Images in Language Space: Exploring the Suitability of Large Language Models for Vision & Language Tasks

Sherzod Hakimov[1] and **David Schlangen**[1,2]
[1]Computational Linguistics, Department of Linguistics
University of Potsdam, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
firstname.lastname@uni-potsdam.de

Abstract

Large language models have demonstrated robust performance on various language tasks using zero-shot or few-shot learning paradigms. While being actively researched, multimodal models that can additionally handle images as input have yet to catch up in size and generality with language-only models. In this work, we ask whether language-only models can be utilised for tasks that require visual input – but also, as we argue, often require a strong reasoning component. Similar to some recent related work, we make visual information accessible to the language model using separate verbalisation models. Specifically, we investigate the performance of open-source, open-access language models against GPT-3 on five vision-language tasks when given textually-encoded visual information. Our results suggest that language models are effective for solving vision-language tasks even with limited samples. This approach also enhances the interpretability of a model's output by providing a means of tracing the output back through the verbalised image content.

1 Introduction

In recent years, large language models have gained significant attention in the natural language processing (NLP) community due to their impressive performance on various tasks such as machine translation, text generation, and language modelling (Vaswani et al., 2017; Devlin et al., 2019). These models, which are trained on massive amounts of data, have been shown to capture complex linguistic patterns and generate coherent text (Brown et al., 2020). Some of the most popular models are trained by OpenAI, a research organization that has released several models, including GPT (Radford et al., 2018), GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020). In addition to GPT models, there are also many open-source or open-access large language models that researchers and organizations around the world have developed, such as BLOOM (Scao et al., 2022), GPT-J (Wang and Komatsuzaki, 2021), OPT (Zhang et al., 2022), Flan-T5 (Chung et al., 2022).

Recent work by Liang et al. (2022) provided an in-depth analysis of many large language models (LLM) across 42 core scenarios.[1] All scenarios are language tasks that are evaluated with multiple metrics by prompting the language models with few-shots from the selected datasets, also known as *in-context learning*. Currently, there are no comparable models directly suitable for tasks that require visual information as part of the context, even though such multimodal tasks have similar practical relevance.

Pre-trained vision-language models (Li et al., 2019; Chen et al., 2020; Dosovitskiy et al., 2021; Radford et al., 2021) have shown great promise by learning joint representation of images and text documents, but so far they have not been optimised for prompting on vision-language tasks but rather using the learned joint representations for fine-tuning on downstream tasks. Moreover, as we note below, many multimodal tasks appear to rely on reasoning capabilities, which larger language models have been shown to perform well on (Talmor et al., 2020; Li et al., 2022b). Hence, in this work, we attempt to utilise such models to do in-context learning on multimodal data, achieving this by encoding the visual information in language.

While there has been some recent work going in this direction (see discussion below), it falls short in terms of evaluating the performance of large language models across multiple dimensions, applying them to a diverse range of vision-language tasks, and comparing the performance of GPT models with open-source or open-access models. As such, there is a need for further research in this area to fully understand the capabilities and limitations of these models for vision-language tasks.

[1]https://crfm.stanford.edu/helm/v0.2.0/

Findings of the Association for Computational Linguistics: ACL 2023, pages 14196–14210

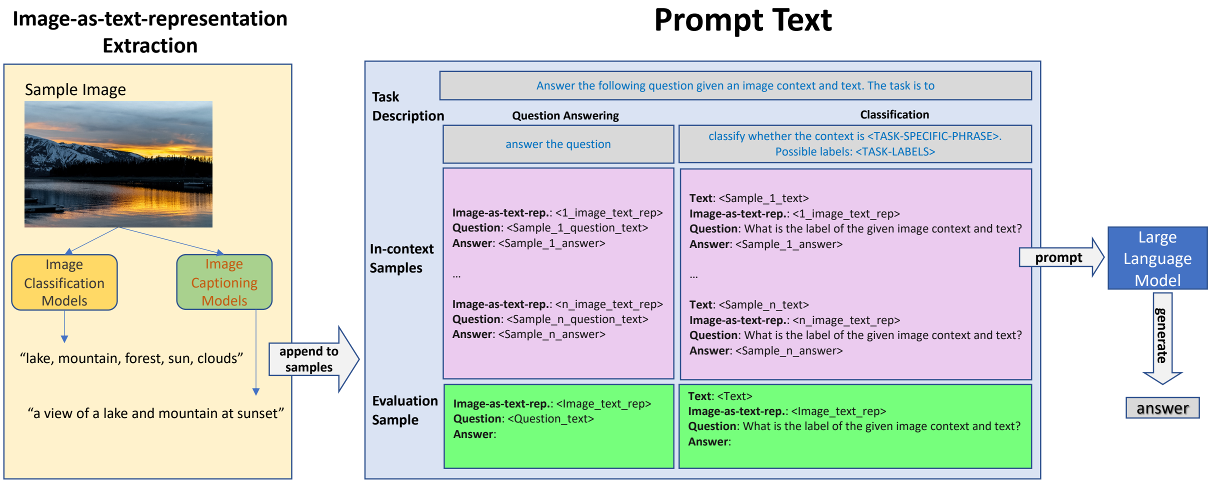

Figure 1: Model architecture for in-context learning for vision-language tasks. Each sample image is converted into its image-as-text-representation by applying pre-trained image captioning and classification models (yellow). The prompt text that is fed into a large language model consists of a task-specific description (blue), in-context samples (pink), and the evaluation sample (green). The language model is expected to generate a text sequence that follows the word *Answer* in the evaluation sample.

