MemeQA: Holistic Evaluation for Meme Understanding

Khoi P. N. Nguyen[1] Terrence Li[1] Derek Lou Zhou[1] Gabriel Xiong[2]
Pranav Balu[1] Nandhan Alahari[1] Alan Huang[3] Tanush Chauhan[4]
Harshavardhan Bala[1] Emre Guzelordu[4] Affan Kashfi[5] Aaron Xu[6]
Suyesh Shrestha[1] Megan Kim Vu[1] Jerry Yining Wang[1] Vincent Ng[1]
[1]University of Texas at Dallas [2]William P. Clements High School [3]Stanford University
[4]University of Texas at Austin [5]Rock Hill High School [6]Texas A&M University
{khoi.nguyen6,terrence.li}@utdallas.edu, vince@hlt.utdallas.edu

Abstract

Automated meme understanding requires systems to demonstrate fine-grained visual recognition, commonsense reasoning, and extensive cultural knowledge. However, existing benchmarks for meme understanding only concern narrow aspects of meme semantics. To fill this gap, we present MemeQA, a dataset of over 9,000 multiple-choice questions designed to holistically evaluate meme comprehension across seven cognitive aspects. Experiments show that state-of-the-art Large Multimodal Models perform much worse than humans on MemeQA. While fine-tuning improves their performance, they still make many errors on memes wherein proper understanding requires going beyond surface-level sentiment. Moreover, injecting "None of the above" into the available options makes the questions more challenging for the models. Our dataset is publicly available at https://github.com/npnkhoi/memeqa.[1]

1 Introduction

Recent years have seen the development of several extensively used benchmarks in the form of multiple-choice questions for evaluating various capabilities of Large Language Models (LLMs), such as Swag (Zellers et al., 2018) and HellaSwag (Zellers et al., 2019). While traditional work has focused on textual Question Answering (QA), recent work has focused on multimodal QA, particularly Visual Question Answering (VQA), where the goal is to answer questions about an image.

Meme-based Multimodal Question Answering (Agarwal et al., 2024), or *MemeMQA*, is a new task in VQA that involves answering questions about a meme. *Memes* are a communicative type of images overlaid with text meant to cause laughter while expressing social commentary. Given the popularity of memes in online communication, there is a growing interest among NLP researchers in modeling complex aspects of memes to keep the internet safe. Such aspects include harmfulness, targeted social groups, offensive cues, and narrative framing.

MemeMQA is arguably more challenging than general VQA. In VQA, the questions are typically designed to elicit understanding of the reality depicted in the images, such as asking "What kind of cheese is on the pizza?" (fine-grained recognition), "How many bikes are there?" (object recognition), "Is this a vegetarian pizza?" (basic knowledge base reasoning), or "Is this person expecting company?" (commonsense reasoning) (Antol et al., 2015). In contrast, memes are crafted by the author to convey *deeper* ideas. For example, the meme in Figure 1 is not simply concerned about the angry woman on the left or the cloud-like cat on the right. The ultimate intention of the meme is to mock "anti-maskers". Understanding the author's intent requires non-trivial subtasks such as retrieving background knowledge, recognizing the sentiment of the target, and deriving implications.

To advance research in MemeMQA, we present MemeQA, a corpus of 9,031 multiple-choice questions about memes that aim to evaluate various aspects of meme understanding. The key innovations of MemeQA include:

Holistic evaluation for meme understanding. As will be discussed in Section 2, the vast majority of existing work on automated meme processing has focused on determining the meme author's communicative *intent* (e.g., the intent can be to provoke hate towards a particular group) or classifying memes based on this intent (e.g., classifying a meme as hateful or not). In contrast, the design of MemeQA is motivated by the reasoning steps that humans used to *derive* the communicative intent. In other words, MemeQA concerns not only *what* the intent is, but also *how* the intent is derived, thus covering a wider range of aspects of meme understanding compared to existing work. The ability to

[1]WARNING: The memes used in this paper are purely for illustration purposes. Some readers may find them offensive.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18926–18946

evaluate a model's understanding of how the intent is derived, though missing from existing work, is important from the point of view of *explainability*: even if a model correctly determines the author's intent, without evaluating whether it understands how the intent is derived, we would not know whether it simply gets the right answer for the wrong reason.

Diagnostic evaluation of model outputs. As mentioned above, MemeQA enables us to evaluate whether a model derives the correct intent using the correct reasoning process. However, if a model derives the wrong intent, it is equally important for us to understand what went wrong. MemeQA allows us to shortlist the possible "culprits" by determining which question(s) related to the meme under consideration were incorrectly answered. For instance, if a model incorrectly answered the question of "what background knowledge is relevant to determining the intent?", then we could attribute the wrong intent it outputted at least in part to missing or misidentification of relevant background knowledge. In contrast, existing work on meme processing has rarely investigated the question of why a model makes a wrong prediction (e.g., why a model misclassifies a meme as hateful).

Evaluation of new model capabilities. In many multiple-choice QA benchmarks, a model is asked to answer a question by picking the (only) correct option out of four given options. This setup, however, does not necessarily test a model's ability to determine which answer is correct: since exactly one answer is known to be correct, all a model needs to do is to rank the options and pick the most plausible one even if it does not believe that the most plausible option is a correct answer. To address this weakness, we present a second version of MemeQA that aims to test the ability of a model to identify the correct answer. Specifically, we create a "None of the above" option that should be chosen if and only if none of the other options corresponds to the correct answer. We believe that this is a more challenging version of MemeQA, as a model that merely returns the most plausible answer without determining whether it is correct may no longer do well on this version of the dataset.

Experiments show that (1) state-of-the-art large multimodal models (LMMs) all perform poorly on MemeQA, and while fine-tuning improves their performance, it is still far from decent; and (2) models achieve poorer results on the version of MemeQA with "None of the above", suggesting that this new question creation methodology indeed presents new challenges to models. We believe MemeQA can pave the way for advancements not only in technical methodologies, but also in ethical applications in content moderation and digital communication.

2 Related Work

Visual Question Answering While earlier formulations of VQA have involved answering simple questions related to images (Antol et al., 2015; Goyal et al., 2017; Ren et al., 2015), later research tackles new aspects of the problem, such as compositional language and elementary reasoning (Johnson et al., 2017; Hudson and Manning, 2019), external knowledge leverage (Wang et al., 2018, 2017; Marino et al., 2019), and diagnostics (Johnson et al., 2017). These tasks spawn follow-up research that creates the class of LMMs that can process both images and texts (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Wang et al., 2024).

Meme Question Answering MemeMQA is a relatively new task that was first studied by Agarwal et al. (2024), who seek to answer questions about the semantic roles of entities in memes. Specifically, they formulated the following task: Given a meme and a semantic role, (1) identify which entity among four options plays such a role in the meme, and (2) provide a concise explanation for the choice. For example, one of their questions asks "What is slandered in this meme?" (i.e., "Who's the villain?"), with "Democratic Party" as the correct choice among four options. Following this formulation, the authors released MemeMQACorpus, the first and by far the only dataset for MemeMQA. Compared to MemeQA, MemeMQACorpus is much smaller (fewer than 2K questions) and narrower in scope, focusing only on role-based queries as opposed to general questions about the intent of a meme.

Meme-related tasks Much existing work on meme processing has facilitated the detection of malicious memes (Suryawanshi et al., 2020a,b; Kiela et al., 2020; Chandra et al., 2021; Pramanick et al., 2021a,b; Fersini et al., 2022). Others proposed to process more general aspects of memes, such as persuasion techniques (Dimitrov et al., 2021), figurative language (Liu et al., 2022), entity roles (Sharma et al., 2022), emotions (Sharma et al., 2020), targeted attacks against groups (Mathias et al., 2021), and overall captions (Hwang and Shwartz, 2023; Park et al., 2024). Details can be found in the survey by Nguyen and Ng (2024).

