A Multi-Modal Large Language Model for Enhancing Health Literacy: Generating Conversational Topics from Images for Educational Use

Kilian Carolan #22266471
Laura Campbell #22269176
School of Computing, Dublin City University (DCU), Ireland
Email: kilian.carolan5@mail.dcu.ie, laura.campbell26@mail.dcu.ie

***Abstract*—This practicum explores using generative AI to produce conversation topics from user-uploaded photos, aimed at enhancing health literacy. The aim is to trigger educational conversations between student and teacher about health literacy issues present in the uploaded photos. We developed a proof-of-concept system to process photos and generate natural language descriptors discussing health literacy aspects. The study compared a unified Multi-Modal Large Language Model (MM-LLM), LLaVA, which processes images and text directly, with an ensemble approach integrating separate vision and language models. We experimented with Parameter-Efficient Fine-Tuning (PEFT) and In Context Learning on models including Falcon-7B-Instruct, LLaMA3-7B-Instruct, and LLaVA for domain-specific tasks. The fine-tuned LLaMA3-7B-Instruct model excelled in generating relevant outputs, while Falcon-7B models struggled with topic relevance and hallucinations. Among vision models, Florence-2 performed comparably to LLaVA and better than BLIP-2 in accuracy and reducing hallucinations. A human evaluation survey favoured the LLaMA3-Florence combination over the unified LLaVA system. Marginal differences were observed between the two ensemble approaches using LLaMA3. While MM-LLMs show promise for health literacy education, practical deployment requires careful consideration of computational resources, regulatory constraints, and information accuracy.**

I. Introduction

The PhotoVoice+ project at the Insight Centre in DCU, funded by Science Foundation Ireland, aims to develop a smartphone application to enhance the health literacy of young people. The idea behind PhotoVoice+ is that schoolchildren upload ordinary everyday photos of their environment taken with their smartphones and these are used to trigger educational conversations with their teacher about health literacy in their own lives. This practicum involves exploring the use of generative Artificial Intelligence (AI) to facilitate conversations about health literacy based on such user-uploaded images. To assist the development of such a smartphone application for PhotoVoice+, our goal is to build a proof-of-concept system that can process an image and provide a natural language response grounded in the concept of health literacy.

An extensive literature review was conducted, focusing on the current state of Large Language Models (LLMs), particularly visual Multi-Modal LLMs (MM-LLMs) [1]. Based on this several research questions were proposed. The primary research question of this work is: "How can MM-LLMs be utilised for health literacy conversations based on user-uploaded images?" The practicum aims to evaluate the feasibility of locally deployed pre-trained open-source LLMs and MM-LLMs for more specific tasks than those for which they were initially trained. In doing this we consider factors such as computational costs and the availability of data to fine-tune models. We aim to assess some shortlisted MM-LLM systems on their ability to process a user-uploaded image and output accurate information grounded in health literacy.

Secondary research questions include:

1) Exploring the best methods to combine vision and text models for the primary goal, specifically whether an application built on either a natively trained MM-LLM or a multi-stage inference pipeline (i.e., an ensemble of a vision model paired with an LLM) can be a viable alternative to proprietary MM-LLMs like Chat-GPT4, particularly when constrained by hardware;
2) Can fine-tuning on a limited set of domain-specific text improve the models output;
3) Investigating whether fine-tuning a model on a set of examples can provide an alternative to few-shot In-Context Learning (ICL) or possibly outperform it.

II. Background

A. Large Language Models

In 2017, the transformer architecture was introduced which laid the groundwork for what would become LLMs. The transformer architecture utilised the attention mechanism for parallel processing of text data which removed the limitation of handling long-range dependencies. This provided Language Models (LMs) with the capabilities to comprehensively analyse input, understanding the relationship and importance of each component [2]. Following this development a year later, Google introduced Bidirectional Encoder Representations from Transformers (BERT), and OpenAI introduced their first Generative Pre-trained Transformer (GPT). Through

vast amounts of data and increased computing power, these transformer-based models evolved LMs into LLMs. These models gained significant attention when OpenAI's ChatGPT was launched in November 2022, showcasing the ability of LLMs to comprehend and generate human language [3], [4].

By 2023, rapid advancements in LLMs spurred competition between industry giants such as OpenAI (GPT-4) and Google (PaLM) to establish dominant private models in the market. Creating open-source LLMs posed significant financial challenges due to the vast data and computational resource requirements. However, this changed when Meta introduced their open-sourced LLaMA model, aimed at promoting innovation, security, and collaboration within the open-source community. Advancements in LLMs have been driven by the success of the open-source community, tackling challenges that even industry leaders like Google have struggled to address [5]–[8].

LLMs claim several advantages over closed models, including cost-effectiveness, transparency, flexibility, and alignment with emerging regulations such as the European Union AI Act due to increased control over data. These models empower developers by providing greater control over their data and enabling diverse applications across various industries. However, despite these benefits, there are notable concerns. These include the absence of formal support agreements, the risk of development discontinuation, and challenges related to adhering to open-source principles, as evidenced by the usage conditions of Meta's LLaMA-2 model [9]–[11]. While some recent work have proposed alternatives to the transformer model such as the state-space approach MAMBA [12], transformer-based models dominate the space.

B. Image Based MM-LLMs

There have been increased efforts to apply transformer-based models to different modalities besides text such as vision tasks. While other modalities like sound have also been explored, these are out of the scope of this research. Image-based MM-LLMs, also known as Large Vision Models (LVMs), in simple terms, integrate the vision and language capabilities of the transformer architecture. LVMs typically achieve this by combining pre-trained vision encoders with a pre-trained language decoder. This encoder/decoder pairing is then linked through the use of a bridge layer sometimes called a projection layer such as found in LLaVA and BLIP models [13], [14]. The advantages of this approach compared to developing the system from scratch include reusing the LLM's ability in language generation and reasoning, increased computational efficiency similar to the transfer learning approach seen across deep learning topics as utilising a model that can already effectively encode visual information, and decreasing the training cost. This leaves a smaller task of training a model to link the image and natural language understanding through the projection layer.