In this paper, we aim to explore the capabilities of large language models for in-context learning and their potential to improve performance on multimodal tasks (see Figure 1). To this end, we conducted a series of experiments to evaluate the performance of various language models (closed, open-source and open-access) on tasks that involve multiple modalities, such as vision and language. The tasks vary from identification of hate speech and sentiment to visual reasoning and question answering. Our results provide insights into the strengths and limitations of these models and highlight methods of expressing visual content through textual descriptions. Our work aims to analyse how much general reasoning is in language models by evaluating them on multimodal tasks where the visual content is only accessible partially or indirectly (since the visual content is verbalised and represented in textual form) accessible. Our main contributions are as follows[2]:

- We examine the impact of in-context learning with large language models on five vision-language tasks, four classification and one question answering tasks;
- We investigate the impact of the textual description generation for the visual content on the model performance for the respective tasks;
- We compare the performance of open-source and open-access models with GPT-3 on the selected vision-language tasks.

[2]Source code and all resources are made publicly available at https://github.com/clp-research/language-models-multimodal-tasks

2 Related Work

Recent work by Liang et al. (2022) provided an in-depth analysis of many 34 large language models (LLM), open, limited-access, and closed. Their analysis revealed the capabilities and limitations of these models across 42 core scenarios. All scenarios are language tasks that are evaluated with 57 metrics by prompting the language models with a few shots from the selected datasets. Such a way of leveraging pre-trained language models for downstream tasks is known as *in-context learning*, where a certain task description and a few shots are presented as a context for the model. A recent survey by Dong et al. (2023) describes the developed techniques for in-context learning where they present a taxonomy that divides the techniques used for prompting such as selection of in-context samples, reasoning step by step (chain of thought) (Wei et al., 2022b), task definition, etc. Moreover, Min et al. (2022) assessed the importance of choosing the in-context samples and its effect on the performance.

So far, a large-scale analysis of large language models and their performances for multimodal data, such as vision-language tasks, has not been done. A handful of methods demonstrated the effectiveness of in-context learning for multimodal tasks. Zhou et al. (2022b,a) modelled the context words in prompts for applying pre-trained vision-language tasks for downstream vision tasks. Tsimpoukelli et al. (2021) trained a vision encoder to represent images as a sequence of continuous embeddings where a prompted pre-trained language model generates a caption. Yang et al. (2022) demonstrated the applicability of GPT-3 on a visual question answering task where they converted the images into textual descriptions by using an image captioning model and extraction of visual tags that correspond to detected objects, landmarks, person, image type, etc. Zeng et al. (2022) follows similar methodology by showing applications on multiple applications that include modalities such as audio, video beside image and text. Gui et al. (2022)'s method is complementary to the previous method with an addition of a contrastive learning module that retrieves knowledge entries from Wikidata knowledge graph (Vrandecic and Krötzsch, 2014). Wang et al. (2022b) applied the method of converting images into textual descriptions to video tasks. The resulting outputs are temporally aligned for a video and then fed into GPT-3 with few shots. More recently, Merullo et al. (2023) aligned image-text encoders by training a linear project layer and keeping the pre-trained image and text encoders frozen. Our paper presents a study that goes beyond these similar approaches by extending the experimental evaluation to multiple datasets, comparing open-source language models with GPT-3, and evaluating different methods of acquiring textual representation for the visual content.

3 Text-Visual In-Context Learning

In this section, we describe the proposed methodology of applying *in-context learning* to vision-language tasks. In-context learning essentially works by prompting a pre-trained language model with the task and expecting it to generate text that solves a particular task. It is performed by giving a few-shots of the respective task at inference time without requiring updating the model weights and expecting the model to generate text corresponding to the expected output.