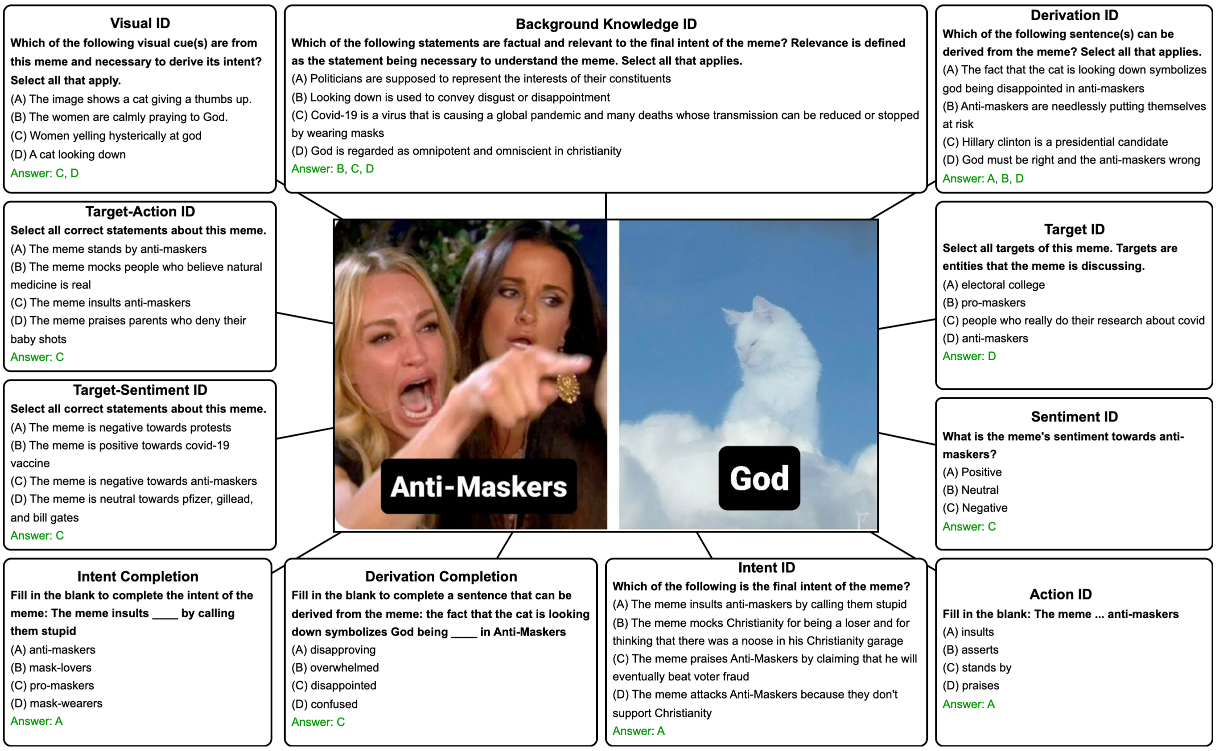


Figure 1: An example meme from `MemeQA` together with the questions created from it.

3 The `MemeQA` Dataset

3.1 Meme Source

To facilitate the creation of questions, we employ the 950 memes in the SemEval2021 Task 6 dataset (Dimitrov et al., 2021). These memes are collected from Facebook in the English language, and cover a variety of social topics, including politics, government and law, science and conspiracies, healthcare, media and corporation, social movements, and international issues.

3.2 Design Methodology

Next, we motivate the design of `MemeQA`, including what aspects of meme understanding and types of questions are to be covered.

3.2.1 Seven Aspects of Meme Understanding

In holistically evaluating a system's understanding of a meme, we break down what it means for a human to "understand" a meme. In doing so, we take inspirations from the cognitive processes a human has to go through when reading a meme.

Consider the meme in the center of Figure 1. To understand this meme, a human first looks at its surface-level details – reading the text and **recognizing the scene** in the image. For this meme, the details are the word "Anti-Maskers" placed under the image of an angry yelling woman on the left, and the word "God" placed under a cloud-like cat looking down towards the woman. After that, they would **connect certain details on the meme with their background knowledge**. Understanding this meme requires knowledge from meme culture that *the pair of images resembles the "Woman Yelling at a Cat"*[2] *meme macro, which is used to make fun of the overreaction and ignorance represented by the woman character (Fact 1)* and the world event that *Covid-19 is a life-threatening disease (Fact 2)*. One or more reasoning steps are then applied to these premises to derive intermediate conclusions (which we call *derivations*) and eventually the final conclusion, which corresponds to the intent of the meme. For this meme, from all the surface-level details and Fact 1, one can make a **derivation** that *the meme portrays God laughing at anti-maskers for being ignorant and overreacting*. From the above derivation and Fact 2, one can conclude the **intent of the meme** to be *criticizing anti-maskers for being ignorant and risking their lives*. So, the **social target** of the meme is *anti-maskers*, the **sentiment** towards them is *negative*, and the **meme's action** towards them is *criticizing*.

The example above illustrates seven aspects in meme understanding: (1) **visual cues** (the visual

[2] https://knowyourmeme.com/memes/woman-yelling-at-a-cat

information that is important for understanding the meme's intent), (2) **background knowledge** (the relevant facts needed to understand the meme's intent), (3) **social targets**, (4) **action towards the targets**, (5) **sentiment towards the targets**, (6) **derivations** (the intermediate conclusions that can be drawn), and (7) **intent** (what the meme author intends to convey). These aspects will be the backbone in our design of the questions in MemeQA to holistically evaluate meme understanding systems.

3.2.2 Question Types

Next, we define the 11 question types in MemeQA, all of which are *multiple-choice* questions.

First, from the seven aspects, we derive seven types of questions: (1) **Visual Identification**, (2) **Background Knowledge Identification**, (3) **Derivation Identification**, (4) **Social Target Identification**, (5) **Sentiment Identification**, (6) **Action Identification**, and (7) **Intent Identification**.

Next, noticing that question types 5 and 6 are specific to one of the (possibly many) social targets of the meme, we create two other types of questions that require selecting a *combination* of target and sentiment, or target and action, namely (8) **Target-Sentiment Identification** and (9) **Target-Action Identification**.

Finally, we design two types of *cloze* questions based on two of the seven aspects of meme understanding, derivations and intent. The two new question types, (10) **Derivation Completion** and (11) **Intent Completion**, are also formulated as multiple-choice questions where the task involves filling in the blank with one of the given options. While there appears to be some overlap between these question types and two of the aforementioned question types (Derivation Identification and Intent Identification), our goal is to examine whether asking the model to fill in a blank in the intent/derivation is easier than having it identify the intent/derivation.

Except for the Sentiment Identification questions, which are presented with three options (POSITIVE, NEGATIVE, and NEUTRAL), each question in the remaining question types has exactly four options. As shown in Table 1 (#A), six types of questions require selecting *all* correct answers (henceforth *multiple-answer questions*), while the remaining types require selecting exactly one (henceforth *single-answer questions*).

Question type	#Q/M	#A	#Q
Visual Identification	1	0-4	463
Background Identification	1	0-4	287
Derivation Identification	1	0-4	215
Target Identification	1	0-4	849
Sentiment Identification	#T	1	617
Action Identification	#T	1	245
Intent Identification	1	1	298
Derivation Completion	#C	1	1786
Intent Completion	#C	1	2520
Target-Sentiment Identification	1	0-4	895
Target-Action Identification	1	0-4	856

Table 1: **Question types in MemeQA.** "#Q/M" is the number of questions per meme; "#C" is the number of content words in the base sentences; "#T" is the number of targets in the meme; "#A" is the number of correct answers; and "#Q" is the final number of questions.

3.3 Question-Answer Creation

Now that we have the question types, we can begin creating the *questions* and the associated *options* (henceforth QO pairs) in MemeQA. Rather than creating the QO pairs in a top-down fashion where the question in each pair is created before the options, we create them in a bottom-up fashion since the cloze questions cannot be created without already knowing the derivations and the intent.

3.3.1 Step 1: Recording Reasoning Processes

We begin by obtaining ground-truth annotations about the seven aspects for each meme. To do so, we asked human annotators to write an *interpretation paragraph* for each meme that represents the reasoning process that they employ to derive the intent of a meme from the surface level cues (i.e., the textual and visual cues present in the meme and any needed background knowledge). Specifically, we distributed each meme to one of the five annotators we trained, and asked them to write interpretation paragraphs that are sufficiently detailed so that the aforementioned seven aspects are covered.[3]

Given an interpretation paragraph, the annotators manually extracted from it the seven aspects mentioned above. Figure 2 illustrates an interpretation paragraph written for the meme shown in Figure 1, and the seven aspects that are being extracted manually. Among the seven aspects, background knowledge and derivations can be absent from a paragraph. Specifically, background knowledge will be absent if no background knowledge is needed to infer the intent, and derivations will

[3]Details of annotator background and the training process can be found in Appendix A. The annotation guidelines are shown in Appendix B.