Many open-source image-based MM-LLMs have been proposed such as MiniGPT-4 [15], m-PlugOwl [16], the previously mentioned LLaVA [13], and BLIP [14]. Leading companies have also introduced proprietary models such as Google's Gemini and OpenAI's ChatGPT-4.

The integration of visual and text models in MM-LLMs varies. For MiniGPT-4, two stages of training were completed using frozen models aligned through a projection layer for each stage. LLaVA also used two stages of training, the first stage but the encoder and decoder remained frozen however, the model fine-tuned the LLM in the second stage training stage. m-PlugOwl used its first stage of training for the vision model, and froze the LLM. In the second stage of training, the LLM was fine-tuned while the visual encoder was frozen. The optimal approach for co-training LLMs with vision models still remains an open research question, in addition the training of these models is prohibitively expensive for example with LLaVA 1.5 was trained on 8 A100 GPU with 80GB of memory [17].

Flash Attention is a method that reports to enhances speed and memory efficient in Transformers. This is achieve through a tiling and re-computation technique that minimises the number of memory read/write operations between levels of GPU high bandwidth memory (HBM) and GPU. As a result, transformers can be trained faster and longer context can be accommodated, leading to higher quality models with improved performance [18]. Additionally, Rotary Positional Embedding (RoPE) is a technique which rotates token embeddings in a high dimensional space, preserves the original information while enhancing positional understanding and it can also optimise memory usage and computational efficiency [19].

C. Fine-tuning LLMs

Research suggests that fine-tuning, the process of modifying the parameters of a pre-trained model, has been shown to improve a model's ability to adapt to a specific task. However research on fine-tuning LLMs for domain-specific tasks remains limited, with no well-established methods developed to ensure exemplary results.

In the context of health literacy and medical applications, the topic of this practicum, several studies have explored fine-tuning LLMs with medical examinations and clinical data, aiming to replicate human-level performance in these domains. However, a systematic review of these studies noted significant challenges in ensuring information accuracy and addressing the issue of hallucinations [20]. One study investigated fine-tuning GPT-4 to transform hospital discharge summaries into a format comprehensible for patients that may struggle with health literacy. Although the LLM generated shorter and more understandable summaries, only 54% were successfully transformed. The remaining summaries were less accurate, with some being flagged as potential safety risks. Despite achieving high confidence in training, the presence of hallucinations was still concerning [21]. Many of these studies on fine-tuning LLMs for domain-specific medical tasks have highlighted the challenges posed by hallucinations and the level accuracy, revealing a notable gap in the comprehensive evaluation of LLMs within medical educational settings [20].

Fine-tuning a pre-trained model typically requires vast amounts of data and significant computational resources [22]. Parameter efficient fine-tuning (PEFT) is a technique that aims to address these challenges by updating only specific components of the model or incorporating additional elements, such as adapter layers, during the tuning process. This approach allows for modifications to be made without altering the original architecture, while mitigating some of the computational cost of fine-tuning [23].

Low Rank Adaption (LoRA) is a PEFT method that trains LLMs on domain-specific data by freezing the weights of a pre-trained LLM and introducing trainable low-rank matrices to capture the key changes needed for the domain-specific adaption. These low-rank matrices create the LoRA adapter. Once fine-tuned the LoRA adapter is integrated into the original pre-trained LLM for inference. This approach reduces the number of trainable parameters that need to adjusted during fine-tuning, which subsequently reduces GPU memory requirements. As a result, computational resources and memory usage are minimised [24].

Quantised Low-Rank Adaption (QLoRA) is an extension of LoRA that adds quantisation into the fine-tuning process, so the low-rank matrices introduced by LoRA are quantised. Quantisation reduces the precision of the model parameters by converting high-precision values, such as floating-point numbers, to lower bit representations. This process further compresses the model, decreasing memory, storage and computational requirements. As a result, the reduced parameter size and lower computational requirements increases the speed of the training [25].

D. In-context Learning (ICL)

Prompts refer to instructions given to a LLM as natural language text. Prompting allows for the LLM to perform what is known as ICL, meaning the LLM can be adapted to a new task via prompts without changing the learned parameters of the model [26]. This is a particularly useful feature of such models when factoring in the vast amount of data needed to fine-tune a LLM [27] [1].

One and Few-shot prompt engineering, refers to the methods of giving the model one or more examples on which to structure its answer on. Zero-shot is when the prompt is not given any examples. There is conflicting evidence on whether PEFT or few-shot engineering is a better option to "customise" LLMs to a specific task. Few-shot engineering has been shown to improve smaller models ability to generate contextually relevant information. However other work claims that PEFT is both computationally cheaper and more accurate than providing in context examples [28].

E. Evaluation of Generative AI

Due to the challenges that arise in LLMs and MM-LLMs, such as hallucinations, inaccurate outputs, and disregarding instructions, it can be difficult to assess the models' outputs quantitatively. Therefore, several qualitative methods have been proposed. The R.A.C.C.A. framework [29], gives six criteria namely relevance, accuracy, completeness, clarity, coherence, and appropriateness, rated on a 1-5 scale. Other qualitative method include expert reviews, where the models' outputs are compared with those of a domain expert for evaluation, and the Elo rating system, which benchmarks LLMs through a chatbot arena where users vote for their favourite answers [30] [31].