Formally, given a query input text $\mathbf{x}$ and a set of candidate answers $Y = \{\mathbf{y_1} ... \mathbf{y_m}\}$, which can be class labels for a particular task or free text, a pre-trained model M outputs a candidate answer with the maximum score conditioned on the task description T, n in-context sample pairs $C = \{(\mathbf{x}_1, \mathbf{y}_1)...(\mathbf{x}_n, \mathbf{y}_n)\}$. The likelihood of the candidate answer $\mathbf{y_j}$ can be represented by a scoring function f with the language model M (Wei et al., 2022a; Dong et al., 2023):

$$\mathrm{P}(\mathbf{y_j}|\mathbf{x}) \triangleq f_M(\mathbf{y_j}, T, C, \mathbf{x}) \quad (1)$$

The final predicted candidate answer of the model ($\hat{\mathbf{y}}$) can be formulated as:

$$\hat{\mathbf{y}} = \mathrm{argmax}_{\mathbf{y_j} \in Y} \mathrm{P}(\mathbf{y_j}|\mathbf{x}) \quad (2)$$

Our proposed methodology for "text-visual in-context learning" is shown in Figure 1. First, all images from all evaluated datasets have been passed through multiple pre-trained image models to obtain the textual description of the visual content, which we refer to as *image-as-text-representation* throughout the paper. The image-as-text-representation is essentially a textual description of the visual content that captures important visual aspects. The prompt text comprises the task description, in-context sample pairs, and the input of the evaluation sample. Given such a prompt text, the language model generates a sequence of text tokens as an output.

We evaluate the proposed methodology on various vision-language datasets that include either classification or question answering tasks. Thus, the task description is different between these two categories. The task description for the classification tasks is further replaced with the task-specific phrase that describes the downstream task and provides the task-specific class labels. More details on the exact prompt text for each dataset are provided in Appendix A. Next, we describe the methods for extracting image-as-text-representation, selecting in-context samples, and aggregating answer predictions in cases where the language model is prompted multiple times with various in-context samples for the evaluation sample.

3.1 Image-as-Text-Representation Extraction

We use two different methods to extract textual representation of images for any vision-language task. The first is to use pre-trained image captioning models that generate a text sequence describing the input image. The second is to employ multiple pre-trained image classification models and extract top-scoring class labels. The extracted class labels from all models are merged to form the set of *visual tags* that describe the image. Specifically, we use pre-trained models to recognise objects, indoor or outdoor scenes, and persons and their facial emotions. These methods yield a different textual description of an input image, which is used as *image-as-text-representation* in the prompt text.

3.2 In-Context Sample Selection

The selection of samples for in-context learning is an essential step for prompting large language models. Each model has its own limitation of the maximum input tokens that a prompt text can have (e.g. 512 tokens for Flan-T5, 4000 for GPT-3). Therefore, only a few samples can be used (few-shot), and the selection directly impacts the model performance (Yang et al., 2022; Min et al., 2022). We experiment with the following sample selection methods.

Random sample selection works by selecting any random n samples from the training split of a dataset (that fit into the maximum input token length of a language model).

Adaptive sample selection uses specific similarity measurement to rank samples with respect to the input sample. Top-ranking n samples are selected (that fit into the maximum input token length of a language model) to prompt a language model.

The in-context samples are selected from the training split of the respective dataset.

4 Experimental Setup

In this section, we describe the details of the building blocks of the methodology, the evaluated datasets, large language models, and methods for obtaining a textual description of images.

4.1 Datasets, Comparison Models & Evaluation Metrics

We use the following five datasets to evaluate the performance of the closed and open-access language models. The best-performing prior models are selected from the leaderboards of the respective datasets. These models are used for comparison with our prompting method.

- *MAMI - Multimedia Automatic Misogyny Identification* (SemEval 2022 Task 5) (Fersini et al., 2022): the dataset consists of *memes* that are classified for being offensive hateful towards women. The train and test splits have 10 000 and 1000 samples, respectively. We use the sub-task A for the experiments to predict a binary class indicating whether the given meme is misogynous.
  Comparison model: Zhang and Wang (2022) proposed to use an ensemble system composed of pre-trained models (CLIP) used for extracting features from multimodal data.
  Evaluation metric: Macro-F1

- *HF - Hateful Memes* (Kiela et al., 2020): is another dataset that focuses on classifying *memes* whether the overall message it is hateful or not towards any group. We use the provided development split for the evaluation since the test split is closed to the community at the time of writing. The train and development splits have 8500 and 500 samples, respectively.
  Comparison model: the best-performing model provided is by Kiela et al. (2020), for which the model performance on the development split is available. The method uses pre-trained ViLBERT model that is later fine-tuned on the dataset.
  Evaluation metric: Accuracy

- *MVSA - Multi-View Sentiment Analysis* (Niu et al., 2016): is a multimodal sentiment analysis dataset collected from Twitter. The task is to classify the sentiment of the given post with an image and tweet text into *positive*, *negative*, or *neutral*.

  Previous work on this dataset has used different train and test splits, making the direct comparison among approaches not feasible. We follow recently provided splits by Cheema et al. (2021) and their evaluation scheme by performing 10-fold cross-validation on the respective train and test splits. Overall, the dataset includes total 3928 samples with 2328, 1167, 433 samples corresponding to *positive*, *negative*, and *neutral* class labels, respectively. We use the version named *MVSA-Single* of this dataset.
  Comparison model: Cheema et al. (2021)'s model uses image features from CLIP and text features from RoBERTa models and fine-tune them on the dataset.
  Evaluation metric: Accuracy averaged over 10-folds.