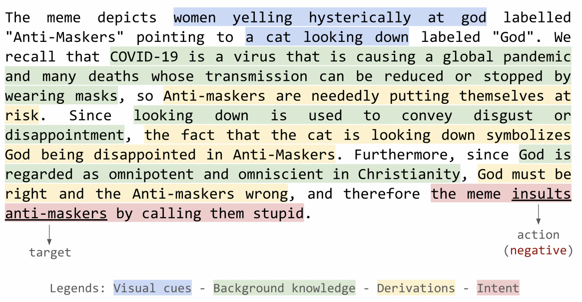


Figure 2: An example interpretation paragraph.

be absent if the intent can be inferred directly from the visual and textual information (and any background information, if applicable) without going through any intermediate derivations.

3.3.2 Step 2: Creating Preliminary QO Pairs

Given the seven aspects annotated for each meme, we created a preliminary version of the QO pairs. Recall that each QO pair is composed of (1) the question, (2) the correct answer(s), and the (3) distractors (i.e., the wrong options), where the number of correct answer(s) and distractors should sum to four for each question. Below we describe how each of these three elements of a QO pair is created. We refer to these as preliminary QO pairs, as the distractors will be refined in the next subsection.

Questions For the cloze question types (Intent Completion and Derivation Completion), we created questions as follows. Given a sentence in the interpretation paragraph that is annotated as an intent/derivation, we create one cloze question by masking out exactly one of the content words (i.e., noun phrases, verbs, adjectives, or adverbs) in it. Hence, the number of content words determines the number of cloze questions to be created.

For Sentiment ID and Action ID, the number of questions created for a meme is equal to the number of social targets that appear in it. Specifically, for each target, we create one Sentiment ID question and one Action ID question, which are the same as the corresponding questions shown in Figure 1 except that the target "anti-maskers" is replaced with the target under consideration.

For the remaining seven question types, we instantiated one question per meme, where the question being asked is the same as the corresponding question shown in Figure 1.

Correct option(s) Recall that except for Sentiment ID questions, all questions have four options. For multiple-answer questions, the correct option(s) are created as follows. For instance, if no background knowledge can be extracted from the interpretation paragraph, then no correct option will be created (and the question will have four incorrect options). If one to four pieces of background knowledge can be extracted, then each piece will be used as a correct option. Finally, if more than four pieces of background knowledge can be extracted, then four pieces will be randomly chosen and used as correct options. For single-answer questions, the correct option is exactly what is extracted from the paragraph.

Distractors The number of distractors to be created depends on the question type. For Sentiment ID, the options are fixed, so no extra distractors are needed. For single-answer questions, three distractors are needed. For multiple-answer questions, the number of distractors needed depends on the number of correct options created above.

The distractors were created heuristically. For cloze questions, the distractors are the antonyms of the correct answer obtained using Spacy (Honnibal and Montani, 2017). For questions asking for social targets, the distractors are the targets randomly sampled from the other questions in `MemeQA` that are distinct from the target in the question under consideration. For the remaining question types, the distractors are the correct answers randomly sampled from the other questions in `MemeQA`.[4]

3.3.3 Step 3: Refining the Questions

On the initial set of questions, "off-the-shelf" Qwen2-VL (Wang et al., 2024) scored around 90% accuracy. This high performance can be attributed to the fact that the distractors heuristically created in the previous subsection were not optimized for difficulty. Therefore, we make the questions more challenging by using Adversarial Filtering (AF) (Zellers et al., 2018) to replace the "easy" distractors. The idea behind AF is to iteratively update the distractors in the questions using two adversarial models, the *discriminator* and the *generator*.[5]

Specifically, in each AF iteration, the discriminator answers all the questions. If a question is incorrectly answered, it will be deemed sufficiently difficult and no change is made to it. Otherwise, AF will increase its difficulty by replacing all the distractors with the new ones generated by the gen-

[4]This makes sense because most memes in our meme collection differ from each other on most of the aspects, including intent, background knowledge, visual cues, and derivations.

[5]A schematic representation of the Adversarial Filtering process can be found in Figure 4 (Appendix C).

erator. The process continues until the discriminator's performance stabilizes. In our implementation, we used Qwen-2 VL (Wang et al., 2024) as the discriminator and Llama-3.1-8B-Instruct (Dubey et al., 2024) as the generator.[6]

3.4 Human Verification

Recall that the correct answer for each question created in the aforementioned three-step process was extracted by a human annotator (henceforth the answer extractor) from the interpretation paragraph in the *absence* of the automatically/heuristically created distractors. For this reason, we perform human verification, where the goal is to determine if additional annotators would still pick the option(s) that are deemed correct by the answer extractor in the *presence* of the distractors.

To avoid the bias carried from the paragraph annotation stage, we recruited 10 new annotators as verifiers. For each question, we had a verifier answer *without* letting them know which option(s) were supposedly correct. A question is discarded if the verifier does not agree on the correct answer(s) or thinks that none of the options in a single-answer question is correct. 88% of the original questions survived the human verification step.

3.5 Two Versions of `MemeQA`

Using the questions that survived human verification, we create two versions of `MemeQA`:

None^{-} Since we are designing a dataset for evaluating meme understanding, we desire questions that cannot be correctly answered without referencing the corresponding meme. Rather than doing this manually, which would be labor-intensive, we approximate this process by having Llama-3.1-8B-Instruct (Dubey et al., 2024) answer all the questions without providing the meme to it. If a question is answered correctly, we assume that those questions are trivial (as it can be answered without the meme) and therefore removed it from the dataset. This process removed another 30% of the questions, resulting in a version of `MemeQA` that we refer to as *None*$^{-}$.[7]

None^{+} A model may be able to answer a question in *None*$^{-}$ correctly simply by ranking the options and returning the k most plausible options (where k=1 for single-answer questions) even if it does not believe that they are correct. This motivates us to create *None*$^{+}$, the second version of `MemeQA`, in which each question has "None of the above" as one of its options. The questions in *None*$^{+}$ are presumably more challenging than those in *None*$^{-}$: in addition to identifying the most plausible option(s), a model will need to determine if these option(s) are indeed the correct answer(s).

We create *None*$^{+}$ from *None*$^{-}$ as follows. To maintain randomness, 25% of the questions in *None*$^{-}$ had its correct answer replaced with "None of the above". For the remaining 75% of the questions, we replaced one randomly chosen wrong option with "None of the above". These substitutions are made to all question types except Sentiment ID, which must have a fixed set of options.

3.6 Final Dataset

The final dataset contains 9,031 questions for each version.[8] The distribution of question types is shown in Table 1 (last column). `MemeQA` is split into training, development, and test with a ratio of 60:20:20. To avoid data leakage, the questions are split at the meme level, meaning that the questions about the same meme appear in the same data split.

4 Evaluation

Next, we conducted benchmarking experiments to gauge the performance of current state-of-the-art (SoTA) vision-language models on `MemeQA`.

4.1 Experimental Setup

Models We selected five most performant open-sourced models, namely **Qwen2-VL-7B-Instruct** (Wang et al., 2024), **BLIP2-Flan-T5-xl** (Li et al., 2023), **InstructBLIP-Vicuna-7B** (Dai et al., 2023), **LLaVA-v1.5** (Liu et al., 2024), and **QVQ-72B-Preview** (Qwen Team, 2024), and a SoTA close-sourced model, **GPT-4o** (OpenAI et al., 2024)[9].

Model outputs Model outputs are represented as text strings. For single-answer questions, the answer must be either "A", "B", "C", or "D". For multiple-answer questions, the answer is a list of characters among A, B, C, D, in alphabetical order (e.g., "ACD"). In version *None*$^{-}$, if no options should be chosen, the answer must be "N".[10] Simple heuristics are applied to extract the answers

[6] See Appendix C for details on why these models were chosen. The prompts used for the discriminator and the generator are shown in Appendices D and E respectively.

[7] We studied the effect of removing these questions in Appendix F.

[8] Example QO pairs can be found in Appendix G.

[9] An overview of these models can be found in Appendix H.

[10] These details are reflected in the prompt templates shown in Appendix D.