Additional research focuses on assessing models against a range of benchmarks, including but not limited to common sense reasoning tasks, yes/no answers, and physical world understanding, such as HellaSwag [1]. However, these type of benchmarks are not without criticism. Research cites "difficulties in measuring genuine reasoning, adaptability, implementation inconsistencies, prompt engineering complexity, evaluator diversity, and the overlooking of cultural and ideological norms in one comprehensive assessment" [32]. This raises questions of the usefulness of benchmarks such as HellaSwag when choosing one LLM over another for real world tasks.

III. Model Overview

The following section will discuss the models we shortlisted for investigation and how we did that shortlisting. We discuss the two specific open source LLMs investigated in this work, Falcon and LLaMA. but first, three image captioning models are discussed, Florence-2, BLIP-2, and LLaVA-Next.

1) ***Falcon-7B-Instruct****:* Falcon-7B-Instruct is an instruction-tuned version of the Falcon-7B model, designed for unrestricted commercial use without royalties or usage fees. The model is trained on 1.5 trillion tokens. It utilises "The RefinedWeb Dataset", a carefully developed high-quality pre-training dataset. We know that the quality of training data can significantly influence the performance of LLMs and the series of three publicly available Falcon LLMs are said to have demonstrated powerful results from leveraging high quality data. However, it is noted in the documentation that the the dataset may contain stereotypes and biases commonly found online. Additionally, the RefinedWeb dataset incorporates alt text and links for images, making it "multi-modal-friendly" [33]–[35]. 15GB are required to run the model for inference and fine-tuning, making it accessible on modest hardware or cloud platforms [1]. The model also incorporates the Multi Query Attention mechanism, which improves handling of large-scale tasks by enhancing inference scalability. It does this by sharing one key head and a single value head across all unique query heads, which reduces GPU memory required for storage, enabling faster processing. Additionally, the model employs both flash attention and RoPE for computational resource efficiency [33], [34], [36].

2) ***LLaMA3-8b-Instruct****:* LLaMA3-8b-Instruct is an instruction-tuned implementation of the LLaMA3-8b model. LLaMA was designed to run on a single GPU, prioritising inference speed over training speed meaning tat on the surface this model should fit our research requirements. LLaMA has the objective of having the best possible performance on different computational budgets; for example, LLaMA 13b outperforms GPT3 on most benchmarks. To ensure stable

training, LLaMA utilises pre-normalisation to normalise the input of the transformers rather than normalising the output. The latest version, LLaMA3, is available in 2 sizes: 8B and 70B parameters [37]. The tokeniser has a vocabulary of 128k tokens, trained on a sequence of 8,192 tokens. The model itself is trained on a dataset seven times larger than LLaMA2.

Of particular interest to this project are LLaMA's safety features such a NSFW filters, given the PhotoVoice+ target audience. Grouped Query Attention (GQA) is used to increase inference efficiency. Some research suggests that fine-tuning a model like LLaMA – due to the quality of its pre-training dataset – can actually degrade performance [38]. However, this would imply that it is a good candidate for ICL, and the performance of both will be investigated in Section 3.

3) ***Florence-2****:* Florence-2 [39] is a new vision-based MM-LLM. What differentiates Florence-2 from other models is its size and the range of vision tasks available. With two main variants, Florence-2-B and Florence-2-L, at 0.223 and 0.77B parameters in size respectively. Florence-2 boasts a wide range of applications, including, but not limited to, image captioning, object detection, OCR, and image segmentation, all controlled through prompting. Trained as a generalist model, it scored comparably with specialist models like BLIP-2, which has 7.8B parameters, and outperformed Flamingo, which has 80B parameters, in image captioning tasks.

Florence-2 uses a vision encoder, DaViT, to convert images into visual token embedding. These embeddings are then combined with text embedding and processed by a transformer-based encoder-decoder. Florence-2 was trained on a curated collection from five separate data sources, resulting in 126 million samples. This highlights once again that the more data you give transformer-based models, the better they tend to perform, further backed by experiments in their research, which shows an increase in zero-shot performance as the number of training samples increase.

4) ***BLIP-2: instructblip-vicuna-7b*** *:* BLIP-2, introduced by Salesforce, is a pre-training strategy aimed at vision language understanding. The first stage involves using frozen image encoders to learn visual-text representations, while the second stage uses a frozen language model to generate vision-to-language understanding. This approach also introduced the Querying Transformer to act as a bridge between the image and text encoders. This strategy is realised by models such as BLIP-2 ViT-g FlanT5XL, and BLIP-2 ViT-g OPT6.7B, as well as InstructBLIP an instruction-tuned version of BLIP-2 aimed at visual question answering [40]. In this practicum, we investigate the instruct blip-vicuna-7b variant. In this model Vicuna-7b is used as the natural language decoder, and is actually an instruction fine-tuned version of LLaMA2 [41].

5) ***LLaVA-NeXt: llava-hf/llava-v1.6-mistral-7b-hf*** *:* Large Language and Vision Assistant (LLaVA) [13] introduced a method for connecting a vision encoder with an LLM to produce a MM-LLM for general purpose vision to language understanding. Using CLIP as a visual encoder and Vicuna [42] as the language decoder, the decoder was fine-tuned on 158k language-vision instructions taken from the MS-COCO data set [43], with the encoder remaining frozen. The fine-tuning on this dataset was trained over 3 epochs, with a learning rate of 2e-5, and a batch size of 32. Full Shard Data Parallel and gradient check pointing was utilised in an effort to save GPU memory.