- *OK-VQA - Outside Knowledge Visual Question Answering* (Marino et al., 2019): is a visual question answering dataset consisting of 14 055 samples where each sample contains an open-ended question and five ground truth answers. The task is to predict one of the expected answers given the question and the image. There are 9009 and 5046 samples in the train and test splits of the dataset, respectively.

Comparison model: Wu et al. (2022)'s method is based on three-stage scheme where the first step generates a set of answer candidates by analysing the syntactic structure of the question. The next step retrieves candidate answers by searching the Wikipedia and ConceptNet, and finally the third step validates the candidate answers.

Evaluation metric: Accuracy

- *NLVR2 - Natural Language for Visual Reasoning for Real* (Suhr et al., 2019): is a dataset for reasoning over two images and a statement where the task is predict whether the statement is *true* or *false*. The dataset includes 86 373 and 6967 samples for the train and test splits, respectively. We used the *test-public* split of the dataset.

  Comparison model: Chen et al. (2020)'s approach is based on first pre-training a joint multimodal model on image captioning datasets and then fine-tune the model on the task.

  Evaluation metric: Accuracy

4.2 Language Models

We experiment with multiple pre-trained open-source and open-access language models and compare them against GPT-3. These language models are as follows:

- *Flan-T5* (Chung et al., 2022): is a language model fine-tuned on multiple tasks with an instruction-specific training paradigm. We use the *flan-t5-xxl* version.
- *T0pp* (Sanh et al., 2022): is a language model that has been fine-tuned on multiple datasets to perform for zero or few-shot prompting.
- *OPT* (Zhang et al., 2022): is a language model trained on multiple large datasets. The language model has various versions with different sizes. We use the *opt-2.7b* version.
- *GPT-3*: we use the *text-davinci-003* version.

4.3 Methods for Extracting Image-as-Text-Representations

The generation of the textual representation of images is carried out in two ways: image captioning and the combination of multiple image classification model outputs.

Image Captioning: we use the following image captioning models to convert the images to textual descriptions:

- ViT-GPT-2 (Vision Transformers GPT-2) (NLP Connect, 2022)
- OFA (One for all) (Wang et al., 2022a)
- BLIP (Bootstrapping Language-Image Pre-training) (Li et al., 2022a)

Visual Tags: we use the following image classification models to build the set of tags extracted from a given image:

- *Image type*: a zero-shot classification with CLIP (Radford et al., 2021) by pairing an image with one of the following text snippets and selecting the one that outputs the highest probability: "This is an image", "This is a sketch", "This is a cartoon", "This is a painting". We select the top-ranking class label that has a probability higher or equal to 0.80.
- *Object*: the pre-trained Detection Transformer (DETR) model (Carion et al., 2020) is used to obtain the bounding boxes of detected objects. We select the top-ranking class labels that have a probability higher or equal to 0.90.
- *Indoor and outdoor scenes*: we use two different pre-trained models to predict the scenes in the given images. The first model is Vision Transformer (ViT) (Wu et al., 2020) pre-trained on Indoor Scene dataset (Quattoni and Torralba, 2009). The second model is a pre-trained ResNet-50 on Places365 dataset (Zhou et al., 2018). We select the top-ranking class labels that have a probability higher or equal to 0.80.
- *Facial expression*: we use the pre-trained MTCNN model (Zhang et al., 2016) to detect faces in images and identify seven facial emotions: angry, disgust, fear, happy, sad, surprise, neutral. (Goodfellow et al., 2015). We select the top-ranking detected faces (probability >= 0.90) and use them to infer the facial expression classes. The top-ranking facial expression class label (probability >= 0.50) is selected for each detected face.

4.4 Prompt Structure

Similarity measurement: As mentioned above in Section 3.2, we employ two different methods for selecting samples for in-context learning: random and adaptive. In order to select the best fitting *n* samples for the adaptive prompting, we use the Sentence Transformers (Reimers and Gurevych, 2019) to calculate the similarities among samples for the adaptive method. The pre-trained *all-mpnet-base-v2* model is used to extract embeddings from two given sample documents and calculates the cosine similarity between them[3]. For any given two samples (one evaluation and the other one from a training split), we calculate the similarity between the text content and image-as-text representation obtained from the methods described before. The similarities from textual content and image-as-text-representation are averaged.

Sample selection: Once the most similar samples to the given evaluation sample are identified, the next step is to select *n* samples out of them. During selection, we ensure that the selected samples are equally distributed across the class labels for any dataset. This only applies to the classification tasks where the labels are predefined (e.g. hateful or not, true/false, positive/negative/neutral). It is to present samples with different labels for in-context learning. We also experiment with a zero-shot (*n*=0) where the prompted text includes only the task description.