		None$^-$							*None*$^+$						
Aspect	Rand	LLaVA	BLIP	IBLIP	Qwen	QVQ	GPT	Hum	LLaVA	BLIP	IBLIP	Qwen	QVQ	GPT	Hum
Visual ID	6.3	25.0	58.3	41.7	77.8	**80.6**	73.6	93.1	31.9	45.8	36.1	62.5	**76.4**	70.8	90.3
Background ID	6.3	13.0	20.4	18.5	27.8	48.1	**61.1**	66.7	22.2	22.2	20.4	27.8	48.1	**64.8**	66.7
Target ID	6.3	12.6	25.1	25.7	32.9	47.9	**55.7**	65.3	25.7	25.7	23.4	24.6	46.7	**58.7**	66.5
Derivation ID	6.3	5.6	22.2	22.2	30.6	41.7	**58.3**	77.8	19.4	19.4	19.4	22.2	47.2	**66.7**	77.8
Target-Sentiment ID	6.3	1.1	27.3	25.1	21.9	41.0	**47.0**	68.3	14.8	27.3	25.7	13.7	35.0	**48.1**	68.3
Target-Action ID	6.3	1.2	18.0	19.8	19.2	41.9	**44.2**	70.3	12.2	23.8	16.9	11.0	39.5	**48.3**	66.9
Deriv. Completion	25.0	34.3	41.9	34.9	54.1	49.7	**64.8**	91.9	30.8	38.4	32.6	42.2	46.2	**66.3**	88.4
Intent Completion	25.0	36.0	47.0	34.6	51.5	54.4	**68.3**	91.0	32.4	41.3	33.3	37.8	48.6	**67.0**	84.5
Intent ID	25.0	36.2	29.8	17.0	44.7	**61.7**	59.6	97.9	36.2	25.5	19.1	27.7	46.8	**68.1**	95.7
Action ID	25.0	22.7	22.7	17.0	14.8	33.0	**59.1**	92.0	23.9	13.6	12.5	4.5	23.9	**55.7**	90.9
Sentiment ID	33.3	47.3	45.6	42.0	43.8	**63.9**	**63.9**	87.0	47.3	45.6	42.0	45.0	65.7	**63.3**	84.6
Macro Average	10.9	21.4	32.6	27.1	38.1	51.3	**59.6**	81.9	27.0	29.9	25.6	29.0	47.7	**61.6**	80.0

Table 2: **Zero-shot results on the two versions of `MemeQA`**. "Rand" shows the expected accuracy for random guessing and "Hum" stands for human performance. The best accuracy in each group and each question type is **boldfaced**. Accuracies lower than random guessing are underlined.

from the responses. If the output is not parseable, its answer will be deemed wrong.[11]

Evaluation metrics We report the performance of a model on each question type in terms of *accuracy*, which is the percentage of questions that are correctly answered. Specifically, for a single-answer question, we consider it correctly answered if and only if the correct option is selected. For a multiple-answer question, we consider it correctly answered if and only if all and only those correct options are selected. In addition, we aggregate the results over different question types by computing the macro-average, which is the unweighted average of the accuracies on all the question types.

Settings We evaluate models in the *zero-shot* setting, where no data from `MemeQA` was used to train models, and the *fine-tuned* setting, where models were fine-tuned on the training split of `MemeQA` with the hyperparameters tuned on development data. Note that we did not fine-tune QVQ and GPT-4o on `MemeQA` since GPT-4o cannot be fine-tuned on images with people and faces due to OpenAI's content moderation policy, and fine-tuning QVQ requires reasoning data not available with `MemeQA`.

Implementation details During LMM inference, greedy generation is used (≤ 10 new tokens). During fine-tuning, we used the Parameter Efficient Fine Tuning technique, attaching and training a LoRA adapter (Hu et al., 2022) to all the linear modules in the models. We trained the models for at most 3 epochs with `batch_size=4`, `lr=`10^{-5}, `lora_alpha=8`, `lora_dropout=0.1`, `r=8`. The models were evaluated on the development set after every 20% of the training set and the checkpoint with the highest accuracy on the development set was chosen as the final one. All experiments took about 20 hours on a computer with 2x RTX A6000.

[11]The parsing failure rate is less than 1.5% of the dataset.

4.2 Results and Discussion

Zero-shot and fine-tuned results are shown in Tables 2 and 3, respectively.

Which LMMs perform the best? The larger-sized models perform significantly better: in the zero-shot setting, GPT-4o scores the highest, followed by QVQ. These two models outperform the smaller models by 10–30 percentage points. Among the smaller models, Qwen and BLIP performed better than LLaVA and InstructBLIP.

Does fine-tuning help? Yes. All models exhibited improvements by a large margin, except LLaVA on *None*$^+$. Qwen consistently performs the best, scoring on average 70% on *None*$^-$ and 65% on *None*$^+$. These are over 30 percentage-point improvements, showing the effectiveness of fine-tuning. However, even the best fine-tuned models are far from perfect. Note, though, that even without fine-tuning, GPT-4o performs competitively with fine-tuned Qwen, lagging behind Qwen by 10% and 3% on *None*$^-$ and *None*$^+$, respectively.

Are some types of questions easier to answer? Yes. Visual ID (which tests entity recognition) and the cloze question types are easier for the models than questions related to background knowledge, targets, sentiments, and actions.

Is *None*$^+$ harder than *None*$^-$? Noticeably. While the best model scored an average accuracy of 69.8% on *None*$^-$, the best model on *None*$^+$ only

Aspect	Rand	*None*$^-$ LLaVA	BLIP	IBLIP	Qwen	QVQ	GPT	Hum	*None*$^+$ LLaVA	BLIP	ILIP	Qwen	QVQ	GPT	Hum
Visual ID	6.3	70.8	65.3	75.0	**94.4**	—	—	93.1	26.4	58.3	72.2	**94.4**	—	—	90.3
Background ID	6.3	18.5	22.2	31.5	**66.7**	—	—	66.7	5.6	25.9	35.2	**72.2**	—	—	66.7
Target ID	6.3	31.7	40.1	41.9	**54.5**	—	—	65.3	6.0	35.3	41.9	**57.5**	—	—	66.5
Derivation ID	6.3	36.1	33.3	55.6	**69.4**	—	—	77.8	16.7	27.8	47.2	**50.0**	—	—	77.8
Target-Sentiment ID	6.3	33.9	39.3	43.7	**57.4**	—	—	68.3	1.6	34.4	47.0	**49.7**	—	—	68.3
Target-Action ID	6.3	27.3	34.9	35.5	**47.1**	—	—	70.3	1.2	31.4	36.6	**46.5**	—	—	66.9
Deriv. Completion	25.0	79.7	73.8	84.0	**85.8**	—	—	91.9	45.6	63.4	71.8	**82.3**	—	—	88.4
Intent Completion	25.0	85.9	79.6	83.4	**87.7**	—	—	91.0	41.6	69.5	75.7	**79.6**	—	—	84.5
Intent ID	25.0	83.0	61.7	74.5	**85.1**	—	—	97.9	57.4	57.4	66.0	**68.1**	—	—	95.7
Action ID	25.0	**65.9**	46.6	45.5	60.2	—	—	92.0	48.9	34.1	31.8	**53.4**	—	—	90.9
Sentiment ID	33.3	55.0	47.3	47.3	**59.2**	—	—	87.0	6.5	47.3	47.3	**59.8**	—	—	84.6
Macro Average	10.9	53.4	49.5	56.2	**69.8**	—	—	81.9	23.4	44.1	52.1	**64.9**	—	—	80.0

Table 3: **Fine-tuned results on the two versions of `MemeQA`**. "Rand" shows the expected accuracy for random guessing and "Hum" stands for human performance. The best accuracy in each group and each question type is **boldfaced**. Accuracies lower than random guessing are underlined.

achieved 64.9%. This inequality holds within almost all models and and question types. The consistent pattern here suggests that substituting "None of the above" into the choices creates significant challenges for the models.

As an exception, GPT-4o performs slightly better on *None*$^+$ than *None*$^-$. Looking more closely, we see that it performs better on *None*$^+$ mostly on multiple-answer questions but worse on single-answer questions. This still aligns with our intuition that *None*$^+$ will be harder at single-answer questions. However, for multiple-answer questions, replacing an option with “None of the above” can make a question easier.

How well do the LMMs perform relative to humans? Humans perform far better than all zero-shot and fine-tuned models, scoring over 80% of macro-average on both dataset versions. Note that fine-tuned Qwen, the best-performing model, underperforms humans by 12–15 percentage points, meaning that `MemeQA` still presents significant challenges to SoTA LMMs.[12]

4.3 Error Analysis

Our analysis is guided by two questions: (1) "what are the challenges from each question type?" and (2) "what effect does *None*$^+$ have on the difficulty of questions?" These questions are answered by sampling `MemeQA`’s questions which our best-performing model, fine-tuned Qwen, answered incorrectly to find patterns within the errors.[13]

[12]Details on how we obtained human performance are in Appendix A.

