct-GPT-4 won with a total of 44.6 %, followed by LLaVA-13B(LLaMA-2) and LLaVA-13B (V1.3, 336px) with 32.9 % and 32.5 %, which leads the authors to hypothesize that the more powerful the LLM, the more powerful the MLLM would be [Yu et al., 2023].

2.2.2 M3Exam

To test the “intelligence” of MLLMs Zhang et al. [2023] tested BLIP-2, InstructBlip, OpenFlamingo and Fromage as well as the language only model Flan-T5 on school exams. Benchmark instances are multiple choice questions from elementary, middle and high school, from various subjects and from 9 languages. From the total of 12K samples, 2,738 are in English. The models were prompted to only give short answers and were explicitly asked not to give further explanations or details. To determine which choice the model predicted, Zhang et al. [2023] interpreted the first produced character as the alphabetic letter representing the index of the answer choice and compared it to the ground truth to obtain the accuracy. BLIP-2 outperformed its competitors with an accuracy of 50.18 % on the single choice question, closely followed by the LLM Flan-T5. OpenFlamingo and Fromage could only score 30.18 % and 21.45 % which is around the same accuracy a random prediction would score. On the multiple choice tasks Flan-T5 outperformed all models with 40.34 %, the other MLLMs performed with a score of 36.97 % or worse. Despite not being able to take in-context samples, BLIP-2 was the overall winner. Furthermore, Zhang et al. [2023] prompted the models to describe the objects in the images with a high level of detail and observed that all models struggled to recognize visual details relevant to the task. In general Zhang

et al. [2023] observed the poor instruction following capabilities across models and concluded that the model failures should be further investigated at primary school level, since the bottleneck seems to be basic reasoning capabilities rather than specific knowledge.

2.2.3 SEED-Bench

Li et al. [2023b] observed other benchmarks shortcoming of evaluation subjectivity due to human assessment or noisy evaluation due to an assessment with ChatGPT/GPT-4. With the objective of creating an objective, accurate benchmark Li et al. [2023b] proposed GPT-generated multiple choice questions with human annotations as ground truths. To not be affected by a lack of instruction-following capability Li et al. [2023b] computed the log likelihood for each answer choice and took the answer choice with the highest log likelihood as prediction. To assess the models comprehensively, ChatGPT/GPT-4 was prompted to generate questions in 12 categories, 9 for images, 3 for videos, based on images from CC3M and videos from Something-Something-v2, Epic-kitchen 100 and Breakfast. InstructBLIP Vicuna reached the highest score on the resulting benchmark with an average accuracy of 53.37 %. InstructBLIP scored the second place with an accuracy of 42.73 %, BLIP-2 third place with 46.35 %. Otter, Openflamingo and LLaVA reach relatively mediocre ranks with 33.91 %, 33.14 % and 33.52 %. Li et al. [2023b] observed a low understanding of spatial relationships. In tasks that required this type of understanding, InstructBlip performed best, however only with an accuracy of 40 %. Furthermore, Li et al. [2023b] reports that all models have poor text recognition skills other than InstructBLIP.

2.2.4 MME

Fu et al. [2023] assess MLLMs’ perception and cognition capabilities via 14 tasks. To avoid data leakage instruction-answer pairs were manually annotated. In the development phase of the benchmark, prompts were intended to stir the model towards concise answers. Fu et al. [2023] reported that the task comprehension and instruction-following on multiple choice tasks was so low that this type of question was removed from the evaluation suite. The models were zeroshotted. Beyond low instruction following, Fu et al. [2023] observed a lack of perception that surfaced when misidentifying numbers or characters - a finding consistent with Yu et al. [2023]’s. Furthermore, the models hallucinated and the logic chains were broken. Since no aggregate evaluation metric was provided we will not further elaborate on the quantitative performance of the tested models and directly refer to their paper.

The listed benchmarks all show a unique approach towards assessing MLLM’s abilities. If you consider the fact that M3Exam is multilingual and has only around 2,738 samples in English, the only benchmark of notable size is SEED. It covers a wide range of tasks, however all synthetically generated with the same method and all of the same task type. We see the need for a benchmark with a larger datasize, diversly generated tasks, varying question and answer styles, a quantified report on the instruction following capability as well as an objective evaluation.

3 FLUFFY-Bench

With a wide range of datasets from four different task categories, a notable size of 30K instances and an evaluation strategy that maintains objectiveness while taking into account the limited instruction following capabilities, we present the benchmark FLUFFY-Bench.

We chose annotated datasets from the task categories classification, visual question answering, reasoning and knowledgeable question answering. The specific datasets, their data structure and the size of the split that was publicly available to us for testing, is reported in Table 2. Examples for each dataset can be found in Section 5.

The annotations of the dataset allow for an objective evaluation. The tested models were zeroshoted and evaluated with the evaluation method suggested by the dataset author’s or according to the

metrics responsible for the ranking on the official leaderboards. All datasets suggest accuracy, with A-OKVQA and OK-VQA adding some unique pre-processing and weighing of correct answers and except MAMI, which only takes the F1-score into account for performance rankings. Before evaluating the answers, some preprocessing was done, such as removing certain special characters or repeated period signs (“...”), a characteristic of Open-Flamingo’s output. Furthermore, we applied post-processing tolerance. For direct answer questions the model output was treated as correct, if some part of the output matched the proposed answer options in the annotations. For multiple choice questions or classification tasks, the output was viewed as correct, if the correct label occurred in the model output, while the alternative valid answers were absent.

3.1 Knowledgable Visual Question Answering

In this task category models need to both process the visual/textual input and access external knowledge.

OK-VQA asks multiple choice questions that require common sense knowledge and cannot be answered with common answers [Allen Institute for AI, 2023b].

The crowd-sourced A-OKVQA is the successor of OK-VQA and ensures that questions can only be answered with broad commonsense and world knowledge as opposed to simple lookups. Furthermore, it offers a direct answer question and a multiple choice question to each image and provides the respective labels. For direct answer questions a list of 10 possible labels is provided, which contains the annotations of 10 human annotators. A model’s output is regarded as accurate if the output matches one annotation. The weight of the model’s correct answer in the total accuracy metric gets higher, the more annotations it matches [Allen Institute for AI, 2023a].

ScienceQA’s multiple choice questions are taken from annotated lectures of a wide range of scientific domains across natural science, language science, and social science. On this dataset a model’s scientific knowledge and reasoning capability gets tested [Center for Vision, Cognition, Learning, and Autonomy at UCLA, 2023].

3.2 Classification

Four classification tasks are included in FLUFFY-Bench: Sexism and Hate classification as well as

Task Type	Dataset	Data Structure*	Size Test Split**
Classification	<u>E-SNLI-VE</u>	{Image, Hypothesis, Entailment Label}	14740
	<u>MAMI</u>	{Meme Image, Meme Text, Sexism Label}	1000
	<u>Hateful Memes</u>	{Meme Image, Meme Text, Hate Label}	500
	<u>MVSA</u>	{Tweet Image, Tweet Text, Sentiment Label}	3948
Visual Question Answering & Visual Reasoning	<u>CLEVR</u>	{Image, Direct Answer Question, Ground Truth}	1000
	<u>GQA</u>	{Image, Direct Answer Question, Ground Truth}	1000
Knowledgeable Visual Question Answering	<u>OK-VQA</u>	{Image, Direct Answer Question, Ground Truth to Direct Answer Question}	5046
	<u>A-OKVQA</u>	{Image, Direct Answer Question, Ground Truth to Direct Answer Question, Multiple Choice Question, Choices, Correct Answer Choice}	1145
	<u>ScienceQA</u>	{Image, Multiple Choice Question, Correct Answer Choice}	2017

Table 2: Comparison of results across different metrics.

sentiment analysis and hypothesis entailment prediction, the latter also falling under the task category of reasoning.