This work was expanded upon in LLaVA 1.5 and later LLaVA-NeXT [44], by adding CLIP-ViT-L-336px with a multi-layer perceptron (MLP) projection. LLaVA-NeXT uses largely the same hyper-parameters as the original LLaVA, excluding the learning rate, which has been halved during pre-training. This is due to the fact the model now uses a MLP projection layer, rather than a linear projection layer as in the original model. Another important consideration is that the model can only handle one image at a time [45]. LLaVA-NeXt incorporates higher input image resolution, specifically four times greater, claiming to allow more detailed image understanding. LLaVA-NeXT highlights once again the prohibitive computational cost of training such a model, with the 34B parameter model utilising 32 A100 GPUs for 1 day of training, which they celebrate as being 100-1000 times faster than similar models. In this research, we will examine the llava-v1.6-mistral-7b variant. Mistral-7B is the natural language decoder in this version of LLaVA, it uses GQA for faster inference and Sliding Window Attention (SWA) to handle longer sequences at smaller cost [46].

IV. Methods

To develop an open-sourced MM-LLM capable of effectively generating conversational topics from images for health literacy education, a comprehensive approach was adopted. This involved exploring various datasets, models, and tuning techniques to ensure alignment with the overall educational goal.

A. Hardware

The experiments in this research utilised a NVIDIA RTX A6000 GPU which has 48GB of VRAM.

B. Data preparation

The data preparation for training the MM-LLM involved gathering several types of datasets. For evaluating the vision models, open-source image datasets were selected to determine and assess the performance of suitable models. Additionally, the PhotoVoice+ research team provided a set of previously unseen photos specific for the PhotoVoice+ use case which were used for further evaluations as many vision models are pre-trained on commonly available open-source datasets. For the language model, text corpora with domain-specific knowledge were utilised.

1) Text Corpus Datasets: The PhotoVoice+ team provided a health literacy ontology, primarily composed of academic papers addressing the definition of health literacy, its significance in public health, adolescent health literacy, and related health behaviours. Additionally, the ontology included documents on the use of inclusive language. To augment the training data for the language models, Question and Answer (Q&A) pairs were

generated from the health literacy ontology. Q&A datasets can enhance focused learning and provide structured context, which is beneficial for model training. Falcon-7B-Instruct was used to generate Q&A pairs. Although the questions produced were of a high standard, the answers were sub-optimal with inaccuracies and hallucinations. As a result, the dataset was altered to provide a question, with the paragraph from which the question was derived being provided as the answer for context.

Gaps were identified in the content of the health ontology that were relevant to the MM-LLM's objectives. While the ontology thoroughly covered various aspects of health literacy, and its related discussions. It lacked educational material aimed at teaching health literacy. To address this, an open-source medical dataset from HuggingFace called "HealthAssistant115" was included in the list of datasets to be used. The medical dataset consists of Q&A related to overall general physical and mental health well being.

As the training of the LLMs progressed and outputs were evaluated, it became evident that a custom dataset, purpose built to the domain-specific use case was necessary. A dataset was create to both fine-tune the LLM model and few-shot prompt tune the MM-LLM. The objective was to determine whether tuning the dataset on many prompts could replace or enhance the ICL approach of few-shot prompting. This example tuning dataset consisted of the system prompt, example image descriptions and corresponding example outputs of conversation topics related to health literacy. This example tuning dataset was developed through a combination of human generated examples and examples produced by ChatGPT4, which were carefully manually reviewed. The goal was to assess the influence of a limited dataset on the model's responses and the extent to which it could affect output quality. By following the LoRA PEFT methodology outline below, the model was fine-tuned on 180 examples, with 20 reserved as a validation set, this was repeated for Falcon. Although the dataset was limited in size, consisting of 200 rows, it was specifically designed to meet the requirements for the PhotoVoice+ application.

2) Image Dataset: To assess the various vision models, 50 images were randomly selected from the Flickr30 dataset [47]. This dataset was selected as it serves as a proxy for the type of everyday photos that may be submitted by PhotoVoice+ users. Additionally, the overall goal of finding a vision model was to aim for a general image captioning model rather then one fine-tuned on a specific task, such as image captioning.

C. Fine-tuning

Initial fine-tuning experiments were conducted with the Falcon model, resulting in various fine-tuned versions, including those using QLoRA or LoRA. The fine-tuning process explored two different methods for comparison. The first method involved combining all three datasets into a single large combined dataset for training the model. The second method involved sequentially fine-tuning the model on each specific dataset, starting with the most representative dataset, the health ontology, followed by the Q&A dataset, and finally the medical dataset.

For QLoRA fine-tuning on the health literacy ontology, the pre-trained Falcon model's precision was quantised to 4 bits. The Torch compute type "bfloat16" was selected to help reduce memory usage and accelerate training computations. The LoRA configuration had the alpha parameter set to 32, which is used as a scaling factor for the weight matrices. The rank parameter, which defines the dimension of the low-rank adapter, was set to 16. While a higher rank accommodates larger datasets, an excessively high rank for a small dataset can lead to over-fitting. A dropout probability of 5% was applied to LoRA. These configuration resulted in 8.9% of trainable parameters being available to train. The training parameters included a learning rate set at 2e-4, with the optimiser "paged adamw 32bit". The learning schedule was constant, with a "warmup ratio" of 3% to allow the learning rate to gradually increase from zero to initial learning rate over the training period set at 400 steps, equivalent to 18 epochs. The model was trained using Hugging Face's "SFT-Trainer", a specialised trainer class for supervised fine-tuning of custom datasets on language models, which aids in minimising the code requirements.

For LoRA fine-tuning on the health literacy ontology, the alpha parameter was adjusted to 64, the rank parameter to 32. These configuration resulted in 9.3% of the parameters being trainable. If the LoRA configurations had been left the same as in QLoRA, only be 4.6% of the parameters would have been trainable. Therefore, the values were increased. The only difference in the training parameters was the training period which was set to 500 steps, equivalent to 22 epochs.