[13]Examples of errors made by fine-tuned Qwen can be found in Appendix I.

4.3.1 Unique Challenges from the Questions

To answer the first question, 30 questions which the model answered incorrectly were randomly sampled from each of the 11 aspects. Three key observations were found across the question types.

Target-related challenges Target-Sentiment ID, Target-Action ID, and Target ID reveal the model’s errors in identifying targets. Particularly, the model fails to identify targets that do not explicitly appear in the meme, because they require complex reasoning steps and/or cultural context to understand. Figure 6a illustrates one of those cases. When targets are partially identified correctly, however, the model fails to identify other targets that are related or complementary to the correctly identified one (see Figure 6b). These behaviors illustrate the model’s struggle both to look deeper than the visuals of the meme and to identify the complementary nature of paired targets (one being praised, the other criticized).

Sentiment-related challenges Sentiment and Action ID questions reveal another pattern: when the target is provided or identified correctly, the model can still fail to correctly identify its sentiment. Particularly in the samples, this occurred if the superficial tone of the meme (the initially evoked emotions, before any deeper reasoning or subtlety) differed from the sentiment of the target. See Figure 6c, where the negative and dark tones are applied incorrectly onto the supplied target.

This pattern manifested beyond Sentiment and Action ID, in aspects which are inherently reliant on sentiments of targets (i.e., all of them except Visual and Background ID). This reliance origi-

	Incorrect replaced		Correct replaced	
	Single	**Multiple**	**Single**	**Multiple**
None$^-$ Acc.	77.3	54.9	95.9	97.2
None$^+$ Acc.	77.2	54.7	67.6	79.1
Difference	0.1	0.2	28.3	16.1

Table 4: **Accuracies of fine-tuned Qwen on *None*$^-$ and *None*$^+$** divided across (1) whether a correct answer choice is replaced, and (2) whether the questions are single-answer or multiple-answer.

nates from the intents of memes, which are built on sentiment-associated action-verbs (e.g., "criticizes", "praises") towards certain targets. Hence, if the model failed to identify the correct sentiment for a given target, it would consistently fail across multiple question types. See Figure 7, which illustrates this behavior. The issue was especially apparent in questions that featured subtle literary elements like irony or sarcasm.

Background knowledge ID Meme annotations often used uncertain phrasing such as "many believe...", "tend to be...". In comparison, the distractors were phrased confidently. Upon closer inspection, the model was biased against (correct) choices which featured *uncertain* wording. See Figure 6d. All four options belong to the answer-key, but the neutral-worded options are not selected.

4.3.2 Effects of *None*$^+$ on Question Difficulty

From Table 4, it is evident that replacing a correct answer choice creates significant difficulty in answering questions. Deeper analysis is performed by sampling 30 questions from each of the categorical combinations of Table 4 (4 total). These questions were answered correctly in *None*$^-$ but incorrectly in *None*$^+$. As such, these errors should reveal model behaviors and why *None*$^+$ is difficult.

The changes between *None*$^-$ and *None*$^+$ further support the observation that identifying a target's sentiment is more difficult than identifying targets. The same pattern of mistaking targets as negative instead of positive appears again, though the introduction of the "None of the above" option in *None*$^+$ exacerbates the evidence of sentiment identification's difficulty. Figure 6e illustrates the removal of the correct target-action pairing, resulting in the model opting for a pairing with the correct target but an incorrect sentiment. It appears as though it is selecting the "second-best" answer, as the other options are considerably less relevant. These questions in *None*$^+$ require the identification of not only a target's correct sentiment, but also an answer choice's incorrect sentiment. This further complicates the task of sentiment identification.

4.3.3 Implications

Our error analysis shows that models struggle with reasoning, particularly when they are required to go beyond surface content and interpret the underlying message of a meme. For example, models often default to literal sentiment cues from text or imagery, failing to grasp irony, sarcasm, or the implied stance. This indicates that reasoning is the key area for improvement, especially in understanding deeper contextual meaning. To improve reasoning capabilities, one could refine prompting strategies for LMMs. For example, prompts could explicitly ask whether a meme is ironic or sarcastic and request an explanation of how the deeper meaning diverges from surface sentiment. The output of such inferences can be used to augment the input to LMMs when answering MemeMQA questions.

Beyond architectural improvements, it is crucial to scale data with richer supervision that reflects how humans interpret memes. As demonstrated in Section 3.3.1, annotators can describe the reasoning steps they take to arrive at a meme's deeper meaning, including how they detect dissonance and infer implied targets and sentiments. Training models on such step-by-step annotations could significantly improve their ability to interpret nuanced or culturally embedded messages that go beyond surface-level sentiment.

5 Conclusion

We proposed `MemeQA`, a novel dataset of multiple-choice questions that holistically evaluates models in their meme understanding capabilities. The design of question types was inspired by the human meme comprehension process. Extensive evaluation on six popular LMMs showed that `MemeQA` is very challenging, particularly when "None of the above" was introduced as an option. A closer analysis of the models revealed they usually failed to go beyond the superficial tone of the meme to reason more deeply about its implications. As such, we believe that `MemeQA` presents new opportunities for researchers as it could facilitate the development of stronger models for meme understanding and enable applications in online communication. Future work includes expanding `MemeQA` to include non-English memes, as well as analyzing potential cultural or contextual biases in meme selection.

Acknowledgments

We thank the three anonymous reviewers for their helpful comments on an earlier draft of the paper. We also thank Lavina Upendram and Rayeed Zarif for demonstrating human performance on `MemeQA`.

Limitations

Our resource is based on English memes in social media, thus not covering other languages and cultures. This may further the gap between high-resourced and low-resourced languages in NLP. However, computational meme processing research is still in its infancy, and researchers have been welcoming any annotated corpora that could advance the computational study of any aspects of meme understanding. Therefore, we believe MemeQA is still a valuable contribution to the development of the field. Furthermore, we do believe the methodology presented in this paper are applicable other languages. We hope that our findings will inspire researchers in other languages to improve MemeMQA in their own languages.

Ethical Considerations

Misuse against free speech MemeMQA models can be used to process memes at scale. It is possible for ill-intentioned actors to use this technology to further suppress unwanted opinions expressed by detecting those via memes. On the other hand, this field of research has also long been motivated for the good purposes. MemeMQA models can also be used to expose propagandist contents (e.g., via Intent ID questions). Furthermore, MemeMQA can enhance online safety, bridge cultural gaps, and help visually impaired people see the world, etc. So these technologies are always double-edged swords that should be used with care.

Terms of use This dataset is consistent with the terms of use and the intellectual property and privacy right of people with SemEval2021 Task 6 (Dimitrov et al., 2021). Instead of redistributing the original images, we refer users to the original data repository for access. There is nothing about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses.

Steps taken to protect annotators from harmful content All annotators were provided with a thorough instructional training session in which they were instructed on how to annotate the data and how to go about the whole task. During training, annotators were shown the types of memes that they will work with so that they have an idea of the dataset's nature. The annotators have full autonomy to withdraw from the project at their own judgment. They also gave consent for the collected data to be used for research purposes. All personally identifiable information was removed from the released data. See Appendix A for more details on annotator treatment.

References

Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. 2024. MemeMQA: Multimodal question answering for memes via rationale-based inferencing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5042–5078, Bangkok, Thailand. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, Santiago, Chile.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "Subverting the Jewtocracy": Online Antisemitism Detection Using Multimodal Deep Learning. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, pages 148–157, New York, New York, USA. Association for Computing Machinery.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25:70:1–70:53.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36*, New Orleans, Louisiana, USA.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2024. Data filtering networks. In *Proceedings of the Twelfth International Conference on Learning Representations*, Vienna, Austria.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. EVA: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, Vancouver, British Columbia, Canada.

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)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6325–6334, Honolulu, Hawaii, USA.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the Tenth International Conference on Learning Representations*, Virtual.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, Long Beach, California, USA.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1997, Honolulu, Kawaii, USA.

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*, volume 33, pages 2611–2624. Curran Associates, Inc.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742, Honolulu, Hawaii, USA.

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204, Long Beach, California, USA.

Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 2021. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 201–206, Online. Association for Computational Linguistics.

Khoi P. N. Nguyen and Vincent Ng. 2024. Computational meme understanding: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21251–21267, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jeongsik Park, Khoi P. N. Nguyen, Terrence Li, Suyesh Shrestha, Megan Kim Vu, Jerry Yining Wang, and Vincent Ng. 2024. MemeIntent: Benchmarking intent description generation for memes. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 631–643, Kyoto, Japan. Association for Computational Linguistics.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qwen Team. 2024. QVQ: To see the world with wisdom.

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*, pages 8748–8763. PMLR.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring Models and Data for Image Question Answering. In *Advances in Neural Information Processing Systems 28*, pages 2953–2961, Montreal, Quebec, Canada. Curran Associates, Inc.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020a. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020b. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. FVQA: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2017. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 1290–1996, Melbourne, Australia.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Annotator Details

Recruitment We recruited undergraduate and graduate students in our institution for the three annotation tasks: interpretation paragraphs (Group 1), human verification of questions (Group 2), and human performance (Group 3). All candidates were assessed based on their performance in doing several sample annotation tasks. Eventually, Group 1, Group 2, and Group 3 had five, eleven, and two members, respectively. The students are from the US, India, China, and Turkey.

Compensation The students participated in this project as part of the "Undergraduate Research in Computer Science" course they signed up for, during which they acquired experience and skills involving data annotation and model training. No additional compensation was thus provided to them.

Training For Group 1, group meetings were held every two weeks to review the annotated paragraphs and discuss ambiguous memes. For Group 2, every two weeks, annotators received written feedback from the second author on 10 randomly sampled questions. Group 3 did not require training as the task of answering multiple-choice questions is easy to understand.

Human performance To measure human performance, each question in the test set of $None^{-}$ and $None^{+}$ was randomly assigned to one annotator among the two in Group 3.

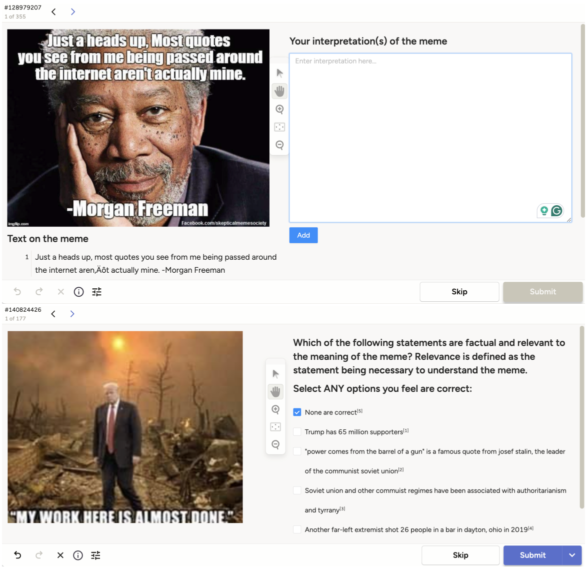


Figure 3: **Annotation interfaces** for interpretation paragraphs (upper) and question verification (lower)

B Annotation Guidelines

For writing interpretation paragraphs (Section 3.3.1), Table 5 presents our full annotation guidelines. For human verification of questions (Section 3.4) and obtaining human performance, we simply presented the questions to the annotators and asked for their answers to the actual questions. Later we checked if their answers matched with the one extracted from the paragraph. The annotation interfaces, shown in Figure 3, were built using Label Studio (Tkachenko et al., 2020-2025).

C Adversarial Filtering

This section gives more details about the Adversarial Filtering (AF) algorithm. Figure 4 illustrates how the generator and discriminator collaboratively generate challenging distractors. Below we describe the rationales behind our model choices for the generator and discriminator.

Discriminator model In AF, the difficulty of the questions is heavily influenced by the performance of the discriminator model. Therefore, we first selected four best open-sourced vision-language models available to us and evaluated them on the initial set of questions. The models are: InstructBLIP2 (Dai et al., 2023), BLIP2 (Li et al., 2023), LLaVA 1.5 (Liu et al., 2024), and Qwen-2 VL (Wang et al., 2024). The best performing model among all, Qwen-2 VL, was chosen as the discriminator model. The prompts used for the discriminator model are shown in Appendix D.

#	Guideline
1	**Overview:** You are asked to annotate a paragraph recording the interpretation process for a meme. In other words, inputs are a meme and output is the reasoning process in paragraph form.
2	**Premises:** The paragraph starts with the surface visual and textual information on the meme.
3	**Derivations:** From those premises, derive higher-level statements about the meme's meaning.
4	**Background knowledge:** When the interpretation involves some background knowledge (i.e., contextual information that is not presented on the meme), explicitly state them in your writing.
5	**Intent:** End the paragraph with the intent of the meme. The intent must be written in the form of "The meme [action] [social targets] ...", where **social targets** are the entities that the meme is discussing, and **action** is what the meme does to such targets.

Table 5: **Annotation guidelines for reasoning process paragraphs**.

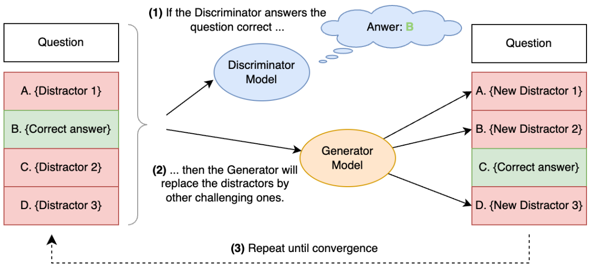


Figure 4: **Adversarial Filtering for creating challenging multiple-choice questions.** Distractors are created by the generator. Any questions the discriminator fails to correctly answer are deemed difficult and kept as part of the dataset. The correctly answered questions are passed back to the generator to regenerate new, more challenging distractors. This cycle repeats for a maximum number of iterations, or until the discriminator's performance converges.

Generator model To generate distractors, we looked for a model that can effectively give distractors that only differ in one or two words from the correct answers. Such distractors should still look relevant to the original meme, but have different or opposite meaning to the correct answer. Observing that this requires excellent language capability, we used the best language model at the time of experiment, Llama-3.1-8B-Instruct (Dubey et al., 2024). As such, to generate new distractors, we fed into this model (1) the context of the meme (including annotations about the scene, the present text, the relevant background knowledge, and the intent), (2) the current question, (3) the correct answers, and (4) all the distractors in previous AF iterations. We then tasked it with generating new challenging distractors. The prompts used for the generator are shown in Appendix E.

D Prompts for Discriminator

Below is the prompt for the discriminator in questions with exactly one correct answer:

```
You are given a meme.
Answer the following question by writing ONLY
one letter A, B, C, or D. DO NOT write anything
else. ONLY write the letter of the correct
answer.
## Question: ...
## Options:
(A) <Option 1>
(B) <Option 2>
(C) <Option 3>
(D) <Option 4>
## Answer:
```

For questions with possibly multiple correct answers, the prompt is as follows:

```
You are given a meme.
Answer the following question by selecting ALL
the correct options and write their letters
consecutively in alphabetical order, such as
'ACD' or 'B'. Write 'N' if none of the options
are correct. DO NOT write anything else. ONLY
write the letters of the correct answers or 'N'.
Remember that you can select multiple options.
## Question: ...
## Options:
(A) <Option 1>
(B) <Option 2>
(C) <Option 3>
(D) <Option 4>
## Answer:
```

For questions with possibly multiple correct answers and the last option is "None of the above", the prompt is as follows:

```
You are given a meme.
Answer the following question by selecting ALL
the correct options and write their letters
consecutively in alphabetical order, such as
'ACD' or 'B'. DO NOT write anything else.
ONLY write the letters of the correct answers.
Remember that you can select multiple options.
## Question: ...
## Options:
(A) <Option 1>
(B) <Option 2>
(C) <Option 3>
(D) None of the above
## Answer:
```

E Prompts for Generator

Each question type requires a unique prompt for generating distractors. This section shows the details of the prompts used.