The prompt structure for each dataset is available in Appendix A.

4.5 Model Parameters & Implementation

We experimented with various configurations of the model parameters. The following values are used for all language models: *max new tokens* set as 10, *number of beams* is set as 10, *temperature* is set to the default value of each language model.

The implementation of the overall architecture and other building blocks (image captioning & classification) is based on the PyTorch library. We used the language models that are available in the HuggingFace directory and queried the backend API of OpenAI for experiments with GPT-3. All experiments have been carried out on two NVIDIA A100 GPUs (80 GB). The estimated runtime of all experiments is approximately 200 hours.

5 Results and Analysis

In this section, we discuss the obtained results from the experiments such as the impact of in-context sample selection, image-as-text representation methods and comparing with fine-tuned vision-language models on the selected datasets.

[3]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

		Number of Samples				
Dataset	**Models**	**n=0**	**n=1**	**n=2**	**n=3**	**S**
MVSA (Acc)	**Flan-T5**	6.5	22.9	28.4	29.1	*r*
	Flan-T5	6.5	21.4	31.8	35.2	*a*
	T0pp	**68.1**	54.2	57.7	45.0	*r*
	T0pp	68.1	57.4	62.3	51.8	*a*
	OPT	0.0	12.9	11.8	14.4	*r*
	OPT	0.0	11.5	11.1	11.3	*a*
OK-VQA (Acc)	**Flan-T5**	27.0	28.1	28.7	29.5	*r*
	Flan-T5	27.0	32.4	34.4	**35.4**	*a*
	T0pp	13.9	18.4	18.4	18.2	*r*
	T0pp	13.9	19.4	20.2	20.3	*a*
	OPT	0.9	5.7	3.6	4.3	*r*
	OPT	0.9	10.0	3.9	3.3	*a*
NLVR2 (Acc)	**Flan-T5**	0.0	12.0	19.5	19.7	*r*
	Flan-T5	0.0	25.6	31.7	31.7	*a*
	T0pp	58.6	57.5	49.2	49.5	*r*
	T0pp	**58.6**	55.2	50.7	50.1	*a*
	OPT	42.2	47.7	46.3	45.3	*r*
	OPT	42.2	26.9	10.8	8.1	*a*
HF (Acc)	**Flan-T5**	55.2	53.2	53.6	53.8	*r*
	Flan-T5	55.2	**57.4**	56.6	56.0	*a*
	T0pp	50.0	48.8	1.6	0.0	*r*
	T0pp	50.0	53.6	49.0	33.0	*a*
	OPT	0.2	44.4	41.4	36.0	*r*
	OPT	0.2	35.2	38.2	30.6	*a*
MAMI (F1)	**Flan-T5**	41.5	51.7	37.7	56.1	*r*
	Flan-T5	41.5	61.9	**64.4**	64.1	*a*
	T0pp	26.0	22.2	6.8	0.0	*r*
	T0pp	26.0	46.0	33.8	21.8	*a*
	OPT	17.0	22.2	22.1	21.9	*r*
	OPT	17.0	22.3	22.0	22.4	*a*

Table 1: Ablation study on the number of in-context samples and the method of selecting them. *n* refers to the number of samples in a given prompt, **S** stands for the selection method: adaptive (*a*) or random (*r*). All runs are based on using captioning from BLIP model. The best result for each dataset is highlighted in bold.

5.1 Impact of In-Context Sample Selection

Sample Selection: we have conducted experiments with different configurations of selecting in-context samples. The results are presented in Table 1. In four out of five datasets, the *adaptive* sample selection yields better performance than *random* method. Only in the MVSA dataset, *random* method yields the best result.

Number of Samples: The presented results for each evaluated language model include the different number of samples in a prompt. We tested numbers 0, 1, 2, 3 where *n=0* essentially means that there are no in-context samples in a prompt, and it is **zero-shot performance** of an evaluated language model. It is a few-shot setting in cases where *n* is bigger than zero. We can observe that

Dataset	Models	Image-as-text Representation			
		BLIP	VG	OFA	VT
MVSA (Acc)	**Flan-T5**	31.8	21.6	27.7	16.1
	T0pp	62.3	61.8	62.4	**63.1**
	OPT	11.1	11.0	19.5	12.7
OK-VQA (Acc)	**Flan-T5**	**34.4**	32.6	31.1	29.2
	T0pp	20.2	19.7	18.3	17.8
	OPT	3.9	4.0	19.5	14.8
NLVR2 (Acc)	**Flan-T5**	31.7	25.6	25.5	23.4
	T0pp	50.7	49.4	50.1	**51.0**
	OPT	10.8	19.3	29.6	3.1
HF (Acc)	**Flan-T5**	56.6	54.8	54.6	**56.8**
	T0pp	49.0	49.2	49.6	48.8
	OPT	38.2	38.4	43.4	42.0
MAMI (F1)	**Flan-T5**	**64.4**	48.6	60.2	60.3
	T0pp	33.8	33.6	33.6	22.6
	OPT	22.0	22.1	22.2	22.2

Table 2: Ablation study on the affect of using different image-as-text representation methods. All runs of each model has been set to *adaptive* sample selection with *n*=2 (number of in-context samples in a prompt). VG: ViT-GPT2, VT: Visual Tags

in three out of five datasets, using $n > 1$ yields better performance, whereas *T0pp* achieves the best performance in *MVSA* and *NLVR2* datasets.