For the other datasets, the QLoRA and LoRA parameters remained the same. The some training parameters where adjusted, the train batch size set to 4, "gradient accumulation steps" set to 4 to average gradients across every 4 steps, learning rate was reduced to 1e-4, and the "warmup ratio" increased to 5%. The training period was set for 500 steps, equivalent to 40 epochs.

LLaMA3 was fine-tuned using LoRA with an alpha parameter set to 16 and a rank parameter of 8 resulting in 1% of trainable parameters. No dropout probability was applied. Stochastic Gradient Descent (SGD) was used as the optimiser, with a learning rate of 4e-4, a cosine learning scheduler, and a warm-up ratio of 10%. The training period was set to 500 steps, equivalent to 22 epochs. Subsequently, both Falcon and LLaMA were fine-tuned using LoRA on the example tuning dataset only, LLaMA employed the parameters previously stated, while Falcon used the parameters applied to the health ontology.

A higher learning rate was employed for the health ontology to ensure the model learned the key features and concepts during the initial training phase. For the next phase of training, the learning rate was lowered for the Q&A and medical data to preserve the knowledge gained in the first stage and to integrate the new data without comprising the core information from the health ontology. Sequential training allowed the

model to focus on learning on broad patterns related to health literacy, with the Q&A and medical dataset used to for further refinement. Additionally, sequential learning helped to stabilise training by enabling the model to gradually adapt to each dataset. This approach also mitigated the risk of catastrophic forgetting in the LLM, ensuring that previously learned information was retained effectively [48].

D. Vision Model Assessment

We aimed to assess how well these models could describe a presented image, i.e. image captioning task and identify potential options for the PhotoVoice+ project. Both members of the project team were to assess the outputs of the three models independently. Three vision models were loaded from hugging face: LlaVA-NeXT (llava-hf/llava-v1.6-mistral-7b-hf), BLIP-2Instruct (Salesforce/instructblip-vicuna-7b), and Florence-2 (microsoft/Florence-2-large). Since Florence-2 requires predefined prompts, the "MORE_DETAILED_CAPTION" option was selected. The BLIP-2 model was prompted to "Describe the image in detail", while the LLaVA model was asked to "Describe what is happening in the image". The reason for the discrepancy between the prompts was due to how the model interpreted the question, these prompts in experimentation appeared superficially to give comparable text and detail to the Florence-2 prompt.

The three models were then each given the same set of 50 photos from the Flickr30k dataset [47] in a loop to generate responses. These responses were collated into a google sheet. Two evaluators independently assessed the responses using three criteria (Accuracy, Completeness and Appropriateness) on a five-point Likert scale, similar to the R.A.C.C.A. [29]. The mean score for each criterion per assessor was calculated, and the mean scores across both assessors was then calculated. Hallucination trends were noted. Inference time was recorded as were the actual VRAM requirements to run inference on the models.

E. MM-LLM Systems: Unified vs Ensemble Approach

This study assesses the performance of a unified MM-LLM, capable of directly processing both images and text, against an ensemble approach that employs a multi-stage pipeline. The pipeline utilises two specialised models: one trained specifically for image processing and the other for natural language processing (NLP). The objective is to compare the efficacy of a comprehensive, generalised system, designed to perform a broad range of tasks at a high level of abstraction, against a system that integrates specialised models, tailored to perform well in domain-specific tasks.

Due to the lightweight nature of Florence-2, with 0.77B parameters, and its comparative performance to heavier methods described in the results section for image captioning, an architecture is proposed that incorporates Florence-2 into the prompting process of the instruct LLMs. In this setup, an image and the pre-trained Florence-2 prompts are used to generate text, which is then passed as part of the prompt to the LLMs, specifically Falcon-7b-instruct, LLaMA3-8b-instruct and their fine-tuned variants. This approach is compared against the LLaVA system, as this version of LLaVA was also trained to accept instruction.

System	Components
System 1	Florence-2 Large & Llama3-8b-Instruct
System 2	LLaVA Next: llava-hf/llava-v1.6-mistral-7b
System 3	Florence-2 & Llama3-8b-Instruct-ExampleTuned

Table I
FINAL THREE SYSTEMS USED IN USER SURVEY.

Few-shot example prompting was used to inform the system of the desired style of output, and the system prompt was iteratively developed. However, it is noted that as the desired output, especially during system development, is largely subjective. There is no entirely objective method to determine that one system prompt is superior to another. A direct comparison between Falcon and LLaMA cannot be made due to Falcon's shorter allowed context length. While the LLaMA model could accept three examples, Falcon only supports a context length of 2024 tokens. Consequently, Falcon was prompted using a one-shot method. Additionally, since Falcon does not have an official prompting guide, and with unofficial advice from the Falcon team at the Technology Innovation Institute, the developers of Falcon have stating that the model does not follow a specific prompting format [49]. Therefore, a working assumption is made that the same System-User-Assistant format used for LLaMA can also be applied to Falcon. Each system was given the same system prompts, with LLaMA receiving three examples and Falcon receiving one. The fine-tuned versions of Falcon and LLaMA were also assessed.

F. Human Evaluation Survey on Final Three Systems

We designed and ran a survey to qualitatively assess the outputs from the three different systems. This section details the survey construction, data collection process, ranking methodology, and statistical analysis of the results.

1) Participants: A total of 34 participants were recruited to evaluate the outputs of the three systems. Each participant was presented with 20 images, each accompanied by three text outputs generated by the systems and thumbnails of the images are shown in Table II. Participants were informed that the images were AI-generated, and were briefed on the objective of the systems. They were then asked to rank the output for each image from a (best) to 3 (worst). Ethical approval with informed consent from participants, was granted by the School of Computing Ethics Committee.

