MemeIntent: Benchmarking Intent Description Generation for Memes

Jeongsik Park[*] **Khoi P. N. Nguyen**[*] **Terrence Li** **Suyesh Shrestha**
Megan Kim Vu **Jerry Yining Wang** **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
{jeongsik.park,khoi.nguyen6,vince}@utdallas.edu

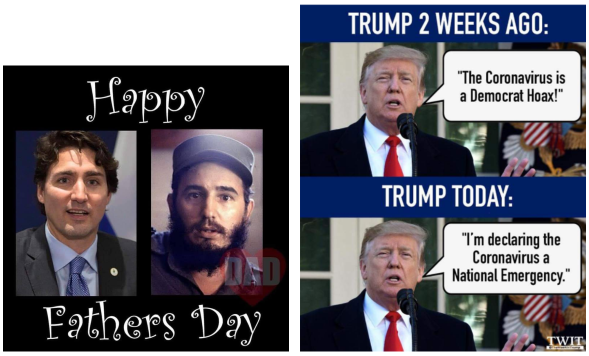
Abstract

While recent years have seen a surge of interest in the automatic processing of memes, much of the work in this area has focused on determining whether a meme contains malicious content. This paper proposes the new task of *intent description generation*: generating a description of the author's intentions when creating the meme. To stimulate future work on this task, we (1) annotated a corpus of memes with the intents being perceived by the reader as well as the background knowledge needed to infer the intents and (2) established baseline performance on the intent description generation task using state-of-the-art large language models. Our results suggest the importance of background knowledge retrieval in intent description generation for memes.

1 Introduction

Memes, which are "amateur media artifacts, extensively remixed and recirculated by different participants on social media networks" (Milner, 2012), are typically created with an intent to perform some "action" (Grundlingh, 2018). While many memes are intended to make a joke (where the author tries to make fun of a celebrity's weird accent, for instance), other memes may have malicious intentions. For instance, a meme author may seek to provoke fear (e.g., by conveying the message that vaccines contain microchips) or manipulate public opinions (e.g., by portraying Hillary Clinton as a corrupt politician before the 2016 presidential campaign with the goal of garnering support for Donald Trump). The core task in automated meme understanding, therefore, involves identifying the intent behind the creation of a meme.

In this paper, an *intent* is defined as an action that the meme author does via the meme. For example, the meme in Figure 1a intends to "mock Justin Trudeau as a communist for being similar to Fidel Castro", while Figure 1b "makes fun of Donald Trump's hypocrisy regarding his view on the severity of COVID-19". As shown in these examples, intents are best represented in textual form. Therefore, intent identification is naturally cast as a generation task, hence will be called *intent description generation*. Automatically generating the intent description of a meme is by no means an easy task, for at least two reasons:


Happy
Fathers Day
TRUMP 2 WEEKS AGO:
"The Coronavirus is a Democrat Hoax!"
TRUMP TODAY:
"I'm declaring the Coronavirus a National Emergency."

(a) (b)

Figure 1: Example memes from Dimitrov et al. (2021a) (left) and Sharma et al. (2023) (right).

First, background knowledge is often needed for proper interpretation of a meme. *Background knowledge* refers to the knowledge that is not present in the meme but is needed to recognize the intent when combined with the information that is explicitly stated in the meme. There are different kinds of background knowledge, including historical knowledge (e.g., '"Make America Great Again" is the slogan used by Trump in his presidential campaigns'), general political ideologies (e.g., 'progressives favor stricter gun control policies'), or knowledge of the meme culture (e.g., 'the meme template Drakeposting[1] funnily expresses an objection and an approval'), etc. For example, Figure 1a requires the knowledge that Castro is a communist

[*] These authors contributed equally to this work.

[1] https://knowyourmeme.com/memes/drakeposting

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 631–643
September 18–20, 2024. ©2024 Association for Computational Linguistics

leader and that there has been a fear of communism in the West since the Cold War. Figure 1b, on the other hand, does not require any special knowledge as all of the information necessary to understand the meme is presented at face value.

Second, in order to derive the intent, complex *inference mechanisms* may be needed to combine background knowledge with different pieces of information extracted from the image and text portions of the meme. Those "combination" steps reflect how a human thinks, such as based on logos (i.e., logical reasoning), ethos (i.e., speaker's authority), or pathos (i.e., emotional appeal). Some of these steps are not about logical reasoning, thus harder to automate (Mondorf and Plank, 2024).