5.2 Evaluation of Image-as-Text Representation Methods

As explained in Section 4.3, we have used four methods of verbalising the visual content and adding the output to the prompt as an image-as-text representation. We have tested these methods on all datasets. The results are presented in Table 2. Based on the outcomes in Section 5.1, we have used *adaptive* sample selection with *n=2* for all runs. We can observe that in the majority of the evaluated datasets, using captions generated by *BLIP* model yields the higher performance on average. The textual descriptions generated by the method *Visual Tags* (collection of multiple image classification high-probability outputs) resulted in the highest performance on three datasets.

5.3 Comparison of Language Models

In Table 3, we have selected the best-ranking configuration of each model for all datasets. All model configurations use image captions generated by the BLIP model to represent the image context in text. To reduce the budget, we ran GPT-3 experiments only on a pre-selected set of parameters (n=2, adaptive) that yielded the best results using open-source language models. The overall comparison of all results shows that GPT-3 achieves the best result

Dataset	Models	Result
MVSA (Acc)	Flan-T5, n=3, adaptive	35.2
	GPT-3, n=2, adaptive	64.3
	OPT, n=3, random	14.4
	T0pp, n=0, adaptive	**68.1**
	Fine-tuned V&L Model Cheema et al. (2021)	***75.3***
OK-VQA (Acc)	Flan-T5, n=3, adaptive	**35.4**
	GPT-3, n=2, adaptive	25.9
	OPT, n=1, adaptive	10.0
	T0pp, n=3, adaptive	20.3
	Fine-tuned V&L Model Wu et al. (2022)	***41.4***
NLVR2 (Acc)	Flan-T5, n=2, adaptive	31.7
	GPT-3, n=2, adaptive	**63.0**
	OPT, n=1, random	47.7
	T0pp, n=1, adaptive	58.6
	Fine-tuned V&L Model Chen et al. (2020)	***79.5***
HF (Acc)	Flan-T5, n=2, adaptive	57.4
	GPT-3, n=2, adaptive	**58.8**
	OPT, n=1, random	44.4
	T0pp, n=1, adaptive	53.6
	Fine-tuned V&L Model Kiela et al. (2020)	***66.1***
MAMI (F1)	Flan-T5, n=2, adaptive	64.4
	GPT-3, n=2, adaptive	**69.2**
	OPT, n=3, adaptive	22.4
	T0pp, n=2, adaptive	46.0
	Fine-tuned V&L Model Zhang and Wang (2022)	***83.4***

Table 3: Overall comparison of the best-ranking configurations for each model. The best result for each dataset using prompting with language models is highlighted in bold. All model configurations use image captions generated by BLIP model. V&L: Vision-Language

on three datasets: *MAMI*, *HF*, and *NLVR2*. T0pp achieves the best result on *MVSA* dataset, whereas the best-ranking model for the *OK-VQA* is Flan-T5.

We have also included the results from the fine-tuned vision-language models for each dataset. By comparing the results obtained via prompting with fine-tuned models, with only a few-shots (n = 1, 2, 3), the language models can generalise to vision-language tasks and achieve comparable results. An important observation is that these models were trained only on text documents. Prompting these models on five downstream vision-language tasks by converting the visual content into textual representation made it possible.

5.4 Qualitative Analysis

We present qualitative examples from each dataset in Figure 2. Each sample includes the image-as-text representation extracted from the BLIP model. We also included the ground truth for each sample and the responses generated from Flan-T5 and GPT-3 models (best configurations as in Table 3). We also added the Visual Tags for each sample (combina-

Figure 2: Qualitative examples for each evaluated dataset. The samples include the ground truth and responses generated via prompting Flan-T5 and GPT-3 models. Samples for the MAMI and HF datasets are prompted including the overlay text embedded in an image, which is excluded in the graphic for spacing reasons.

tions of multiple image classification predictions) to show the the comparison against captions generated by the BLIP model.