2) Survey Construction and Data Collection: The survey was administered via Google Forms and was organised into 20 sections, each corresponding to one of the 20 images, plus an informed consent section. Each section contained three outputs related to the image generated by the three systems. Participants were asked to rank the output generated from each of the three systems, in order of preference using a drop down

Table II
THUMBNAILS OF THE 20 IMAGES USED IN OUR USER EVALUATION.

selection to avoid ranking two systems the same. Participants were required to answer all questions, and the underlying systems associated with each output were not disclosed. To mitigate any potential bias, the order of the systems was shuffled within each section. The responses were recorded and mapped back to the respective systems.

3) Statistical Analysis: A statistical analysis was carried out using the SciPy package in python [50]. Given the ordinal nature of the ranking data, with only three possible ranks to chose from, the mode was used to determine the central tendency. The distribution of ranks was also analysed. Non-parametric statistical tests were employed. Specifically, the Friedman test was used to analyse if there is a significant difference in the ranking across all participants for each image, as there are three systems, two degrees of freedom are used [50], [51]. Post-hoc analysis was carried out by performing Wilcoxon signed rank test [50], [52]. As there are three comparisons here it may not be necessary to apply a correction, however the Holm–Bonferroni [53] was applied. Kendall's W [54] was calculated to assess the level of agreement between participants.

V. RESULTS AND DISCUSSION

A. Evaluation of Language Models

The first method, which involved training the model on a single large combined dataset, resulted in the model producing hallucinations and gibberish outputs. Due to the high levels of inaccuracies and hallucinations in outputs, this model was not considered for further evaluation against other fine-tune models, particularly as it performed worse than the base models. In contrast, the models fine-tuned sequentially produced more coherent output, though some inaccuracies and relevance issues persisted. Notably, the example tuning dataset yielded favourable results on both Falcon and LLaMA models using LoRA. The LLMs were evaluated using 10 image captions and 10 questions as inputs into the LLMs, to test the model outputs produced.An example of image caption: "A person lying on a sun lounger by a pool, wearing sunglasses and sunscreen.". Examples of a questions asked: "Do babies like whisky","Is smoking cool". We ranked the model outputs for each question and image cation from a (best) to 4 (worst) taking into consideration the accuracy, relevance and appropriateness of each output.

Table III
RANKED LLMS.

Rank	LLM
1	LoRA-Falcon-7B-Instruct-ExampleTune
2	QLoRA-Falcon-7B-Instruct-Sequential
3	LoRA-Llama3-7B-Instruct-ExampleTune
4	LoRA-Falcon-7B-Instruct-Sequential

B. Evaluation of Vision Models

Both assessors from the project team tended to give all three models a score of 5 for appropriateness, with an average score of 4.84, and each model scoring above 4.75. The LLaVA model achieved the highest overall score across the metrics, with a score of 13.02, followed by Florence with 12.98 and BLIP2 with 12.39. No significant differences were observed between each models' ability to remain on topic in image captioning or for usage of appropriate language. However, further analysis revealed that Florence-2 outperformed BLIP-2 in terms of accuracy. Among the three models, BLIP-2 had the highest propensity for hallucinations; for example, it frequently described objects such as handbags and people that were not present in the image. For example in Figure 1, BLIP-2 states there is a handbag on the ground near the car. In contrast, while Florence-2 did not achieve the same level of performance as LLaVA across the three evaluation criteria, it performed comparably with LLaVA despite having an order of magnitude fewer parameters.

Figure 1. Sample image from the Flickr30 dataset [47]

C. A comparison of different MM-LLM Systems

Through experimentation with Falcon-based MM-LLM systems, using both the one-shot format and fine-tuning on the example tuning dataset as well as a combination of these methods, it was found that the Falcon-7b variant did not perform as effectively as the LLaVA and LLaMA based MM-LLM systems. The Falcon-7b version prompted using the one-shot approach, without fine-tuning, performed best using the 50 example images from the Flickr30 dataset [47]. However, it struggled to maintain topic relevance concerning the images. For instance, when shown an image of a child shown in Figure 2, the system initially provided information about the photo, the child's surroundings, and the importance of play,

but subsequently defaulted to an example related to vaping that was trained on in ICL. It is noteworthy that the vision component of the system did not identify the crucial detail of the image, that the child is upset. While it is an important issue, it is somewhat secondary to the primary concern: that system's tendency to revert to an unrelated example. The image has nothing to do with vaping, yet the system references this unrelated example. This suggests that Falcon-7b-based system cannot generalise from examples to new photographs effectively.

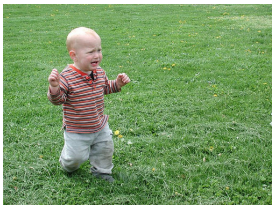

Figure 2. Sample image from the Flickr30 dataset [47]

Given the same photo in Figure 2, the Falcon-7b-instruct-based system appeared to completely misrepresent the image, repeating the system prompt, and fabricating another off-topic example: "You appear to have had a few too many drinks at a party, throwing up in a corner. Friends leaving you alone, embarrassed. Photo courtesy of Pixabay." While providing an example of ICL with the Falcon system that was fine-tuned with the example tuning dataset, slightly improved the response, it remained off-topic and included fabricated details. While Falcon showed some potential for development, it appeared to lack the sophistication required to achieve a general input-to-specific output system. Additionally, the lack of documentation for prompt engineering the model, means educated guesses are needed to be made, based on prompting methods used with other instruct models. This challenge will hopefully be addressed by Falcon-2.