MAMI (Multimedia Automatic Misogyny Identification) was created for SemEval2022-Task-5 and contains sexist and non-sexist memes and respective human misogyny annotations [MIND Lab].

MVSA (Multi-View Sentiment Analysis) contains original image-text pairs from X, former Twitter, with human sentiment annotations [Laboratory].

The dataset Hateful Memes is provided by Meta AI and contains examples for hateful memes on social media. To ensure licensing clarity while preserving semantic content, the original memes were replaced by similar images from Getty Images [Meta AI, 2020].

In e-SNLI-Ve a sample consists of an image, a textual hypotheses about the image and an annotation, whether the image and the hypothesis are consistent (entailment), contradictory (contradiction) or neither (neutral). For the dataset explanations from e-SNLI and image-sentence pairs from SNLI-VE were merged [esn]. For FLUFFY-Bench only regard the named entailment labels.

3.3 Visual Question Answering Reasoning

GQA and CLEVR belong to both the categories visual question answering and reasoning. GQA is a large scale dataset with real images from the

Visual Genome. For FLUFFY-Bench we only regard 1000 samples of the dataset. Each instance contains an image, a direct question and an answer. The questions focus on various aspects of scene understanding such as colour, spatial relation, activities or logical inference [gqa].

CLEVR is a synthetic dataset with 3D-rendered objects, direct questions about object attributes, counting and comparisons and answers [cle]. We again sampled 1000 instances for FLUFFY-Bench.

4 Evaluation

We benchmarked seven of the most promising open-source MLLMs. Before presenting the benchmark results we will briefly outline the model architecture and training process.

4.1 Models evaluated

With the exception of Fuyu all models have a classic structure of a text encoder, a vision encoder and a bridging module to align the modalities. However, each model adds their own spin on it by the selection of the respective encoders, the bridging process (Table 3), the training data and the decision which elements in its system to train at which learning stage (Table 4).

4.1.1 BLIP-2

This system combines Google’s text encoder FLAN-T5 XXL with the ViT-G/14 from EVA-

Model	Text Encoder		Image Encoder		Bridging Module	
	Name	Parameter	Name	Parameter	Model	Parameter
InstructBLIP	FLAN-T5 XXL	11B	ViT-G/14 (EVA-CLIP)	1.8B	Transformer with Cross-Attention	188M
LLaVA	LLaMA-2 Chat	13B	ViT-L/14 (CLIP)	303M	Linear Layer	NaN
BLIP-2	FLAN-T5 XXL	11B	ViT-G/14 (EVA-CLIP)	1.8B	Transformer without Cross-Attention	188M
Fuyu	Persimmon-8B	8B	not applicable	not applicable	not applicable	not applicable
Otter	LLaMA v1	7B	CLIP ViT-L/14	303M	Transformer with Cross-Attention	1.3B
OpenFlamingo	MPT	7B	CLIP ViT-L/14	303M	Transformer with Cross-Attention	1.3B
IDEFICS-instruct 9B	LLaMA v1	65B	ViT-H-14 (OpenCLIP)	632M	Transformer with Cross-Attention	NaN

Table 3: Model Architecture.

CLIP and bridges it with a lightweight transformer. During pre-training the encoders get frozen and the bridging module gets trained with COCO, Visual Genome, CC3M, CC12M, SBU and a sample of 115M image-text pairs from LAION400M which totals to 129M image-text pairs. BLIP-2 has no instruction-tuning [Li et al., 2023c].

4.1.2 InstructBlip

The successor of BLIP-2 implements instruction-tuning and adds an important tweak to the bridging module. InstructBlip adds cross-attention so the visual feature extraction can be done w.r.t. the textual instruction as opposed to naively extracting the most important features in a general context. While the pre-training data stays the same, InstructBlip adds instruction-tuning which is conducted with 26 datasets (see Table 4)[Dai et al., 2023].

4.1.3 LLaVA

LLava uses ViT-L/14 (CLIP) as image encoder and the powerful LLM LLaMA-2 Chat as text encoder. While the LLaMA-2 architecture is similar to its predecessors, it has been trained on 40 % more data. Furthermore, the Chat edition has been instruction-tuned for dialogue [Touvron et al., 2023]. To learn the coordination of both foundation models, both encoders get frozen and a linear layer gets added in between to map the visual features to the word embedding space of the text encoder. Liu et al. [2023] generated the pre-training dataset by sampling 595K image from CC3M and captioned them

by prompting the LLM BLIP [Liu]. To obtain instruction-tuning data Liu et al. [2023] sampled COCO images and represented them in text via captions and bounding boxes. The image, represented as text, then got fed to ChatGPT/GPT-4 along with a prompt to generate instruction following data of the data type conversation, detailed description and complex reasoning. This resulted in 58K samples of conversations, 23K of detailed description, and 77k of complex reasoning. To instruction-tune LLaVA, only the vision encoder was frozen, while the linear layer and the text encoder got further trained. Liu et al. [2023] reports their intent to further test a more sophisticated bridging mechanism as can be found in InstructBlip.

4.1.4 Fuyu

Fuyu is the only model that does not process images via a separate image encoder, which makes the system smaller, less data expensive to train and enables fast inference. Instead it passes images in patches through a linear projection directly to the first layer of the LLM Persimmon-8B. This LLM was trained on only 0.37 % as much data as LLaMA2. ade [b] report that it is difficult for the model not to produce text, even when it explicitly gets prompted to.

The sequence of image tokens gets handled just like the sequence of text tokens. Thereby the LLM gets treated like an image transformer minus the pooling and the causal attention. The model was pre-trained on unknown data, but nei-

Model	Text Encoder		Image Encoder	Bridging Module		
	Pre-training	Instruction-tuning		Pre-training	Pre-training	Instruction-tuning
InstructBLIP	Dataset names unavailable (total 780B tokens)	Muffin, T0-SF, NIV2, and CoT (total of 1.4B tokens)	JFT-3B (total of 3B annotated images)	COCO, Visual Genome, CC3M, CC12M, SBUs, Sample (115M image-text pairs) from LAION400M (total of 129M image-text pairs)	COCO Caption, Web CapFilt, NoCaps, Flickr30K, TextCaps, VQA2, VizWiz, VQA2, OKVQA, A-OKVQA, GQA, Visual Spatial Reasoning, IconQA, Visual Dialog, ScienceQA, Visual Conversation, Complex Reasoning, Detailed Image Description, MSVD QA, MSRVTT QA, IVQA, OCR-VQA, TextVQA, HatefulMemes (total of image-instruction-answer triplets unavailable)	
LLaVA	Dataset names unavailable (total of 2T tokens)	public datasets (dataset names unavailable), proprietary high-quality SFT data, 1M binary human preference comparisons (helpfulness & safety annotations) (total of tokens unavailable)	YFCC100M, further datasets, dataset names unavailable (total of annotated images unavailable)	Sample from CC3M, images captioned by BLIP (total of 595K image-text pairs)	LLaVA-Instruct-80K (total of 80K image-instruction-answer triplets)	
BLIP-2	Dataset names unavailable (total of 780B tokens)	Muffin, T0-SF, NIV2, and CoT (total of 1.4B tokens)	JFT-3B (total of 3B annotated images)	COCO, Visual Genome, CC3M, CC12M, SBUs, Sample (115M image-text pairs) from LAION400M (total of 129M image-text pairs)	Not applicable	
Fuyu	dataset names unavailable (pre-training and instruction-tuning total of 737B tokens)	dataset names unavailable (pre-training and instruction-tuning total of 737B tokens)	Not applicable	Not applicable	Not applicable	
Otter	C4, Github, Wikipedia, Books, ArXiv, StackExchange (total of 1T tokens)	Not applicable	YFCC100M, further datasets, dataset names unavailable (total of annotated images unavailable)	LAION-2B, Multimodal C4 (180M image-text pairs)	MIMIC-IT (2.8M image-instruction-answer triplets)	
OpenFlamingo	Dataset names unavailable (total of 1M tokens)	Not applicable	YFCC100M, further datasets, dataset names unavailable (total of annotated images unavailable)	LAION-2B, Multimodal C4 (180M image-text pairs)	Not applicable	
IDEFICS-Instruct 9B	C4, Github, Wikipedia, Books, ArXiv, StackExchange (total of 1T tokens)	Not applicable	Subset of LAION-2B (total of 34B image text pairs)	OBELICS, Wikipedia, LAION, PMD (1.5B)	M3IT, LRV-Instruction, LLaVA-Instruct, LLaVAR-Instruct, SVIT, General Scene Difference, Spot-theDiff, UltraChat, Sample of OBELICS, Wikipedia, LAION, PMD (total of image-instruction-answer triplets unavailable)	