As an example of such inference mechanisms, consider Figure 1a again. To arrive at the final intent, we first have to recognize that the person on the left is Justin Trudeau (Canada's prime minister) and the person on the right is Fidel Castro (Cuba's former leader). In addition, Trudeau has his mouth open whereas Castro has his mouth closed, which signifies that Trudeau is speaking and Castro is listening. When combining this information with the text "Happy Father's Day", one can infer that Trudeau either admits that Castro is his father or simply likes Castro enough to send his greetings to him. Then, combining the background knowledge that Castro was a staunch communist with the fact derived earlier that Trudeau admires Castro, the meme poster is trying to transfer the communist nature of Castro to Trudeau to damage Trudeau's reputation. Given the negative sentiment towards communism in the Western public, the final intent is thus "mocking Trudeau as a communist for being similar to Castro".

Intent description generation, though challenging, is a task whose solution has both practical and theoretical significance. From a practical perspective, knowledge of the intent being perceived through the meme could be useful for other meme-related processing tasks. For instance, knowing what the intent is could facilitate the determination of whether a meme contains harmful content (Pramanick et al., 2021a) or the detection of persuasion techniques (Dimitrov et al., 2021a). Theoretically speaking, being able to generate intents like humans requires that a machine read between the lines and achieve a deeper level of understanding of perceptual input, enabling machine perception to get one step closer to human perception.

Our contributions in this paper are four-fold. First, we propose the new task of intent description generation. Second, we construct the first benchmark for intent description generation, called *MemeIntent*, which shows the background knowledge required for each meme and its final intent(s). Third, we produce preliminary results on MemeIntent from two state-of-the-art language and vision-language models. Finally, based on the experimental results, we justify the need for more careful treatments of background knowledge in meme processing. To stimulate future research in intent description generation for memes, MemeIntent has been made publicly available[2].

The rest of this paper is organized as follows. Section 2 provides an overview of related work on automated meme processing. In Section 3, we describe our intent description generation benchmark, MemeIntent. To get an idea of how challenging intent description generation is, we conduct experiments on MemeIntent, discussing our experimental setup in Section 4 and showing preliminary evaluation results of two state-of-the-art large language models on MemeIntent in Section 5. Finally, we present our conclusions in Section 6.

2 Related Work

There has been a recent surge of interest in meme processing. Table 1 summarizes our survey on the tasks that have been proposed up to date for meme processing. These tasks can be classified into three groups: categorization, interpretation, and explanation, which will be described next.

2.1 Categorization Tasks

A growing effort has been made to assemble internet memes and categorically label them along various dimensions. These tasks can be broadly categorized into two groups.

The first group is composed of tasks that ask to classify malicious memes, including the offensive (Suryawanshi et al., 2020a), trolling (Suryawanshi et al., 2020b), hateful (Kiela et al., 2020), anti-semistic (Chandra et al., 2021), harmful (Pramanick et al., 2021b,c), and misogynous (Fersini et al., 2022). The second group is composed of tasks about detecting other aspects of memes such as persuasion techniques (Dimitrov et al., 2021b), figurative language (e.g., allusion, irony, sarcasm, contrast, etc.) (Liu et al., 2022), people's roles (e.g.,

[2] https://github.com/JeongSikPark1998/MemeIntent

Task	Dataset name	Topics	Size
Offensiveness Identification	MultiOFF (Suryawanshi et al., 2020a)	US Election	743
Troll Classification	TamilMemes (Suryawanshi et al., 2020b)	Tamil Memes	2,969
Hatefulness Detection	HatefulMemes (Kiela et al., 2020)	N/A	10K
Antisemitism	Jewtocracy (Chandra et al., 2021)	Antisemitism	3,102+3,509
Harm Detection	HarMeme (Pramanick et al., 2021b)	Covid	3544
Harm Detection	HARM-C&P (Pramanick et al., 2021c)	Covid, Politics	3,544; 3,552
Persuasion Technique Detection	SemEval-2021-T6 (Dimitrov et al., 2021b)	Mixed	950
Emotion Classification	Memotion (Sharma et al., 2020)	N/A	10K
Fine-grained Hatefulness Detection	WOAH-5 (Mathias et al., 2021)	N/A	10K
Misogyny Identification	MAMI (Fersini et al., 2022)	N/A	15K
Figurative Language Detection	FigMemes (Liu et al., 2022)	Politics	5,141
Role Labelling of Entities	HVVMemes (Sharma et al., 2022)	Covid, Politics	7K
Explaning Hate	HatReD (Hee et al., 2023)	N/A	3,228
Explaining Role of Entities	ExHVV (Sharma et al., 2023)	Covid, Politics	4,680
Meme Captioning	MemeCap (Hwang and Shwartz, 2023)	No offensive/sexual	6,387
Intent Description Generation	**MemeIntent**	**Mixed**	**950**