The LLaVA-Mistral-7V-based system produced better results after expanding the instruction prompt. Initially, when provided only with a definition of health literacy, the system described the photo in detail but did not adequately discuss the concept of health literacy. However, after updating the prompt with a larger set of instructions, the output became more focused on healthy living and health literacy. Unlike the other approaches investigated, this method significantly influenced the system's ability to focus on specific aspects of the photo. However, because LLaVA-NeXT can only accept one photo at a time, few-shot prompting could not be employed.

The LLaMA-based systems were capable of accepting multiple examples for few-shot prompting. Similar to the other MM-LLM systems evaluated, the responses of the LLaMA based MM-LLM systems were highly influenced by the prompt. Notably, the responses were improved using the variant fine-tuned on the example tuning dataset compared to that without. However, the fine-tuned system only produced acceptable responses when combined with few-shot prompting. In zero-shot scenarios, where no examples are provided, the fine-tuned LLaMA-based system described the image in some detail but then began including example photo descriptions from its training data. It also hallucinated some information; for example, it separated the on-topic photo description and the numerous examples with "Pass Code: 1234567890". The output did not explain or address this hallucination. However, with few-shot prompting, the system appeared to behave as expected, with a slightly different style of response compared to the non-fine-tuned system. It seems that training the model on the example set has influenced the system. Though, a larger and more diverse set of examples is most likely needed to shape the response in the desired manner.

An investigation into the deliberate over-fitting during LoRA training could yield valuable insights and potentially produce improved results. However, as previous work has indicated, few-shot prompting on the LLaMA3-instruct variant remains the most effective strategy when working with LLaMA. The three systems shortlisted for the user survey are described in Table I.

D. Human Evaluation Survey on MM-LLMs

Table IV
FREQUENCY DISTRIBUTION OF RANKS ASSIGNED TO EACH SYSTEM.

Rank	System_1	System_2	System_3
1	244	209	246
2	262	185	231
3	174	286	203

At the close of the survey, there were 34 respondents who completed assessing and ranking the outputs generated for all 20 images shown in Table II. The mode for each system are System_1 = 2, System_2 = 3 and System_3 = 1. As per Table IV, System_3 was most likely to be ranked best (246 times), with System_3 second at 244 and System_2 Worst at 290. Additionally, System_2 appeared more polarising in its output with a higher count as best and worst system than rank 2.

These findings suggest that the LLaMA3-Florence combination was preferred by participants over the LLaVA system, though the difference in preference between the LLaMA3 versions was marginal. To support this claim, a Friedman statistic of 17.1845 was calculated, with a p-value of 0.0001855, indicating a failure to reject the null hypothesis that there is no significant difference between the rankings of the systems. Post-hoc analysis of the systems, as shown in Tableẽftab:sigs, further supports the claim that there is no significant difference between the rankings of System_1 and 3. However, there was a significant difference between System_1 versus System_2 and System_2 versus System_3. A Kendalls's W value of 0.0234 suggests a low level of agreement among participants, indicating a wide variation in how participants assessed the systems against the images. This support our hypothesis that the manual assessment of which system is superior, is highly subjective.

Table V
WILCOXON SIGNED-RANK TEST WITH HOLM-BONFERRONI CORRECTION.

Comparison	Statistic	p-value	Significance
System_1 vs System_2	91604.5	0.0001	Significant
System_1 vs System_3	104520.5	0.3644	Not Significant
System_2 vs System_3	94900.0	0.0009	Significant

VI. CONCLUSIONS AND RECOMMENDATIONS

LLMs and LVMs have demonstrated remarkable potential, although some results, especially with smaller models, may be over-hyped. Although these models may appear to intuitively understand text and images, they fundamentally predict the most likely next sequence of tokens based on learned representations of image and text. The multistage pipeline's modularity allows for flexibility, such as replacing LLaMA3-8b-Instruct with LLaMA3-70b-Instruct without fine-tuning the projection layer. Models like Florence-2 enable the use of more powerful language models within computational budgets. Although we ran inference on a single GPU, fine-tuning was limited. For handling children's data, sending photos to an API is unfavourable, especially with upcoming EU AI regulations. Fine-tuning on a GPU cluster or in the cloud should be considered if budget permits. However, the systems presented in our research show promise for prototyping before moving to more powerful foundation models.

Our research encountered several limitations, including the absence of real-world health literacy-led responses for fine-tuning. Although the provided health literacy ontology was comprehensive, it lacked accompanying educational material. Additionally, there was no existing dataset of appropriate responses to various images. Despite these challenges, the artificially created example tuning dataset did influence the model responses, and users showed a marginal preference for these responses. This could serve as a baseline for further exploration by gathering a much wider example set with more diverse examples. Future work should focus on expanding this dataset with health education resources, such as the Irish SPHE curriculum or similar materials, to further improve the model's accuracy and context-appropriateness.

Due to context length limitations with Falcon and LLaVA only allowing one photo in context, exploring alternative models with LLaMA-based systems is advisable. Fine-tuning vision models proved computationally expensive, and our experiments were constrained by hardware limitations, particularly the visual encoder requiring 100GB of VRAM.

Given the importance of health information, human oversight is crucial. In the case of PhotoVoice+, a teacher or healthcare professional should monitor the system's outputs to ensure accuracy and appropriateness, especially when used with children. Adult supervision is necessary, as MM-LLM systems, while helpful in educational discussions, cannot fully replace human expertise.

In conclusion, LLMs and LVMs offer significant potential for enhancing health literacy education through conversational AI. Practical deployment requires careful consideration of computational resources, regulatory constraints, and ensuring information accuracy. Future work should address these challenges to ensure effective and secure implementation. Our research lays a strong foundation for further advancements in this field.

REFERENCES

[1] Kilian Carolan, Laura Fennelly, and Alan F. Smeaton. A Review of Multi-Modal Large Language and Vision Models. *arXiv preprint arXiv:2404.01322*, 2024.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. Technical report, OpenAI, 2018.