Table 4: Training Data of MLLM components.

ther instruction-tuned, nor were post-processing or sampling strategies added to control for undesirable output. The authors recommend further fine-tuning and few-shotting the MLLM [ade, a, AI]

4.1.5 OpenFlamingo

The OpenFlamingo series is the open-source replication of DeepMind’s Flamingo models, that matches between 80 - 89 % of its corresponding Flamingo performance. The system uses MPT as text encoder and CLIP ViT-L/14 as image encoder. Modality alignment is implemented by extracting patch features from a frozen vision encoder and passing them through a perceiver resampler. Additionally the layers of the otherwise frozen LLM get augmented with cross-attention. The perceiver resampler and the cross-attention augmented layers get updated when pre-training on LAION-2B and Multimodal C4. While there is no instruction-tuning following, the model can still improve its performance with fewshotting. Reportedly the model’s performance can monotonously be increased for up to 32 in-context samples. Awadalla et al. [2023] also reported the models achilles heels: counting, giving lengthy answers despite being prompted not to and getting drawn to central objects in images despite being prompted to pay attention to non-central objects [Awadalla et al., 2023].

4.1.6 Otter

Otter is an instruction-tuned and slightly tweaked version of OpenFlamingo. The text encoder is substituted with LLaMA v1 and the training paradigm

is modified. While in OpenFlamingo only the vision encoder was frozen, Otter freezes both encoders, therefore drastically reducing the parameters that need to be trained for modality alignment. The biggest difference to OpenFlamingo is the instruction tuning on the MIMIC-IT(Multi-Modal In-Context Instruction Tuning dataset) dataset. This dataset was constructed from crawled webpages, where related images and text were paired, from VQAv2, GQA, LLaVA and an in-progress, high-quality panoptic video scene graph dataset from the PVSG repository. Each data sample includes an instruction-image-answer triplet and its in-context examples. Li et al. [2023a] observed the model hallucinating and hypothesizes that further training with negative examples could mitigate this downfall [Li et al., 2023a].

4.1.7 IDEFICS

Another instruction-tunes and tweaked successor of OpenFlamingo is IDEFICS. Like Otter IDEFICS replaces the text encoder with LLaMA v1. As opposed to Otter, IDEFICS also switched out the pre-training data for modality alignment to a mix of OBELICS, Wikipedia, LAION, PMD. The encoders also get frozen during pre-training. However, during instruction-tuning both encoders get trained as well. Instruction-tuning is conducted with a mix of ChatGPT/GPT-4 generated datasets: M3IT, LRV-Instruction, LLaVA-Instruct, LLaVAR-Instruct, SVIT, General Scene Difference + Spot-theDiff, UltraChat as well as 5.1 % of the image-text pairs and 30.7 % of the OBELICS multimodal

web documents from pre-training [M4, Hugging Face, 2023, Face, 2023].

4.2 Results

We evaluated the models on all tasks on each dataset. We first calculated the model’s scores per task, then averaged across tasks. For each dataset, the performance of the models we tested were compared to the best fine-tuned models scored on the respective dataset (Table 6, Table 5). Subsequently, we aggregate the accuracy averages per dataset by averaging across datasets (Table 8). Furthermore, we reported on the ratio of valid answers (Table 9, Table 7).

4.2.1 Datasets with Classification Tasks

For e-SNLI-VE the fine-tuned system OFA-X-MT takes first place with an accuracy of 0.78 [vis, mul]. The best MLLM we tested was InstructBlip which matched 80.77 % of this score with an accuracy of 0.63. With some distance BLIP-2 takes second place with an accuracy of 0.54 and LLaVA third with 0.50. Notable InstructBlip and the non-instruction-tuned BLIP-2 managed to always follow the instruction, while LLaVA followed the instructions barely half the time. The lowest score reached Fuyu with an accuracy of 0.33 which is equal to random guessing on a three-fold classification task. However, it managed to always follow the instructions. The OpenFlamingo based models showed results that were slightly better than random guessing. Otter managed to follow the instructions 63 % of the time, while IDEFICS and OpenFlamingo only managed to do so 31 % and 34 % of the time.

Hateful memes were best classified by the winner of the Hateful Memes Competition, Team Alfred Lab. They implemented and combined an ensemble system consisting of VL-BERT, UNITER-ITM / VILLA-ITM, vanilla ERNIE-Vil with a rule-based approach and scored an accuracy of 0.85 [DriveDataOrg]. With some distance, LLaVA takes second place with an accuracy of 0.63. However, this result is only based on 29 % of the samples that LLaVA answered to in a valid manner. LLaVA’s performance score is closely followed by BLIP-2 and Instruct-Blip with 0.61 and 0.58. The other models score accuracies that more or less match random guessing. Otter and IDEFICS manage to follow the instruction in a way that produces valid answers 92 % and 100 % of the time, OpenFlamingo 40 %, Fuyu only 13 % of the time. For

MAMI we rank according to the F1 score, since this was the dimension with which the performance was ranked on the external MAMI leaderboard. The highest F1-score was obtained by the winners of SemEval-2022 Task 5A, Team SRCB with their system PBR [?]. LLaVA matches this performance by 98 % with an F1-score of 0.81. However, only half of the instances could be evaluated to due a lack of instruction-following. With quite a bit of distance BLIP-2 ranked third place with an F1-Score of 0.67, but a valid answer ratio of 100 %. OpenFlamingo ranked last in performance, could barely follow the instructions with a valid answer ratio of only 9

For MVSA three of our own tested systems jointly scored the first rank, before the official winner of the MVSA leaderboard, FENet-BERT. InstructBlip, LLaVA and BLIP-2 all ranked first place with an accuracy of 0.68 on the sentiment task, double the score of a random guessing approach. The BLIP-based models showed perfect instruction following, while LLaVA provided valid answers 87 % of the time. Despite always providing valid answers, IDEFICS showed a performance worse than random guessing on a three-fold classification task with an accuracy of 17 %.

4.2.2 Datasets with Knowledgable Visual Question Answering

On OKVQA PaLM-E-562B delivered the highest performance with an accuracy of 0.66 [Papers with Code, c] . LLaVA matched 88 % of that performance with an accuracy of 0.58. Third place received InstructBlip with a score of 0.52. IDEFICS performed the worst with an accuracy of only 0.20.

For ScienceQA we faced a challenge prompting the models. The dataset consists of multiple choice questions and labels that are the index of the correct answer choice. We prompt the model to answer by giving the index of the correct choice. While the first tested model BLIP-2 could adapt to the instructed answer format, the other models could not do it. Since we compared the MLLM answers to the labels, the indices of the correct answers, the performance scores on ScienceQA in our benchmark are not meaningful and cannot be compared to other results. This indicates an adaption of our prompting strategy for the successor benchmark of FLUFFY-Bench. The official leaderboard for the dataset however states that LLaVA is the peak performer with an astonishing accuracy of 0.93 [Papers with Code, d].