Below is the generator's prompt for Intent/Derivation Completion question types.

```
You are given a meme as follows.
The meme is composed of the following images:
<Image caption>
The meme contains the following text: <Text>
List 3 words or phrases that are the most
sensible to be filled in the blank of
the following sentence: 'The meme supports
Trump and ____ that gun laws should be less
restrictive'.
The words or phrases must have OPPOSITE or
IRRELEVANT meaning from '<Option 3>'. Also,
don't use the following words or phrases:
<Old distractor 1>, <Old distractor 2>, <Old
distractor 3> Answer by listing the words or
phrases separated by commas, and write NOTHING
ELSE. Remember, write NOTHING ELSE but the 3
things.
```

Below is the generator's prompt for the Derivation ID question type.

```
 You are given a meme as follows.
The meme is composed of the following images:
'<Image caption>'
The meme contains the following text: '<Text>'
Someone thinks the following statements can be
derived from the meme:
- <Option 3>
- <Option 4>
List 2 other statements that look derivable
from the meme but are actually wrong. The new
statements must have OPPOSITE or IRRELEVANT
meaning from the original statements. Also,
don't repeat the following sentences:
- <Old distractor 1>
- <Old distractor 2>
- <Old distractor 3>
Answer by listing each statement as a sentence
on one line, and write NOTHING ELSE. Remember,
write NOTHING ELSE but the 2 new statements.
```

Below is the generator's prompt for the Intent ID question type.

```
 You are given a meme as follows.
The meme is composed of the following images:
<Image caption>
The meme contains the following text: <Text>
Someone thinks the meme's intent is that
'<Option 3>'.
List 3 other possible intents of the meme.
The new intents must have OPPOSITE or
IRRELEVANT meaning from the original intent.
Also, don't repeat the following intents:
- <Old distractor 1>
- <Old distractor 2>
- <Old distractor 3>
Answer by listing each intent on one line, and
write NOTHING ELSE. Remember, write NOTHING
ELSE but the 3 new sentences.
```

Below is the generator's prompt for the Visual ID question type.

```
 You are given a meme as follows.
The meme is composed of the following images:
'<Image caption>'
The meme contains the following text: '<Text>'
Someone thinks the following details are
visually visible on the meme and are important
to understand its meaning:
- <Option 3>
- <Option 4>
List 2 other statements that seem to be from
the meme but are actually not. The new
statements must have OPPOSITE or IRRELEVANT
meaning from the original statements. Also,
don't repeat the following sentences:
- <Old distractor 1>
- <Old distractor 2>
- <Old distractor 3>
Answer by listing each statement as a sentence
on one line, and write NOTHING ELSE. Remember,
write NOTHING ELSE but the 2 new statements.
```

Below is an example of the generator's prompt for the Background Knowledge ID question type where the question has two correct answers (Options 3 and 4) and two distractors.

```
 You are given a meme as follows.
The meme is composed of the following images:
'<Image caption>'
The meme contains the following text: '<Text>'
Someone thinks the following are relevant
facts that need to be known to understand the
meme:
- <Option 3>
- <Option 4>
List 2 other statements that seem to be
both factual and relevant to the meme but
are actually not. The new statements must
have OPPOSITE or IRRELEVANT meaning from the
original statements. It can be a non-factual
statements, or a factual statement that is not
relevant to the meme. Also, don't repeat the
following sentences:
- <Old distractor 1>
- <Old distractor 2>
- <Old distractor 3>
Answer by listing each statement as a sentence
on one line, and write NOTHING ELSE. Remember,
write NOTHING ELSE but the 2 new statements.
```

For the remaining question types (e.g., Action ID, Target ID, etc.), distractors are randomly sampled from other memes or clusters. Finally, Sentiment ID questions have a fixed set of options, i.e., POSITIVE, NEGATIVE, and NEUTRAL.

F On QO Pairs where Llama Answered Correctly

Recall that those questions where Llama answered correctly were removed from the original QO sets to obtain *None*$^-$. To verify this design choice, we compared the performance of the six models used in Section 4.1 on *None*$^-$ and on these questions.

Results are shown in Table 6. As can be seen,

	LLaVA	BLIP	IBLIP	Qwen	QvQ	GPT
Visual ID	31.6	85.5	61.2	97.4	85.8	77.0
Background ID	46.4	72.5	62.3	85.5	86.4	87.0
Target ID	23.4	49.4	44.2	68.8	66.2	63.6
Derivation ID	29.2	66.7	58.3	79.2	74.4	85.4
Target-Sentiment ID	2.1	51.1	46.8	70.2	56.1	59.6
Target-Action ID	1.8	42.9	33.9	76.8	81.2	58.9
Deriv. Completion	71.0	75.0	66.8	87.0	55.7	88.6
Intent Completion	70.0	83.8	69.2	89.0	63.0	92.2
Intent ID	48.4	28.1	25.0	82.8	76.7	85.9
Action ID	21.6	50.0	43.2	17.6	42.1	85.8
Sentiment ID	96.5	92.7	83.5	91.3	90.5	98.6
Macro Average	40.2	63.4	54.0	76.9	70.7	80.2

Table 6: Performances of zero-shot models on questions where Llama answered correctly.

these questions are much easier for all models. This shows that using Llama to identify trivial questions for removal is appropriate.

G More Question-Options Examples

To enable the reader to gain a deeper understanding of the challenges presented by `MemeQA`, we provide 11 examples, each illustrating one question type, in Figure 5.

H Model Overview

This section describes the state-of-the-art LMMs selected for our evaluation.

Qwen Qwen2-VL-7B-Instruct[14] (Wang et al., 2024) follows the common approach in vision-language models: *visual encoder* $\rightarrow$ *cross-modal connector* $\rightarrow$ *LLM*. Innovations here include "Naive Dynamic Resolution" for flexibly fine-grained visual processing and "Multimodal Rotary Position Embedding" for effective modality fusion. Its vision encoder and LLM were initialized from Data Filtering Network's ViT (Fang et al., 2024) and Qwen2 (Yang et al., 2024), respectively. It was trained via three stages with 1.4 trillion tokens.

BLIP2 BLIP2-Flan-T5-xl[15] (Li et al., 2023) is the first model that employs Querying Transformer (Q-Former), which is a type of cross-modal connector. The authors only trained the Q-Former and froze both the vision encoder and the LLM, thus being much more efficient than fellow models. This model variant uses ViT-g/14 from EVA-CLIP (Fang et al., 2023) as the vision encoder and Flan-T5-xl (Chung et al., 2024) as the LLM.

InstructBLIP InstructBLIP-Vicuna-7B[16] (Dai et al., 2023) extends BLIP2 and adds instruction-aware Query Transformer, which extracts informative features tailored to the given instruction. While the vision encoder is still ViT-g/14, the LLM is Vicuna-7B (Chiang et al., 2023). It was trained on 13 held-in datasets and tested on 13 held-out ones.

LLaVA LLaVA-v1.5[17] (Liu et al., 2024) uses CLIP-ViT-L-336px as the vision encoder and Vicuna v1.5 13B as the LLM. It was pre-trained on 500K image-text pairs before being fine-tuned on instruction and academic-oriented data.

QVQ QvQ-72B-Preview[18] (Qwen Team, 2024) is a multimodal reasoning model, extending the Qwen2-VL-72B architecture. It was optimized for "visual understanding and complex problem-solving", which are much emphasized competencies in meme understanding. The model outperformed GPT-4o in MMMU[19] and math-related benchmarks.

GPT GPT-4o (OpenAI et al., 2024) is OpenAI's first unified multimodal model capable of processing text, images, and audio in a single neural network. It was trained end-to-end across modalities, and was attributed for its high speed and performance. It achieved state-of-the-art results in multilingual understanding, vision, and audio tasks, outperforming GPT-4 Turbo.

I Examples of Errors

Figures 6 and 7 illustrate some of the questions where fine-tuned Qwen answered incorrectly.

[14] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

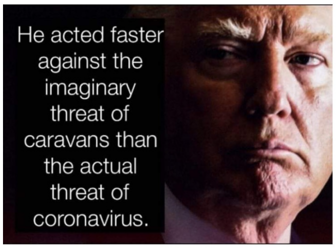
[15] https://huggingface.co/Salesforce/blip2-flan-t5-xl

[16] https://huggingface.co/Salesforce/instructblip-vicuna-7b

[17] https://huggingface.co/llava-hf/llava-1.5-7b-hf

[18] https://huggingface.co/Qwen/QVQ-72B-Preview

[19] https://mmmu-benchmark.github.io/



Background Knowledge ID

Which of the following statements are factual and relevant to the meaning of the meme? Relevance is defined as the statement being necessary to understand the meme. Select all that applies.