[5] Krissy Davis. The Best Large Language Models on The Market — wearedevelopers.com. https://www.wearedevelopers.com/magazine/best-large-language-models#toc-7, 2023. [Accessed 22-Feb-2024].

[6] ODSC - Open Data Science. 2023 Was the Year of Large Language Models: Then and Now — odsc.medium.com. https://odsc.medium.com/2023-was-the-year-of-large-language-models-then-and-now-924d34f3b6a9, 2023. [Accessed 22-Feb-2024].

[7] Ari Chanen. What I learned from Bloomberg's experience of building their own LLM — linkedin.com. https://www.linkedin.com/pulse/what-i-learned-from-bloombergs-experience-building-own-chanen-phd/, 2023. [Accessed 21-Feb-2024].

[8] Nikita Khudov. The Future of LLMs: Proprietary versus Open-Source — linkedin.com. https://www.linkedin.com/pulse/future-llms-proprietary-versus-open-source-nikita-khudov/, 2023. [Accessed 22-Feb-2024].

[9] Sara Guaglione. The case for and against open-source large language models for use in newsrooms — digiday.com. https://digiday.com/media/the-case-for-and-against-open-source-large-language-models-for-use-in-newsrooms/, 2023. [Accessed 22-Feb-2024].

[10] IBM Data and AI Team. Open source large language models: Benefits, risks and types - IBM Blog — ibm.com. https://www.ibm.com/blog/open-source-large-language-models-benefits-risks-and-types/. [Accessed 22-Feb-2024].

[11] Indumathi Pandiyan. Open Source or Proprietary LLMs — indukishen. https://medium.com/@indukishen/open-source-or-proprietary-llms-fbaa8dae2b6d, 2023. [Accessed 21-Feb-2024].

[12] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023.

[15] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[16] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[17] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA: Large Language and Vision Assistant. https://github.com/haotian-liu/LLaVA, 2023. Accessed: 2024-07-21.

[18] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

[19] Andrew Johnson. Understanding Rotary Position Embedding: A Key Concept in Transformer Models. https://https://medium.com/@andrew_johnson_4/understanding-rotary-position-embedding-a-key-concept-in-transformer-models-5275c6bda6d0, 7 2023. [Accessed 05-Mar-2024].

[20] Jamie R. Robinson Harrison C. Lucas, Jeffrey S. Upperman. A systematic review of large language models and their implications in medical education. https://asmepublications.onlinelibrary.wiley.com/doi/10.1111/medu.15402. [Accessed 23-07-2024].

[21] Enhancing Health Care Communication With Large Language Models-The Role, Challenges, and Future Directions - PubMed — pubmed.ncbi.nlm.nih.gov. https://pubmed.ncbi.nlm.nih.gov/38466311/. [Accessed 23-07-2024].

[22] Najeeb Nabwani. Full Fine-Tuning, PEFT, Prompt Engineering, or RAG? — deci.ai. https://deci.ai/blog/fine-tuning-peft-prompt-engineering-and-rag-which-one-is-right-for-you, 2023. [Accessed 17-Feb-2024].

[23] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. *arXiv preprint arXiv:2312.12148*, 2023.

[24] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021.

[25] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs, 2023.

[26] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

[27] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[28] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[29] Andrew Maynard. Prompt and response evaluation. https://andrewmaynard.net/prompt-and-response-evaluation/, 2024. Accessed: 2024-05-14.

[30] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[31] Ying Sheng Wei-Lin Chiang, Lianmin Zheng. Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings — LMSYS Org — lmsys.org. https://lmsys.org/blog/2023-05-03-arena/. [Accessed 23-07-2024].

[32] Timothy R McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv preprint arXiv:2402.09880*, 2024.

[33] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The Falcon Series of Open Language Models. *arXiv preprint arXiv:2311.16867*, 2023.

[34] Leandro von Werra, Younes Belkada, Sourab Mangrulkar, Lewis Tunstall, Olivier Dehaene, Pedro Cuenca, Philipp Schmid, and Omar Sanseviero. The Falcon has landed in the Hugging Face ecosystem — huggingface.co. https://huggingface.co/blog/falcon, 2023. [Accessed 23-Feb-2024].

[35] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. *arXiv preprint arXiv:2306.01116*, 2023.

[36] Minhajul Hoque. Exploring the Falcon LLM: The New King of The Jungle — minh.hoque. https://medium.com/@minh.hoque/exploring-the-falcon-llm-the-new-king-of-the-jungle-5c6a15b91159, 2023. [Accessed 23-Feb-2024].

[37] AI@Meta. Llama 3 model card. 2024.

[38] Wei Huang, Xudong Ma, Haotong Qin, Xingyu Zheng, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. How good are low-bit quantized llama3 models? an empirical study. *arXiv preprint arXiv:2404.14047*, 2024.

[39] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024.

[40] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

[41] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[42] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* ChatGPT quality, March 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 3(5), 2023.

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[44] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, January 2024.

[45] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[46] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[47] Aditya Jain. Flickr30k, July 2023.

[48] Weijieying Ren, Xinlong Li, Lei Wang, Tianxiang Zhao, and Wei Qin. Analyzing and reducing catastrophic forgetting in parameter efficient tuning. *arXiv preprint arXiv:2402.18865*, 2024.

[49] Hugging Face. Discussion on falcon 7b instruct model. https://huggingface.co/tiiuae/falcon-7b-instruct/discussions/1, 2023. Accessed: 2024-07-19.

[50] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[51] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

[52] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer, 1992.

[53] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979.

[54] Andy Field. Kendall's coefficient of concordance. *Encyclopedia of statistics in behavioral science*, 10 2005.