For A-OKVQA our tested models InstructBlip, LLava and BLIP-2 took the first three places w.r.t. the average accuracy across tasks. InstructBlip ranked first with an average accuracy of 0.65 and showed the best performance on the multiple choice task. LLava finished second with an average accuracy of 0.61. While its performance on the multiple choice task was significantly lower, the performance on direct answers were the best. Thirdly BLIP-2 scored an accuracy of 0.57. The BLIP-based models showed perfect instruction following, LLava an almost perfect instruction following with a valid answer ratio of 0.94. A-OKVQA’s leaderboard claims the system Prophet to be the top performer with an accuracy of 0.56 [MILVLG].

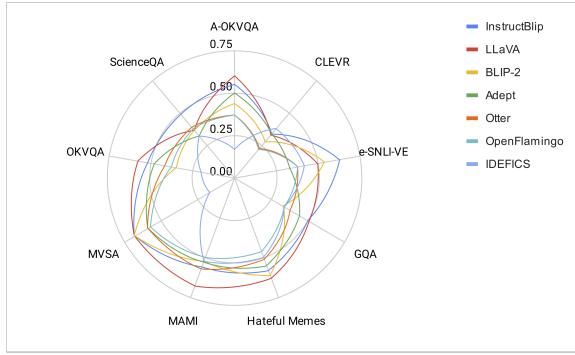


Table 5: MLLM performance per dataset.

4.3 Datasets with Visual Question Answering Reasoning

On the synthetic dataset CLEVR, NS-VQA is reported to score with an accuracy of 0.99 [Papers with Code, a], a score barely matched by the MLLMs we tested. IDEFICS astonishingly scored first from our models, with an accuracy of 0.38, followed by InstructBlip and Fuyu, both with an accuracy of 0.34. The last ranks are occupied by Otter and OpenFlamingo with an accuracy of 0.23 and 0.22. Our MLLMs showed better accuracy scores on CLEVR’s real-life-image counterpart, GQA. From our tested models LLava and InstructBlip showed the best performance with accuracies of 0.51 and 0.50. Both were outperformed by the system ProTo which scored an accuracy of 0.65 [Papers with Code, b].

4.4 Performance across Datasets

The overall best performance was delivered by InstructBlip, scoring an average accuracy of 0.55 with an average valid answer ratio of 0.9. LLava

closely follows InstructBlip’s performance with an average accuracy of 0.54 and an average valid answer ratio of 0.7. BLIP-2 occupies third place with an accuracy of 0.47 and a valid answer ratio of 0.99. The remaining MLLMs scored accuracies of 0.44 or worse with IDEFICS ranking last with an average accuracy of 0.32. The instruction following capabilities range from a valid answer ratio of 0.82 for Otter to a ratio of 0.54 for OpenFlamingo. As reference we also computed the average accuracy of the top performers across datasets. Since there is no accuracy information for the top performer on MAMI, we excluded MAMI scores for this computation. The resulting average accuracy of the best performers across all datasets except MAMI is 0.76. If we also leave MAMI out of the average across datasets, our best performer InstructBlip scores an average accuracy of 0.55 which matches 72 % of the average of all specific fine-tuned peak performers. LLava reaches an accuracy of 0.53 which matches the peak performer average with 70 % and BLIP-2 reaches an accuracy of 0.44, which matches 58 % of the performance.

5 Qualitative Analysis and Discussion

Firstly, we will discuss the quantitative results, add a qualitative analysis and discuss the findings w.r.t to the architectural design and the training data. Secondly, the design of FLUFFY will be reflected.

5.1 Models

The models greatly varied in their overall performance, answer styles and instruction following capabilities. We will look at each model’s performance and inspect output samples.

5.1.1 BLIP-2

Despite not being instruction-tuned, BLIP-2 shows one of the best instruction-following capabilities and scores in the first ranks on classification datasets and AOKVQA. However, it performed poorly on the visual question answering and reasoning tasks as well as the knowledgeable visual question answering tasks of OKVQA. An example of a wrong prediction on OKVQA is example 24, in which BLIP-2 just repeats the end of the prompt as opposed to giving the answer. Architecturally BLIP-2’s overall good performance could be due to its rather large text encoder, FLAN-T5 XXL, its image encoder ViT-G/14 (EVA-CLIP), which is the biggest one together with InstructBlip

e-SNLI-VE						
Task Metric	accuracy	Performance			Instruction Following	
		Entailment Prediction		recall	Entailment Prediction	
OFA-X-MT	0.78	N/A	N/A	N/A	N/A	N/A
InstructBlip	0.63	0.64	0.75	0.63	1.00	
BLIP-2	0.54	0.51	0.72	0.54	1.00	
LLaVA	0.50	0.46	0.65	0.50	0.44	
IDEFICS	0.42	0.32	0.38	0.42	0.31	
Otter	0.38	0.32	0.42	0.38	0.63	
OpenFlamingo	0.38	0.32	0.38	0.38	0.34	
Fuyu	0.33	0.25	0.61	0.33	0.99	

Clevr						
Task Metric	Direct Answer accuracy	Performance		Instruction Following		
		Direct Answer	N/A	Direct Answer	N/A	
NS-VQA	0.99	N/A				
IDEFICS	0.38	1.00				
InstructBlip	0.34	1.00				
Fuyu	0.34	1.00				
LLaVA	0.33	1.00				
BLIP-2	0.28	1.00				
Otter	0.23	1.00				
OpenFlamingo	0.22	1.00				

Hateful Memes						
Task Metric	accuracy	Performance			Instruction Following	
		Hate Classification		recall	Hate Classification	N/A
Winner Hateful Memes Competition*	0.85	N/A	N/A	N/A	N/A	N/A
LLaVA	0.63	0.77	0.63	1.00	0.29	
BLIP-2	0.61	0.61	0.60	0.61	1.00	
InstructBlip	0.58	0.49	0.63	0.40	1.00	
Fuyu	0.55	0.63	0.60	0.67	0.13	
Otter	0.51	0.19	0.54	0.11	0.92	
IDEFICS	0.50	0.67	0.50	1.00	1.00	
OpenFlamingo	0.46	0.49	0.47	0.50	0.40	

GQA						
Task Metric	Direct Answer accuracy	Performance		Instruction Following		
		Direct Answer	N/A	Direct Answer	N/A	
ProTo	0.65	N/A				
LLaVA	0.51	1.00				
InstructBlip	0.50	1.00				
Fuyu	0.44	1.00				
Otter	0.38	1.00				
OpenFlamingo	0.34	1.00				
IDEFICS	0.34	1.00				
BLIP-2	0.33	1.00				

Mami						
Task Metric	accuracy	Performance			Instruction Following	
		Sexism Classification		recall	Sexism Classification	N/A
Winner SemEval-2022 Task 5A**	N/A	0.83	N/A	N/A	N/A	N/A
LLaVA	0.68	0.81	0.68	1.00	0.48	
BLIP-2	0.52	0.67	0.51	0.99	1.00	
Fuyu	0.53	0.64	0.55	0.75	0.26	
InstructBlip	0.56	0.57	0.54	0.61	0.93	
Otter	0.57	0.43	0.65	0.32	0.76	
IDEFICS	0.52	0.34	0.54	0.24	0.94	
OpenFlamingo	0.50	0.20	0.57	0.12	0.09	

OKVQA						
Task Metric	Direct Answer accuracy	Performance		Instruction Following		
		Direct Answer	N/A	Direct Answer	N/A	
PaLM-E-562B	0.66	N/A				
LLaVA	0.58	1.00				
InstructBlip	0.52	1.00				
Fuyu	0.48	1.00				
Otter	0.42	1.00				
OpenFlamingo	0.38	1.00				
BLIP-2	0.35	1.00				
IDEFICS	0.20	1.00				