Table 1: **Tasks related to memes processing and associated benchmarks.** *Mixed* means *politics, vaccines, COVID-19, gender equality*. The three groups (separated by horizontal lines) are about categorization, explanation, and interpretation tasks, respectively.

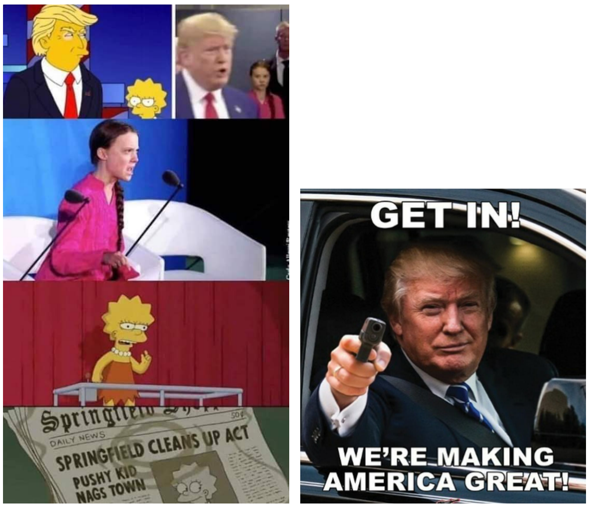
hero, villain, or victim) (Sharma et al., 2022), emotion (e.g., humor, sarcasm, motivation, or offensiveness) (Sharma et al., 2020), and attacked targets (e.g., religion, race, sex, nationality, or disability) (Mathias et al., 2021).

2.2 Interpretation Tasks

The second category of work on meme processing involves the relatively new task of meme interpretation, which involves generating text that captures the final meaning of a meme. Because intent description generation is a meme interpretation task, this category is the central interest of this paper.

To the best of our knowledge, meme interpretation has only been studied by Hwang and Shwartz (2023), who proposed the task of *meme captioning*, which means "describing the meaning of the meme". MemeCap, the dataset they produced as part of their work, contains memes images from Reddit. For each meme, they manually annotated the meme captions, the *literal captions* (i.e., the caption of the image excluding the text), and the *visual metaphors* (i.e., associations between entities on the meme and its actual target).

Intent description generation, while being a meme interpretation task, can be seen as the next level of meme captioning. Grundlingh (2018), a linguist studying memes, has argued that a meme, like an utterance, has both *illocutionary* and *perlocutionary* acts. In other words, a meme *says something* to *do something*. As such, while meme captioning is about what the meme is *saying* (the illocutionary act), intent description generation is concerned with what the meme is *doing* (the perlocutionary act).



(a) (b)

Figure 2: (a) A meme from MemeCap (Hwang and Shwartz, 2023), with title "Simpsons predicted it yet again". (b) A meme from Dimitrov et al. (2021a).

For example, for the meme in Figure 2a, the caption from MemeCap is "The Simpsons was correct about its use of Trump and Greta Thurnberg." However, the intent requires one reasoning step further to show that "the meme insults Greta Thurnberg as a pushy kid".

2.3 Explanation Tasks

The third category of work, like the second category, also involves generating text, but the focus here is generating a textual *explanation* of the message conveyed in a meme, as described below.

Sharma et al. (2023) defined the task of generating an explanation of *why* an entity plays the given role in the meme, where the role can be one of "hero", "villain", and "victim". Hee et al. (2023), on the other hand, addressed the task of explaining the reason why a meme is hateful toward a general target group.