5.5 Discussion

We have presented experimental results that prompted large language models for five vision-language tasks. The prompting was made possible by representing the visual content in any task using methods such as image captioning or applying multiple image classification methods and combining their outputted high-ranking class labels. We have shown that such a method can achieve impressive results by presenting only two or three samples from a respective dataset compared with fine-tuned models on the entire train splits. It is worth mentioning that the gap between prompted models and fine-tuned models (in some evaluated datasets) is still there (margins of 10-20 points). One way of closing the such gap is by making the image-as-text representation methods achieve performance closer to how humans verbalise the visual content. Our paper essentially aims to highlight that given such image-as-text representations, which are only partial representations given the image models' capabilities, whether language models can be used for multimodal tasks by relying on their (imperfect) general reasoning mechanisms such as chain-of-thought (Wei et al., 2022b). Another way to achieve better performance (and close the gap with task-specific fine-tuned models) is to train vision-language models that are capable of in-context learning via prompting (Alayrac et al., 2022).

We have also shown that the choice of in-context samples impacts the results. Using samples similar to the evaluated one (adaptive method) yields better performance than choosing them randomly.

Another critical observation to mention here is that different language models obtained the best results on various tasks. Overall, *GPT-3* is the best-ranking model for three datasets. Among open-source models, *T0pp* and *Flan-T5* obtained the highest overall performance. Even though their performance was not the highest for many tasks, it is still possible to achieve comparable results or even the best ones in some cases. For the *MVSA* dataset, *T0pp* achieved the best performance even in a zero-shot setting. Thus, the language model's choice makes a difference in applying such models for any downstream tasks.

6 Conclusion

In conclusion, our study has demonstrated the suitability and effectiveness of using large language models via prompting on vision-language tasks.

Our approach relies on verbalising the visual content employing image captioning and classification models and prompts the language models along with the textual content. We have also shown that the choice of in-context samples and the method of verbalising the visual content impact the results. Our experimental evaluation suggests that this approach can achieve impressive results by presenting only a few samples from a dataset compared to models that are fine-tuned on entire train splits of the evaluated datasets. Furthermore, our study has also highlighted the importance of considering the choice of language models when applying them to such downstream tasks. We have demonstrated that different models perform better on various tasks, with GPT-3 achieving the highest overall performance across three datasets and open-source models T0pp and Flan-T5 achieving the best overall performance among them. Even though the performance of these models may not have been the best across all evaluated tasks, they still have the potential to be used in such cases and even achieve comparable results. For instance, T0pp yielded the best performance on the MVSA dataset, even in a zero-shot setting. Thus, the choice of language models is crucial for achieving optimal results in vision-language tasks.

Limitations

Limitations on the evaluated language models and obtained results: The presented model architecture utilises various pre-trained language or image models. The main limitation of the experimental evaluation is not using other language models. Due to the limited budget and processing power, we have included the language models that have been shown to perform better based on the previous work. Another limitation is that we excluded language models that exceeded the 80 GB memory of an NVIDIA A100 GPU. Our experiments led to different results for the GPT-3 compared to Yang et al. (2022). It can be explained by using different methods for converting images to textual representations and slightly varying prompting structures.

Limitations on the used image models: The limitation concerning the pre-trained image models is that we selected a handful of methods based on their success for related tasks. Including other pre-trained models would increase the parameter space and thus increase the budget for the study.

Limitations on the selected datasets: All datasets are multimodal tasks where the underlying text is only in English. The choice of the dataset is related to the fact that there are limited multimodal datasets in other languages. The evaluation metric for the OK-VQA dataset requires the output to match exactly one of the expected answers. It counts as a wrong answer even if a slight change in the answer or another paraphrase is given as an output, e.g. “race” vs “racing”. We applied the same evaluation criterion and left this improvement as future work.

Ethics Statement

There might arise ethical issues as part of this work. The used pre-trained language models inherit particular biases as part of their learning process, which might affect the generated outputs. Another concern is the use of pre-trained image models for captioning or classification. The generated outputs from these models might predict certain visual concepts and thus leading to inaccurate text descriptions for the given images are generated. Another concern directly concerns using large language models as few-shot models. Such models have demonstrated high performance for many downstream tasks. However, the interpretation of the model predictions is still ongoing research.

Acknowledgements

We want to thank the anonymous reviewers for their comments. This work was supported by BMBF (German Federal Ministry of Research), project “COCOBOTS” (01IS21102A) and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 423217434 (“RECOLAGE”) grant.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *CoRR*, abs/2204.14198.

Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33:*

Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer.

Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. 2021. A fair and comprehensive comparison of multimodal tweet sentiment analysis methods. In *MMPT@ICMR2021: Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding, Taipei, Taiwan, August 21, 2021*, pages 37–45. ACM.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, and et al. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey for in-context learning. *CoRR*, abs/2301.00234.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2015. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968, Seattle, United States. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022a. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022b. How pre-trained language models capture factual knowledge? a causal-inspired analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1720–1732, Dublin, Ireland. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, and et al. 2022. Holistic evaluation of language models. *CoRR*, abs/2211.09110.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *CoRR*, abs/2202.12837.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El-Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling - 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II*, volume 9517 of *Lecture Notes in Computer Science*, pages 15–27. Springer.