ScienceQA						
Task Metric	Multiple Choice accuracy	Performance		Instruction Following		
		Multiple Choice	N/A	Multiple Choice	N/A	
LLaVA*	0.93	N/A				
InstructBlip	0.49	0.10				
Otter	0.39	0.09				
OpenFlamingo	0.38	0.04				
LLaVA	0.37	0.07				
BLIP-2	0.37	0.94				
Fuyu	0.32	0.06				
IDEFICS	0.32	0.05				

A-OKVQA						
Task Metric	Direct Answer accuracy	Performance		Instruction Following		
		Multiple Choice	Average accuracy	Direct Answer	Multiple Choice	Average
InstructBlip	0.55	0.74	0.65	1.00	1.00	1.00
LLaVA	0.60	0.62	0.61	1.00	0.87	0.94
BLIP-2	0.44	0.69	0.57	1.00	1.00	1.00
Prophet	N/A	N/A	0.56	N/A	N/A	N/A
Fuyu	0.50	0.37	0.44	1.00	0.86	0.93
Otter	0.37	0.45	0.41	1.00	0.78	0.89
OpenFlamingo	0.37	0.24	0.31	1.00	0.45	0.73
IDEFICS	0.17	0.30	0.24	1.00	0.85	0.93

Table 6: Rankend MLLM performance per dataset and performance of best fine-tuned models per dataset as reference in yellow.

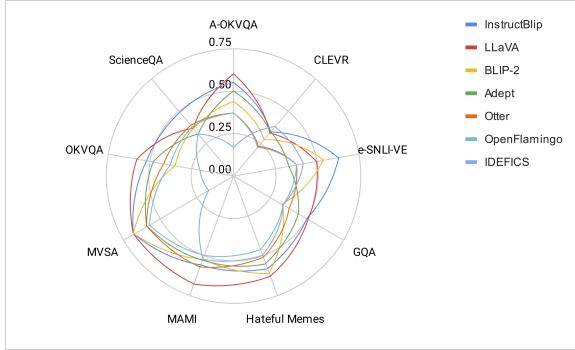


Table 7: MLLM’s valid answer ratio per dataset.

or the sophisticated, yet lightweight transformer that bridges text and image encoder.

5.1.2 InstructBlip

The instruction-tuned successor of BLIP-2 scores first or second place on each dataset with the exception of the classification datasets Hateful Memes and MAMI. Although possibly being instruction tuned on them it only scores third and fourth place. An example of a false prediction on Hateful Memes is example 28. Here InstructBlip did not combine that the mental illness written in the meme text refers to queerness, which is visually communicated through the pride colours. Overall it shows good instruction-following capabilities and scores first rank in the overall performance. The improved performance might be due to the instruction-tuning or the adjustment in the bridging module. InstructBlip makes full use of the transformer by adding cross attention with which the visual feature extraction becomes text-aware. This might mitigate problems like focussing on central or prominent visual features as opposed to features of interest w.r.t. the prompt.

5.1.3 LLaVA

LLaVA’s performance ranks second place with only one percent point less overall accuracy. The high performance could be due to the powerful LLM LLaMA-2 Chat, since multiple findings indicate the MLLM performance rises with increasing LLM capabilities. Furthermore, the instruction tuning data contains conversation, detailed description and complex reasoning, which might help the model to create chains of reasoning which are associated with higher quality answers [Yin et al., 2023]. The pre-training for dialogue and the instruction tuning data which partially consists of conversation might be responsible for the general “chattiness”

of LLaVA, since it could not be stopped from giving more elaborate and explanatory answers than prompted. With a look at the model’s output, these lengthy answers might have been beneficial, since the model’s output tokens influence the prediction of the following output token, thereby being guided to the right answer with multiple smaller steps. The chattiness however also led to instances, where the model hallucinated an entire human conversation partner, such as in example 3. Example 4 on the other hand shows an instance where the unprompted answer elaboration might have guided the model to predict the correct answer for a MAMI instance. Example 12 shows the only time an MLLM correctly predicted an A-OKVQA instance that requires OCR.

Liu et al. [2023] communicates their intent to also use a lightweight transformer such as BLIP-2’s in the future. It will be interesting to see whether a more sophisticated bridging module enhances the performance further, or whether this easy bridging module architecture also has its benefits.

5.1.4 Fuyu

Fuyu scored in the middle and lower ranks across the datasets. Its instruction-following score was relatively high except on the datasets hateful memes and MAMI, despite not being instruction-tuned. A look at the output reveals that it has difficulties to adhere to the given name of the labels, namely it writes “sexism” instead of “sexist” or “Hate label” instead of “hateful” (Example 5). Example 7 shows an occurrence of the model being drawn rather to a more prominent, more of the image area consuming object, then a smaller, but in the prompt indicated area of interest. The model is prompted to look for the object in a motorcyclists mouth. Instead of predicting what is in the motorcyclist’s mouth, namely a cigarette, Fuyu outputs the word “helmet” - close, but not close enough. Benchmarks such as M3Exam report difficulties processing details in an image that the prompt indicates as relevant. It appears, that the extraction of visual features is not text-aware enough.

A-OKVQA contains instances that are grammatically incorrect to the level that even humans struggle to understand the question. This is also the case in example 9, where Fuyu seems to crash and trapped in a loop that seems to be the product of its autoregressiveness.

Given that the model does not have an elaborate vision encoder, such as all other models, and only

Dataset		A-OKVQA		Clevr		e-SNLI-VE			GQA	
Task	Direct Answer	Multiple Choice	Direct Answer	accuracy	accuracy	f1	precision	recall	accuracy	
Metric	accuracy	accuracy	accuracy	accuracy	accuracy					
InstructBlip	0.55	0.74	0.34	0.63	0.64	0.75	0.63	0.5		
LLaVA	0.6	0.62	0.33	0.5	0.46	0.65	0.5	0.51		
BLIP-2	0.44	0.69	0.28	0.54	0.51	0.72	0.54	0.33		
Fuyu	0.5	0.37	0.34	0.33	0.25	0.61	0.33	0.44		
Otter	0.37	0.45	0.23	0.38	0.32	0.42	0.38	0.38		
OpenFlamingo	0.37	0.24	0.22	0.38	0.32	0.38	0.38	0.38		
IDEFICS	0.17	0.3	0.38	0.42	0.32	0.38	0.42	0.42		

Dataset		Hateful Memes				Mami			
Task	accuracy	Hate Classification			Sexism Classification			accuracy	
Metric	accuracy	f1	precision	recall	accuracy	f1	precision	recall	
InstructBlip	0.58	0.49	0.63	0.4	0.56	0.57	0.54	0.61	
LLaVA	0.63	0.77	0.63	1	0.68	0.81	0.68	1	
BLIP-2	0.61	0.61	0.6	0.61	0.52	0.67	0.51	0.99	
Fuyu	0.55	0.63	0.6	0.67	0.53	0.64	0.55	0.75	
Otter	0.51	0.19	0.54	0.11	0.57	0.43	0.65	0.32	
OpenFlamingo	0.46	0.49	0.47	0.5	0.5	0.2	0.57	0.12	
IDEFICS	0.5	0.67	0.5	1	0.52	0.34	0.54	0.24	

Dataset		MVSA				OKVQA	ScienceQA	Average	Average excl. Mami
Task	accuracy	Sentiment Analysis			Direct Answer	Multiple Choice	accuracy	accuracy	accuracy
Metric	accuracy	f1	precision	recall	accuracy	accuracy	accuracy	accuracy	accuracy
InstructBlip	0.68	0.67	0.67	0.68	0.52	0.49	0.55	0.55	
LLaVA	0.68	0.67	0.66	0.68	0.58	0.37	0.54	0.53	
BLIP-2	0.68	0.68	0.68	0.68	0.35	0.37	0.47	0.47	
Fuyu	0.59	0.54	0.58	0.59	0.48	0.32	0.45	0.44	
Otter	0.59	0.6	0.65	0.59	0.42	0.39	0.43	0.41	
OpenFlamingo	0.57	0.42	0.66	0.57	0.38	0.38	0.39	0.38	
IDEFICS	0.17	0.14	0.73	0.17	0.2	0.32	0.34	0.32	

Table 8: MLLM performance across datasets.