Note that these explanation tasks are different from the interpretation tasks. The explanation tasks can be viewed as *constrained* interpretation tasks: in Sharma et al.'s task, both the entity and the role are given, whereas in Hee et al.'s task, the general target is given. In contrast, such constraints are not present in the interpretation task. As an example, consider the meme in Figure 1b again. The final intent that we would have produced for this meme (as the output of interpretation) is "The meme poster makes fun of Trump for the change in his recognition of the severity of the Coronavirus". However, when the target is constrained to be "the Democratic Party", the explanation would be "The Democratic Party is portrayed as a victim of false allegations", which is entirely different in meaning.

3 Benchmark Creation

In this section, we will show details about MemeIntent.

3.1 SemEval 2021 Task 6

We chose to annotate the meme collection of SemEval 2021 Task 6 (Dimitrov et al., 2021b). This dataset contains 950 memes, each of which has the image, the text extracted from the image, and the persuasion techniques used. Based on these memes, we built the MemeIntent benchmark. This dataset is favored due to its wide range of opinionated topics, including politics, vaccines, COVID-19, and gender equality. Moreover, each meme in this dataset cannot be properly interpreted without relying on both the visual cues and the textual cues. Therefore, the dataset asks for a 'true' multimodal processing ability in the solutions, as well as the capacity to retrieve relevant world knowledge to interpret contents on such topics.

3.2 Annotation Scheme

Our annotation scheme and procedure is illustrated in Figure 3, while further details are shown in Appendix B. For each meme, we annotate two fields:

- **Intent**: what the author might be trying to do through the meme (e.g., "The meme encourages people to get vaccinated because they are safe"). A meme can have multiple intents, representing its multiple meanings.

Annotation guideline
Intent: Write one sentence about what the author ultimately wants to do with the meme, as perceived by the annotator. This must be written in good English (complete sentence, with a period at the end) **BK:** Write the additional knowledge, besides the visible image and text, you needed to use to derive the intent. Examples are information about a public figure or an explanation for a related event.

Table 2: Annotation Guidelines.

- **Background knowledge (BK)**: the knowledge that is not present in the meme, but is needed to recognize the intent when combined with the information that is explicitly stated in the meme. That includes historical knowledge, general political ideologies, or knowledge of the meme culture, etc.

Note that, we allowed multiple intents in a meme to respect the subjectivity of meme interpretation. Built on top of theories from Bach and Harnish (1984), Grundlingh (2018) argued that a meme, like an utterance, could have more than one inference, which depends on the context of communication. Therefore, it is necessary to collect different intents perceived by different annotators, which is a natural consequence of the difference in their backgrounds and personalities. For example, consider the meme in Figure 2b. Depending on how one thinks about gun use, they may interpret the intent of the meme as "accusing Trump of being violent" or "praising Trump for his policies".

Additionally, the annotations include BK to provide extra guidance for learning algorithms in generating intents. As memes usually require a high level of cultural understanding (Milner, 2012), learning systems should be able to store and appropriately retrieve truthful knowledge about the world. The BK was collected to support testing such capabilities.

3.3 Annotation Procedure

Now, we seek to design an annotation procedure to label high quality intents and BKs. To control quality, dataset creators typically maintain *inter-annotator agreement* scores – the higher the score, the more reliable the dataset is (Artstein, 2017).

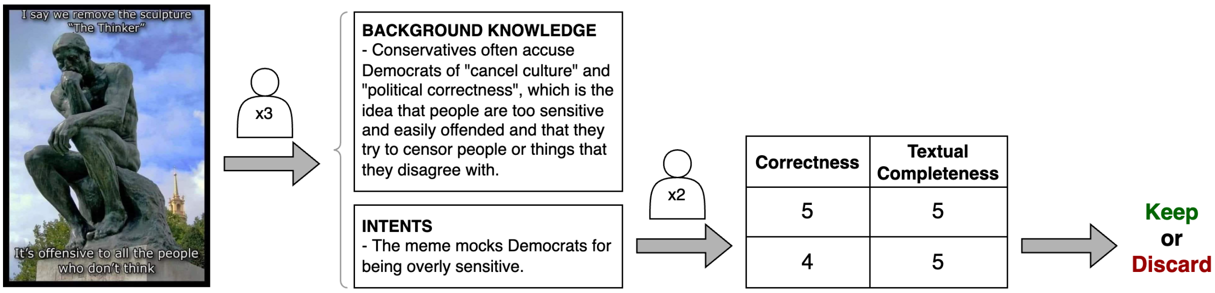

I say we remove the sculpture "The Thinker"
It's offensive to all the people who don't think
x3
BACKGROUND KNOWLEDGE
- Conservatives often accuse Democrats of "cancel culture" and "political correctness", which is the idea that people are too sensitive and easily offended and that they try to censor people or things that they disagree with.
INTENTS
- The meme mocks Democrats for being overly sensitive.
x2
Correctness
Textual Completeness
5
5
4
5
Keep or Discard

Figure 3: Annotation scheme and procedure of MemeIntent.