(A) Chris wallace moderated the 2020 presidential debate
(B) People believe trump did not handle the coronavirus pandemic properly and with enough seriousness
(C) During his presidency, trump suggested that undocumented central american immigrants would cross the southern border of the united states illegally in caravans, despite there being no evidence to support these claims
(D) The 2020 presidential debate was infamous for its lack of etiquette
Answer: B, C

Target-Action ID

Select all correct statements about this meme. Select all that applies.

(A) The meme supports Trump
(B) The meme criticizes Trump
(C) The meme discourages liberal media
(D) The meme praises people that oppose to wearing masks
Answer: B

Intent Completion

Fill in the blank to complete the intent of the meme: The meme criticizes Trump for ____ more about the imaginary threat of immigrants rather than COVID-19

(A) caring
(B) ignoring
(C) dismissing
(D) downplaying
Answer: A

Visual ID

Which of the following visual cue(s) are from this meme and necessary to derive its intent? Select all that applies.

(A) Protesters in California set fire to a courthouse, damaged a police station and assaulted officers after a peaceful demonstration resulted in a significant increase in police brutality and a decrease in community engagement.
(B) The image of Stalin with laser eyes is a symbol of a brutal and oppressive regime that crushes dissent and opposition.
(C) An image of stalin with laser eyes
(D) A peaceful protest in California resulted in a significant increase in crime rates and a decrease in community engagement due to the protesters' love of chaos and destruction.
Answer: C

Target-Sentiment ID

Select all correct statements about this meme. Select all that applies.

(A) The meme is negative towards protesters
(B) The meme is neutral towards the supreme court in the united states
(C) The meme is positive towards movement tracing of covid positive people
(D) The meme is negative towards people who want to ban glyphosate
Answer: A

Derivation Completion

Fill in the blank to complete a sentence that can be derived from the meme: ____ are causing people to be brainwashed

(A) junk food
(B) propaganda
(C) vaccines
(D) reality tv
Answer: C

Action ID

Fill in the blank: The meme ____ COVID-19 vaccine

(A) encourages (B) attacks (C) urges (D) asserts
Answer: B

Sentiment ID

What is the meme's sentiment towards COVID-19 vaccine?

(A) Positive (B) Neutral (C) Negative
Answer: C

Target ID

Select all targets of this meme. Targets are entities that the meme is discussing.

(A) the 2020 election
(B) republicans
(C) Biden
(D) the new york times
Answer: C

Intent ID

Which of the following is the final intent of the meme?

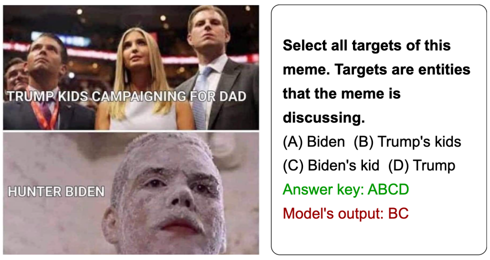
(A) insults that slavery is universally bad
(B) discourages for Child Sex Trafficking . In other words : He was arrested for Modern Day Slave Trading . I hope the irony has not been lost on you
(C) asserts a police photograph of Charles Wade
(D) accuses BLM of being a bad movement since it's founder was a criminal
Answer: D

Derivation ID

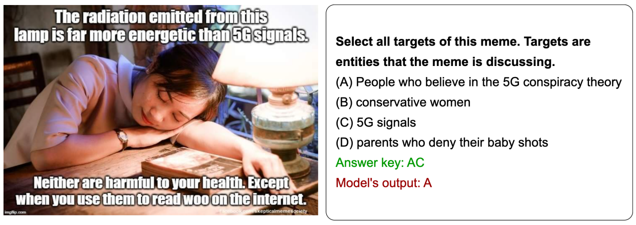
Which of the following sentence(s) can be derived from the meme? Select all that applies.

(A) This meme's topic is likely related to facebook's moderation policy
(B) The meme is trying to imply that facebook has a agenda to delete these memes
(C) Contradictory for facebook to delete memes which the author doubts are truely "hate speech" while being ok with child porn and graphic violence
(D) Trump in unfazed by his haters and has plans to be triumphant in his return
Answer: A, B, C

Figure 5: **Examples of questions in MemeQA**. Some examples share the same base meme on their left.

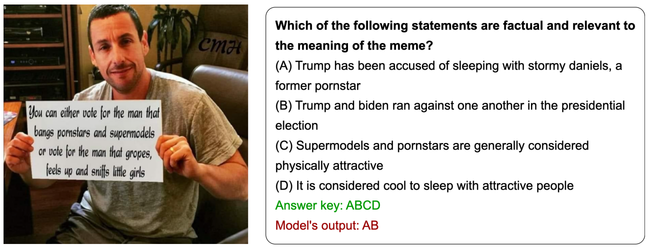


(a) A Target ID question example. Fine-tuned Qwen fails to identify the meme's implicit targets Trump and Biden, which it speaks on by extension of their children.

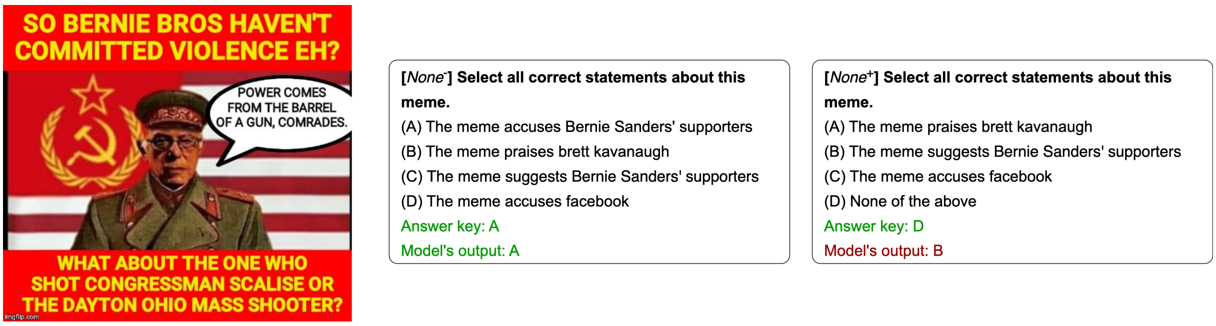


(b) A Target ID question example. Fine-tuned Qwen correctly identifies "People who believe in the 5G conspiracy theory" as a target, but fails to identify the related target of "5G signals" in general.

(c) An Action ID question example. The model attributes the aggressive wording, tone, and nature of the meme as a negative sentiment towards the target, guns.



(d) A Background ID question example. The model opts to not select the neutral options which appear argumentatively uncertain.



(e) A Target-Action ID question example. Here, the model recognizes that Bernie Sanders' supporters are the targets, but fails to recognize that the sentiment is negative.

Figure 6: **Example questions in which Qwen failed to answer correctly**.

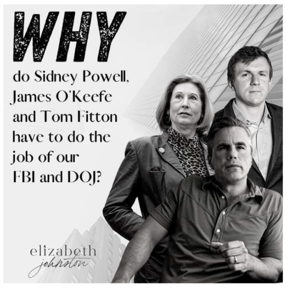
Sentiment ID

What is the meme's sentiment towards Sidney Powell?

(A) Positive (B) Negative (C) Neutral

Answer key: A

Model's output: C



Action ID

Fill in the blank: The meme ____ Sidney Powell

(A) criticizes (B) urges (C) asserts (D) encourages

Answer key: D

Model's output: A

Derivation ID

Which of the following sentence(s) can be derived from the meme?

(A) People should easily be able to breathe through medical masks

(B) The fbi and doj are not adaquately doing their jobs

(C) The meme praises sidney powell, james o'keefe, and tom fitton for doing the jobs of the fbi and doj

(D) tear down history refers to removing confederate statues

Answer key: BC

Model's output: B

Intent Completion

Fill in the blank to complete the intent of the meme: The meme attacks the FBI and DOJ for not doing their job and praises Sidney Powell, James O'Keefe, and Tom Fitton for being reliable and ____ the jobs of the FBI and DOJ

(A) criticizes

(B) urges

(C) asserts

(D) encourages

Answer key: A

Model's output: D

Figure 7: **Example meme in which Qwen failed to identify the correct sentiment for the social target.** Qwen incorrectly assumes the negative tone of the meme to be a negative sentiment towards Sydney Powell. Since the sentiment itself was mistaken, the meme makes a mistake at every different level of question answering that required understanding the sentiment.