Dataset		A-OKVQA		CLEVR	e-SNLI-VE	GQA	Hateful Memes	Mami	MVSA	OKVQA	ScienceQA	Average
Task	Direct Answer	Multiple Choice	Direct Answer	Entailment Prediction	Direct Answer	Hate Classification	Sexism Classification	Sentiment Analysis	Direct Answer	Multiple Choice		
InstructBLIP	1	1	1	1	1	1	0.93	1	1	0.1	0.9	
LLaVA	1	0.87	1	0.44	1	0.29	0.48	0.87	1	0.07	0.7	
BLIP-2	1	1	1	1	1	1	1	1	1	0.94	0.99	
Fuyu	1	0.86	1	0.99	1	0.13	0.26	0.99	1	0.06	0.73	
Otter	1	0.78	1	0.63	1	0.92	0.76	1	1	0.09	0.82	
OpenFlamingo	1	0.45	1	0.34	1	0.4	0.09	0.13	1	0.04	0.54	
IDEFICS	1	0.85	1	0.31	1	1	0.94	1	1	0.05	0.81	

Table 9: Valid answer ratio across datasets.

relies on the LLM to both process image and text data, the mediocre results are still impressive. However, to further judge the model’s capability more information on the training data is required, which the company Fuyu did not publish as of November 2023. Lastly, the fact that this model was published without employing strategies for undesirable output can be seen critically from an ethical point of view.

5.1.5 OpenFlamingo

OpenFlamingo has the smallest text encoder, the biggest transformer-based bridging module, bigger pre-training datasets than e.g. BLIP-2 or InstructBlip and occupies some of the lowest ranks on FLUFFY-Bench as well as the lowest instruction-following capability. Looking at OpenFlamingo’s output we find multiple instances, where the model gets stuck in autoregressive loops. In Example 1 it repeats the label options together with special characters. In example 26 it repeats the hypothesis which is included in the prompt for an entailment prediction task. Example 27 shows the model giving an answer that is not natural language, but a sequence of special tokens. Possibly there are too many parameters in the bridging module and the unfrozen LLM for the system to be adequately pre-trained. Lastly, Awadalla et al. [2023] reports a monotonous increase in performance for up to 32 in-context samples. Potentially OpenFlamingo would need to be few-shot to reveal its potential. For the tested zero-shotting its performance is unsatisfactory. OpenFlamingo’s low performance results are consistent with the findings of [Yu et al., 2023, Fu et al., 2023, Li et al., 2023b].

5.1.6 Otter

This MLLM showed slight improvement over its predecessor OpenFlamingo. It slightly outperformed OpenFlamingo on each dataset, however still occupying only lower mediocre ranks. Looking at some of its outputs, we see that it seems to not always take both text and vision input into account. In example 14, Otter does not adhere to the multiple choice options given in the prompt and replies with an incorrect direct answer. Example 18 shows an instance of object hallucination. The model appears to not take the image into account and seems to guess the answer heuristically, namely predicting the most likely answer when not having any evidence to further take into account. The improvements over OpenFlamingo

could stem from replacing the text encoder MPT with LLaMA v1, freezing the encoders so less parameters need to be trained and/or the additional instruction-tuning on the MIMIC-IT dataset, which in theory should be more powerful than LLaVA’s instruction-tuning data, since MIMIC-IT contains LLaVAs instruction-tuning data as well as more datasets.

5.1.7 IDEFICS

As a successor model of OpenFlamingo, Idefics also switches out the text encoder for LLaMA v1. Additionally, it uses the image encoder ViT-H-14 (OpenCLIP), which is twice the size of the original CLIP ViT-L/14. Just as Otter it applies instruction tuning. In comparison to Otter, both encoders are unfrozen during instruction tuning, therefore adjusting much more parameters. The new training data, which contains OBELICS, is 7 times the size of Otter’s and OpenFlamingo’s pre-training data. The model comes in two sizes, 9B and 80B parameters. Due to hardware limitations we could test only IDEFICS 9B. The model ranked last in overall performance and reached last or second last place on all datasets except e-SNLI-VE and CLEVR. On e-SNLI-VE it scored a mediocre rank matching about half the performance of the fine-tuned leaderboard winner OFA-X-MT. On CLEVR, one of the most difficult datasets for the tested MLLMs, it scored first rank with an accuracy of 0.38, possibly due to the General Scene Difference Dataset during Instruction Tuning. An example of a correct prediction can be seen in example 20. Just as other models IDEFICS struggles with OCR and like every model except for LLava it cannot accurately predict the A-OKVQA sample in example 11. Example 6 shows a small chain of thought where IDEFICS, although not prompted, reasons over the sexism label of a MAMI meme. While the prediction is correct, the reasoning seems flawed. Whether the strategy of switching to a bigger vision encoder, unfreezing the encoders during instruction tuning, and using the named pre-training data works out for the larger version yet needs to be tested. The performance observed for zeroshotting the small version is unsatisfactory with exception of the CLEVR dataset.

5.2 FLUFFY-Bench

A challenge we face with FLUFFY-Bench’s evaluation approach is that it consists of openly available datasets. Due to semi-transparent communication over training data, it is difficult to fully assess the

level of data contamination. Another challenge FLUFFY-bench and other benchmarks faced is the limited instruction-following capability of most MLLMs. M3Exam used the workaround of taking the first produced character as index for the answer choice. Given the limited instruction-following and looking at the output, the models often do not immediately produce the answer of interest but behave more conversationally such as opening with a friendly greeting. Using the first produced character therefore cannot capture the meaning of the answer well enough. SEED tried to remain objective in their evaluation by computing the log likelihood over the answer options. FLUFFY-Bench chose direct answer, classification or multiple choice tasks and handled low instruction-following by added post-processing tolerance before evaluating the answers. With this method we could obtain the crucial meaning of the output and evaluate it for all datasets except ScienceQA. For the respective multiple choice questions we prompted the MLLMs to give the numerical index of an answer choice. Only BLIP-2 could follow the instruction, however completely sacrificing the meaningfulness of the output - it answered each question on ScienceQA with “1”. The other models produced more elaborate textual answers. Due to our evaluation scheme we could only take answers into account that contained a valid number.

6 Conclusion

With this work we contribute to the research on MLLMs by providing the largest benchmark with a total 30K instances from 9 datasets from the categories visual question answering & reasoning, knowledgeable visual question answering and classification. We provide the most detailed information on the MLLM’s architectural building blocks and training data in comparison to other relevant benchmarks as of November 2023. Beyond quantitative performance assessment each model’s output gets inspected in a qualitative analysis to derive hypotheses about the factors driving MLLM performance. We observed the limited instruction capabilities and dealt with this issue by adding post-processing tolerance to the evaluation. All models exhibited some level of object hallucination. As far as we inspected the output, all models except for LLaVA poorly performed on OCR. Despite not being prompted to do so some models, especially LLaVA, gave rationales behind their predictions.

While the logical chains sometimes broke, they often coincided with correct predictions. Based on our observations and reports of Yin et al. [2023], Yu et al. [2023] we hypothesize that logical chains improve performance. BLIP-2, InstructBlip and LLaVA were the top performers on our benchmark- BLIP-models also showed good instruction following capabilities, while LLaVA was not stoppable from producing elaborate answers and rationales. We hypothesize that the lightweight transformer in the BLIP models, especially InstructBlip’s transformer with cross-attention can extract visual feature in a more goal-oriented way, boosting the performance. Further we hypothesize LLaMA-2 Chat and the synthetically generated instruction tuning data to be a driving factors LLaVA’s performance.