In order to obtain such scores, datasets must only involve *categorical* labels, which makes it easy to determine if two annotations agree. However, MemeIntent contains unstructured text annotations where there is no trivial way to check if two sentences (e.g., two intents) are identical. For reference, Hwang and Shwartz (2023) created the MemeCap dataset with only one round of annotations.

We questioned what a reliable procedure to annotate free-text data looks like. Towards that goal, we referred to the work of Wiegreffe and Marasovic (2021), who surveyed 65 papers that produce datasets for explainable NLP. For improving annotation quality, the authors advocate for "a two-stage COLLECT-AND-EDIT" approach, where annotations are first collected (stage 1), and then edited by a new annotator (stage 2). This approach is recommended due to its potential "to increase linguistic diversity via multiple annotators per-instance, reduce individual annotator biases, and perform quality control", and thus has been employed in constructing various free-text datasets (Parikh et al., 2020; Do et al., 2021; Li et al., 2018). Although COLLECT-AND-EDIT does not return any scores at the end, it has been shown to yield high quality annotations without further steps. Therefore, we used COLLECT-AND-EDIT as our annotation procedure.

We recruited five students in computer science, all of whom are native speakers of English, to label the dataset. All annotators went through roughly two hours of initial training and received regular feedback to adhere to our annotation guidelines. Each meme was annotated by three annotators in a sequential COLLECT-AND-EDIT manner: each of the three annotators, given the annotations of the previous person (which was initially empty), could add new intents, add new background knowledge sentences, or modify existing ones, based on their own interpretation of the meme. At the end of this stage, each meme had one or more unique intents, along with a list of BK sentences that is relevant to the understanding of the meme.

To control quality, we asked two reviewers to review each intent. To avoid biases, those reviewers were made sure to review memes that they did not annotate. The reviews were recorded as answers in the 5-point Likert scale[3] to two questions:

- *Correctness*: How much do you agree that this is the author's intent?
- *Textual Completeness*: How much do you agree that this sentence has complete English writing with good grammar?

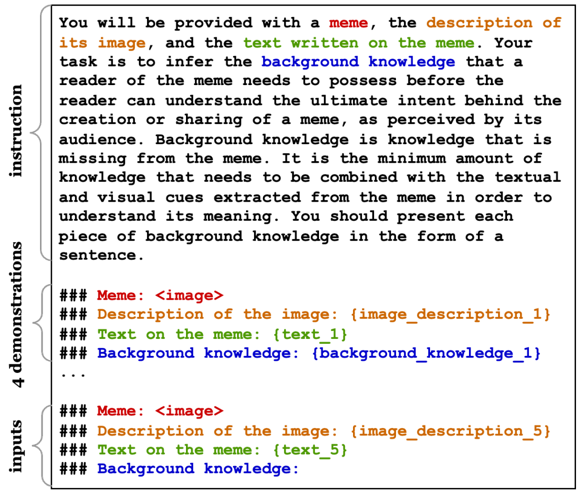
We removed all intents that received at least one correctness score lower than 4 from any of the reviewers. If no intents remained for a meme, we would restore the intents(s) with the highest average Correctness score.

Overall, 11.4% of memes in MemeIntent have more than one intent. The mean number of words in the intents is 10.6 ± 4.8. For background knowledge, the list of BK for each meme has an average of 1.7 ± 1.3 sentences. The mean correctness scores of the intents are 4.76 ± 0.37 (on the 1-5 Likert scale), while the mean Textual Completeness is 4.54 ± 0.71. These statistics provide suggestive evidence that the COLLECT-AND-EDIT procedure indeed produces high-quality annotations. Appendix A shows further qualitative analysis of the memes in MemeIntent, while Appendix C presents our ethics statement regarding the dataset.