NLP Connect. 2022. vit-gpt2-image-captioning (revision 0e334c7).

Ariadna Quattoni and Antonio Torralba. 2009. Recognizing indoor scenes. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 413–420. IEEE Computer Society.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and et al. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, and et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6418–6428. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olmpics - on what language model pre-training captures. *Trans. Assoc. Comput. Linguistics*, 8:743–758.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 200–212.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, Shih-Fu Chang, Mohit Bansal, and Heng Ji. 2022b. Language models with image descriptors are strong few-shot video-language learners. In *Advances in Neural Information Processing Systems*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision. *CoRR*, abs/2006.03677.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based VQA. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 2712–2721. AAAI Press.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 3081–3089. AAAI Press.

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael S. Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *CoRR*, abs/2204.00598.

Jing Zhang and Yujin Wang. 2022. SRCB at SemEval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 585–596, Seattle, United States. Association for Computational Linguistics.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, and et al. 2022. Opt: Open pre-trained transformer language models.

Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16795–16804. IEEE.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348.

A Prompt Structures

We provided the prompt structures for all datasets in Figure 3. Given a sample from a respective dataset, the prompt structures are initialised to create a prompt text. Each prompt text includes a task description followed by task-specific labels (only for classification tasks). In the middle of the prompt text are the selected in-context samples. The bottom part includes the evaluation sample, which is represented by only its input text, image-as-text representation (image context), and the task-specific question. The prompted language model is expected to generate the next text sequence that starts after the last occurrence of the word *Answer*.

MAMI

Given an image context and text, answer the given question. The task is to classify whether the context is hateful towards women (includes misogynistic tone).
Possible labels:
1. **Hateful**
2. **Not**

Text: <sample_text>
Image context: <sample_image_as_text_repr>
Question: What is the label of the given image context and text (hateful, not)?
Answer: <sample_answer>

Text: <sample_text>
Image context: <sample_image_as_text_repr>
Question: What is the label of the given image context and text (hateful, not)?
Answer: <sample_answer>

…
Text: <sample_text>
Image context: <sample_image_as_text_repr>
Question: What is the label of the given image context and text (hateful, not)?
Answer:

HF

Given an image context and text, answer the given question. The task is to classify whether the content is hateful towards any group of people.
Possible labels:
1. **Hateful**
2. **Not**

Text: <sample_text>
Image context: <sample_image_as_text_repr>
Question: What is the label of the given image context and text (hateful, not)?
Answer: <sample_answer>

Text: <sample_text>
Image context: <sample_image_as_text_repr>
Question: What is the label of the given image context and text (hateful, not)?
Answer: <sample_answer>

…
Text: <sample_text>
Image context: <sample_image_as_text_repr>
Question: What is the label of the given image context and text (hateful, not)?
Answer:

MVSA

Given an image context and text, answer the given question. The task is to classify whether the context has positive, negative or neutral sentiment.
Possible labels:
1. **Positive**
2. **Negative**
3. **Neutral**

Text: <sample_text>
Image context: <sample_image_as_text_repr>
Question: What is the label of the given image context and text (positive, negative, neutral)?
Answer: <sample_answer>

Text: <sample_text>
Image context: <sample_image_as_text_repr>
Question: What is the label of the given image context and text (positive, negative, neutral)?
Answer: <sample_answer>

…
Text: <sample_text>
Image context: <sample_image_as_text_repr>
Question: What is the label of the given image context and text (positive, negative, neutral)?
Answer:

(a) Prompt structures for MAMI, HF and MVSA datasets

NLVR2

Given an image context for two images (left and right) and text, answer the given question. The task is to classify whether the given text is true or false.
Possible labels:
1. **True**
2. **False**

Text: <sample_text>
Left Image context: <sample_image_as_text_repr>
Right Image context: <sample_image_as_text_repr>
Question: What is the label of the given left image context, right image context, and text (true, false)?
Answer: <sample_answer>

…

Text: <sample_text>
Left Image context: <sample_image_as_text_repr>
Right Image context: <sample_image_as_text_repr>
Question: What is the label of the given left image context, right image context, and text (true, false)?
Answer:

OK-VQA

Answer the following question given an image context.

Image context: <sample_image_as_text_repr>
Question: <sample_question>
Answer: <sample_answer>

Image context: <sample_image_as_text_repr>
Question: <sample_question>
Answer: <sample_answer>

…

Image context: <sample_image_as_text_repr>
Question: <sample_question>
Answer:

(b) Prompt structures for NLVR2, OK-VQA datasets

Figure 3: Prompt structures for each evaluated dataset. Each prompt structure includes a task description, which also includes possible labels, selected *n* in-context samples, and the evaluated sample. The prompted language models are expected to generated the next text sequence that starts after the last occurrence of the word *Answer*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4

☑ B1. Did you cite the creators of artifacts you used?
4

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.