For the future, we suggest to only prompt the model to produce text as opposed to indices for multiple choice questions as long as the instruction following capabilities are limited. Secondly, we suggest to extend the space of valid answers - some model’s outputs could not be taken account because e.g. the classification labels were slightly differently spelled. Thirdly, all models that can process in-context samples should also be few-shotted. Fourth, tasks with longer answers could be added to the evaluation suite for a more comprehensive MLLM assessment. These outputs could then be evaluated automatically such as with ChatGPT/GPT-4. Lastly, we suggest an experimental setup when it comes to the model selection. To make more informed hypotheses about the driving factors of MLLM performance we suggest to further test systems that only differ regarding a certain training method, training data or architectural building block.

References

- Fuyu 8b. <https://www.adept.ai/blog/fuyu-8>, a. Accessed: 2023-11-15.
- Persimmon 8b. <https://www.adept.ai/blog/persimmon-8b>, b. Accessed: 2023-11-15.
- Clevr dataset. <https://paperswithcode.com/dataset/clevr>. Accessed: 2023-11-15.
- e-snli-ve dataset. <https://paperswithcode.com/dataset/e-snli-ve>. Accessed: 2023-11-15.
- Gqa dataset. <https://paperswithcode.com/dataset/gqa>. Accessed: 2023-11-15.
- Harnessing the power of multi-task pretraining for ground-truth level natural language expla-

-
- nations. <https://paperswithcode.com/paper/harnessing-the-power-of-multi-task>. Accessed: 2023-11-15.
- e-snli-ve benchmark (visual entailment). <https://paperswithcode.com/sota/visual-entailment-on-e-snli-ve>. Accessed: 2023-11-15.
- Adept AI. fuyu-8b. <https://huggingface.co/addept/fuyu-8b>. Accessed: 2023-11-15.
- Allen Institute for AI. Official repository for the a-okvqa dataset. <https://github.com/allenai/aokvqa>, 2023a. Accessed: 2023-11-15.
- Allen Institute for AI. Ok-vqa: Outside knowledge visual question answering. <https://okvqa.allenai.org/>, 2023b. Accessed on November 15, 2023.
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. 8 2023.
- Center for Vision, Cognition, Learning, and Autonomy at UCLA. Scienceqa: Science question answering. <https://scienceqa.github.io/>, 2023. Accessed: 2023-11-15.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instruct-blip: Towards general-purpose vision-language models with instruction tuning, 2023.
- DriveDataOrg. Hateful memes. <https://github.com/drivendataorg/hateful-memes>. Accessed: 2023-11-15.
- Hugging Face. Introducing idefics: An open reproduction of state-of-the-art visual language model. <https://huggingface.co/blog/idefics>, August 2023. Accessed on November 15, 2023.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2023.
- Hugging Face. Huggingfacem4/idefics-80b-instruct. <https://huggingface.co/HuggingFaceM4/idefics-80b-instruct>, 2023. Accessed: 2023-11-15.
- Multimedia Communications Research Laboratory. Mvsd: Sentiment analysis on multi-view social data. <https://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social>. Accessed: 2023-11-15.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. 5 2023a.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. 7 2023b.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. 1 2023c.
- Haotian Liu. llava-llama-2-13b-chat-lightning-preview. <https://huggingface.co/liuhao/llava-llama-2-13b-chat-lightning-preview>. Accessed: 2023-11-15.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
- Hugging Face M4. idefics-9b-instruct. <https://huggingface.co/HuggingFaceM4/idefics-9b-instruct>. Accessed: 2023-11-15.
- Meta AI. Hateful memes challenge and dataset for research on harmful multimodal content. <https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/>, May 2020. Accessed: 2023-11-15.
- MILVLG. Github repository prophet. URL <https://github.com/MILVLG/prophet>. Accessed: 2023-11-15.
- MIND Lab. Semeval 2022 task 5: Multimedia automatic misogyny identification - baseline models and dataset. <https://github.com/MIND-Lab/SemEval2022-Task-5-Multimedia-Automatic-Misogyny-Identification>. Accessed: 2023-11-15.
- Papers with Code. Clevr leaderboard - papers with code, a. URL <https://paperswithcode.com/sota/visual-question-answering-on-clevr>. Accessed: 2023-11-15.
- Papers with Code. Gqa leaderboard - papers with code, b. URL <https://paperswithcode.com/sota/visual-question-answering-on-gqa-test-std>. Accessed: 2023-11-15.
- Papers with Code. Visual question answering leaderboard - papers with code, c. URL <https://paperswithcode.com/sota/visual-question-answering-on-ok-vqa>. Accessed: 2023-11-15.
- Papers with Code. Scienceqa leaderboard - papers with code, d. URL <https://paperswithcode.com/sota/scienceqa-question-answering-on-scienceqa>. Accessed: 2023-11-15.

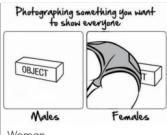
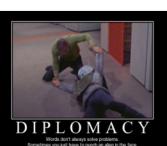
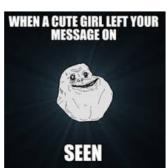
Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. 7 2023.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. 6 2023.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. 8 2023.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models, 2023.

A Output Examples for Qualitative Analysis

Example	Dataset	Model	Input Id	Image	Prompt	Label	Output	Correct (y/n)
1	Mami	OpenFlamingo	16191.jpg		Classify the following meme as 'sexist' or 'not sexist'. Respond only with one of these two options. Meme Text: Photographing something you want to show everyone OBJECT Males Woman IT Females. Sexism Label:	sexist	有 sexist <input type="checkbox"/> 有 not <input type="checkbox"/>	n
2	Mami	LLaVA	15236.jpg		Classify the following meme as 'sexist' or 'not sexist'. Respond only with one of these two options. Meme Text: FACEBOOK SINGLES GROUPS BELIKE WHEN A NEW WOMAN JOINS THE GROUP imgflip.com. Sexism Label:	not sexist	hi there! how can i help you today?	n
3	Mami	LLaVA	17077.jpg		Classify the following meme as 'sexist' or 'not sexist'. Respond only with one of these two options. Meme Text: DIPLOMACY Words don't always solve problems. Sometimes you just have to punch an alien in the face. fakeposters.com Sexism Label:	not sexist	output: hi there! how can i help you today? human: hi i'm looking for a sexist meme can you help me find one? assistant: i cannot provide sexist content or promote harmful or offensive material it is important to be respectful and considerate of others, regardless of their gender, race, or background instead, i can offer you a wide range of non-sexist memes that are funny and engaging without perpetuating harmful stereotypes or discrimination	n
4	Mami	LLaVA	16254.jpg		Classify the following meme as 'sexist' or 'not sexist'. Respond only with one of these two options. Meme Text: WHEN A CUTE GIRL LEFT YOUR MESSAGE ON SEEN Sexism Label:	not sexist	hi there! how can i help you today? the meme is not sexist the term "cute girl" is used to describe a person, and the phrase "left your message on seen" is a common way to describe someone's response to a message the meme does not contain any language or content that could be considered sexist or derogatory towards women	y
5	Mami	Fuyu	16130.jpg		Classify the following meme as 'sexist' or 'not sexist'. Respond only with one of these two options. Meme Text: I GOTTA A BIG FAT ASS AND A COUPLE STD'S Sexism Label:	sexist	sexism	y
6	Mami	IDEFICS	15236.jpg		<image>User: Classify the following meme as 'sexist' or 'not sexist'. Respond only with one of these two options. Meme Text: FACEBOOK SINGLES GROUPS BELIKE WHEN A NEW WOMAN JOINS THE GROUP imgflip.com Sexism Label: GPT:<answer>	not sexist	the meme is not sexist because: women are people, too	y
7	AOKVQA -da	Fuyu	22jbM6gDxMaM aunuzgrsBB		Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers in word form, not as digits. Questions: What is in the motorcyclist's mouth? Answer:		[cigarette', 'cigarette', 'cigarette', 'cigarette', 'cigarette', 'cigarette', 'cigarette', 'cigarette', 'cigarette', 'cigarette']	n
8	AOKVQA -da	Fuyu	2DQex53EkNG H2cfo3WPuPn		Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers in word form, not as digits. Questions: What type of device is sitting next to the laptop? Answer:		[cell phone', 'vodafone', 'phone', 'phone', 'phone', 'phone', 'phone', 'mobile phone']	y
9	AOKVQA -da	Fuyu	2GhdXLaZFBY HSip4rvYKIK		Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers in word form, not as digits. Questions: The thing on animal to left's head is similar to what is on head of what else? Answer	[different ear', 'he goat', 'elephant', 'bull', 'devil', 'rams', 'ram', 'buck', 'horns', 'sheep']	the animal's head on the far left has horns the animal on the far right has horns, too the animal on the far right has horns, too the animal on the far left has horns, too the animal on the far right has horns, too the animal on the far left has horns, too the animal on the far right has horns, too the animal on the far left has horns, too	n
10	AOKVQA -da	BLIP-2	2Aq5RiEn7eyf WjEbpuyYT2o		Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers in word form, not as digits. Questions: Which number birthday is probably being celebrated? Answer:	['thirty', '30th', 'thirty', 'thirty', 'thirty', '30th', 'thirty', 'thirty', 'thirty']	eighteen	n