4 Experimental Setup

With MemeIntent constructed, we now evaluate the performance of state-of-the-art models on intent generation for memes. The experiments were set up to answer the following research questions (RQs):

[3] The 5-point Likert scale is a numerical scale for recording agreement level, going from 1 (strongly disagree) to 5 (strongly agree).


instruction
4 demonstrations
inputs

```
You will be provided with a meme, the description of
its image, and the text written on the meme. Your
task is to infer the background knowledge that a
reader of the meme needs to possess before the
reader can understand the ultimate intent behind the
creation or sharing of a meme, as perceived by its
audience. Background knowledge is knowledge that is
missing from the meme. It is the minimum amount of
knowledge that needs to be combined with the textual
and visual cues extracted from the meme in order to
understand its meaning. You should present each
piece of background knowledge in the form of a
sentence.

### Meme: <image>
### Description of the image: {image_description_1}
### Text on the meme: {text_1}
### Background knowledge: {background_knowledge_1}
...

### Meme: <image>
### Description of the image: {image_description_5}
### Text on the meme: {text_5}
### Background knowledge:
```

Figure 4: **Prompt template used for BK generation in Llava few-shot learning setup.** For zero-shot learning, the demonstrations are omitted.

RQ1: What is the effect of adding background knowledge to the input on models' performance in intent description generation for memes?

RQ2: How do state-of-the-art models perform in intent description generation?

To that end, we designed a 3-factor experimental setup, consisting of 2 models × 2 learning setups × 3 input types, resulting in 12 settings. Finally, models' outputs from all settings are evaluated automatically and by humans. This section describes those factors and the evaluation metrics.

4.1 Three Input Types

We designed three input types that vary only in the treatment of background knowledge in the input.

In the first type, NoBK, only the *surface* information of the meme is fed to the model, including the meme itself[4], the extracted text on the meme, an automatically generated caption of the image without the text.

In the second type, AutoBK, we introduced automatically generated BK into the process. In particular, the BK is generated from a different model (BK generation model) in the same setting as the intent (i.e., the same model type and learning setup), using the prompt template in Figure 4. The BK is then fed to the intent description generation model along with surface information to generate the intents.

Finally, the HumanBK type replaces the generated BK with the BK annotated by humans. The prompt template for this input type is shown in Figure 5. This input setting is to gauge the upper bound on performance improvement given the human-annotated BK.

4.2 Two Models

Next, we selected two of the best open-sourced models for experiments.

Vision Language Model (Llava 1.6) Because intent description generation is a vision-language task, it is natural to use a vision-language model (VLM) to generate intent descriptions. In our experiments, we used Llava 1.6 (Liu et al., 2023), one of the most popular open-source vision language models with state-of-the-art performance in many visual reasoning tasks. We chose the variant `llava-v1.6-mistral-7b-hf`[5] for its superior performance among the Llana-Next variants with model size no more than 10B. It contains `Mistral-7B-Instruct-v0.2`[6] as the base language model and `CLIP-ViT-L-336px` (Radford et al., 2021) as the vision encoder.

Aided Large Language Model (Llava 1.6 + Llama 3) We also experimented with a pure large language model (LLM) with the aid of an image captioner. In other words, we employed a two-staged pipeline including (1) image captioning and (2) text-based intent description generation. For image captioning, we again leveraged Llava 1.6 to generate the captions for the memes. These captions, which describe the images themselves, would act as a proxy for the actual image to the LLM[7]. We then used Llama 3[8] to generate intents from the caption and other textual inputs. Llama 3, the most capable open-source LLM as of now (May 2024), has a decoder-only transformer architecture and was trained on 15 trillion tokens from public data. We used the variant `Meta-Llama-3-8B-Instruct`[9].

In all experiments, we kept the hyperparameters of the models the same with the default values and only tuned the `max_new_tokens`, setting its final value to 100 for intent description generation and

[4] Note that the meme is ignored by LMMs because it does not take images as input.