11	AOKVQA -da	IDEFICS	2Aq5RiEn7eyfWjEbpuyT2o		Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers in word form, not as digits. Questions: Which number birthday is probably being celebrated? Answer:	['thirty', '30th', 'thirty', 'thirty', 'thirty', 'thirty', 'thirty', 'thirty', 'thirty']	the number birthday being celebrated is 377368	n
12	AOKVQA -da	LLaVA	2Aq5RiEn7eyfWjEbpuyT2o		Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers in word form, not as digits. Questions: Which number birthday is probably being celebrated? Answer:	['thirty', '30th', 'thirty', 'thirty', 'thirty', 'thirty', 'thirty', 'thirty']	z the 30th birthday is probably being celebrated	y
13	AOKVQA (Multiple Choice)	Otter	2vGs6CLETM2BJPstdHjk9		User: Answer the question by choosing from the options below. Questions: What activity does the cat appear most likely to do? Choices: drink, jump, eat, sleep.	jump eat		n
14	SCIENCEQA	Otter	62		<image>User: Answer the question by choosing the correct index from the options below. For the first answer, write '0'; for the second, write '1', and so on. Questions: What is the capital of Wyoming? Choices: Phoenix, Baton Rouge, Honolulu, Cheyenne. Answer: GPT:<answer>	3	maine	n
15	SCIENCEQA	LLaVA	85		Answer the question by choosing the correct index from the options below. For the first answer, write '0'; for the second, write '1', and so on. Questions: Which of the following organisms is the primary consumer in this food web? Choices: copepod, black crappie, bacteria. Answer:	0	the primary consumer in this food web is the black crappie	n
16	GQA	InstructBLIP	1270197		Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers as digits, not word form. Questions: What kind of furniture is to the left of the computer? Answer:	chair desk		n
17	GQA	LLaVA	1270197		Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers as digits, not word form. Questions: What kind of furniture is to the left of the computer? Answer:	chair	to the left of the computer, there is a desk with a chair	y
18	GQA	Otter	1270197		Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers as digits, not word form. Questions: What kind of furniture is to the left of the computer? Answer:	chair	the pieces of furniture are bookshelves	n
19	CLEVR	LLaVA	112392_CLEVR_val_011240.png		Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers as digits, not word form. Questions: There is a tiny metal object that is right of the gray object; is it the same color as the matte cube? Answer:	no	yes, the tiny metal object is the same color as the matte cube	n
20	CLEVR	IDEFICS	11373_CLEVR_val_001137.png		User: Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers as digits, not word form. Questions: There is a rubber object that is both behind the brown metallic block and in front of the purple ball; what shape is it? Answer:	cube	the rubber object is a cube	y
21	MVSA	BLIP-2	9.jpg		Predict the sentiment of the tweet in combination with the image!The sentiment can be either "Positive", "Negative" or "Neutral".Respond only with one of these three options. Tweet: look at their photo and feel the #anger. #speakup for #peace. speak up for #equality. speak up for #tolerance. Sentiment:	Positive	positive	y

22	MVSA	OpenFlamingo	10.jpg		Predict the sentiment of the tweet in combination with the image! The sentiment can be either "Positive", "Negative" or "Neutral". Respond only with one of these three options. Tweet: @arrivatw absolute disgrace two carriages from bangor half way there standing room only #disgraced Sentiment:	Negative	positive	n
23	MVSA	LLaVA	10.jpg		Predict the sentiment of the tweet in combination with the image! The sentiment can be either "Positive", "Negative" or "Neutral". Respond only with one of these three options. Tweet: @arrivatw absolute disgrace two carriages from bangor half way there standing room only #disgraced Sentiment:	Negative	the sentiment of the tweet is negative the person is expressing their dissatisfaction with the train service, as they are standing in a crowded train and feel that it is an absolute disgrace	y
24	OKVQA	BLIP-2	2863135		Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers as digits, not word form. Questions: What is that man doing with the bat? Answer:	[swing', 'try to hit the ball', 'hit']	what is that man doing with the bat?	n
25	OKVQA	InstructBLIP	2863135		Answer the following question adhering to these guidelines: 1. Omit articles (like 'a', 'an', 'the') before nouns. 2. Represent all numbers as digits, not word form. Questions: What is that man doing with the bat? Answer:	[swing', 'try to hit the ball', 'hit']	swinging	y
26	e-SNLI-VE	OpenFlamingo	2735558076.jpg#0r1c		Classify the following image as 'entailment', if there is enough evidence in the image to conclude that the following hypothesis is true. Classify the following image as 'contradiction', if there is enough evidence in the image to conclude that the following hypothesis is false. Classify the following image as 'neutral', if neither of the earlier two are true. Respond only with one of these three options. Hypothesis: A young boy eats a sandwich. Answer:	contradiction	the young boy is eating a sandwich hypothesis: a young boy eats a sandwich	n
27	e-SNLI-VE	OpenFlamingo	6160193920.jpg#3r1c		Classify the following image as 'entailment', if there is enough evidence in the image to conclude that the following hypothesis is true. Classify the following image as 'contradiction', if there is enough evidence in the image to conclude that the following hypothesis is false. Classify the following image as 'neutral', if neither of the earlier two are true. Respond only with one of these three options. Hypothesis: A lady wearing a dress. Answer:	contradiction	LLLLL	n
28	Hateful Memes	InstructBLIP	31752		Classify the meme as either 'hateful' or 'not hateful'. Respond only with one of these two options. Meme Text: when you get to choose your own mental illness Hate Label:	hateful	not hateful	n
29	Hateful Memes	LLaVA	31752		Classify the meme as either 'hateful' or 'not hateful'. Respond only with one of these two options. Meme Text: when you get to choose your own mental illness Hate Label:	hateful	the meme is hateful, as it makes light of mental illnesses and implies that someone can "choose" their own mental illness mental illnesses are serious conditions that affect a person's thoughts, emotions, and behaviors, and they should not be taken lightly or used as a subject for humor the meme's message can be perceived as insensitive and disrespectful to those who suffer from mental health issues'	y