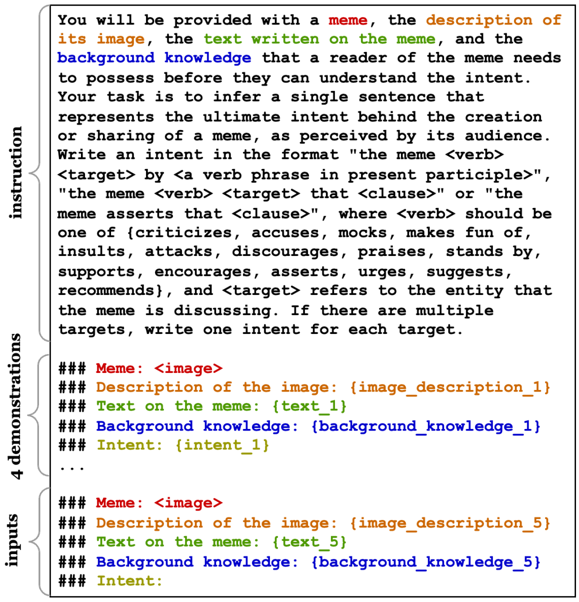
[5] https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

[6] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[7] To be fair with the VLM setting, we also feed the image caption to the VLM model.

[8] https://ai.meta.com/blog/meta-llama-3/

[9] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct


instruction
4 demonstrations
inputs

```
You will be provided with a meme, the description of
its image, the text written on the meme, and the
background knowledge that a reader of the meme needs
to possess before they can understand the intent.
Your task is to infer a single sentence that
represents the ultimate intent behind the creation
or sharing of a meme, as perceived by its audience.
Write an intent in the format "the meme <verb>
<target> by <a verb phrase in present participle>",
"the meme <verb> <target> that <clause>" or "the
meme asserts that <clause>", where <verb> should be
one of {criticizes, accuses, mocks, makes fun of,
insults, attacks, discourages, praises, stands by,
supports, encourages, asserts, urges, suggests,
recommends}, and <target> refers to the entity that
the meme is discussing. If there are multiple
targets, write one intent for each target.

### Meme: <image>
### Description of the image: {image_description_1}
### Text on the meme: {text_1}
### Background knowledge: {background_knowledge_1}
### Intent: {intent_1}
...

### Meme: <image>
### Description of the image: {image_description_5}
### Text on the meme: {text_5}
### Background knowledge: {background_knowledge_5}
### Intent:
```

Figure 5: **Prompt template used for intent description generation in Llava, with HumanBK input type, in few-shot learning setup.** For zero-shot learning, the demonstrations are omitted.

500 for background knowledge generation in both models.

4.3 Two Learning Setups

For each of the input types and models, we further experimented with two learning setups: zero-shot and few-shot learning (Mann et al., 2020)[10]. Overall, the two setups differ in the existence of the demonstrations. In the **zero-shot** setup, the *prompt* to the model includes an *instruction* and the *inputs* for the current meme. Meanwhile, in the **few-shot** setup, the prompt also includes 4 *demonstrations*, which are carefully crafted examples of input-output for 4 randomly chosen memes from MemeCap's test set. We illustrated the prompt used in few-shot learning in Figure 5.

4.4 Evaluation Metrics

To automatically evaluate model-outputted intents, we employed four metrics that are commonly used for text generation tasks, namely BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), BERT-F1 (Zhang et al., 2020), and SelfCheckGPT-NLI (Manakul et al., 2023). When there are multiple ground-truth intents, we took the maximum (i.e., best) of the scores when comparing the generated intent with the ground truths[11]. When making comparisons between settings on a metric, we performed the two-sided T-test with significant level $\alpha = 0.05$. Finally, we conducted human evaluation on the outputs of some selected settings to verify observations from automatic evaluation.

Model	Setup	BK	Metrics			
			BLEU	ROU.	BERT	Self.
Llama	zero-shot	No	0.011	0.243	0.89	0.354
		Auto	0.006	0.214	0.884	0.321
		Human	0.014	0.232	0.887	<u>0.475</u>
	few-shot	No	<u>0.015</u>	<u>0.287</u>	<u>0.899</u>	0.339
		Auto	0.013	0.282	0.899	0.34
		Human	**0.024**	**0.312**	**0.904**	0.439
Llava	zero-shot	No	0.006	0.231	0.885	0.352
		Auto	0.004	0.21	0.88	0.405
		Human	0.011	0.255	0.891	**0.506**
	few-shot	No	0.003	0.214	0.88	0.248
		Auto	0.002	0.134	0.867	0.459
		Human	0.003	0.225	0.883	0.313

Table 3: **Automatic Evaluation Results on Intent Description Generation.** For each metric, the overall best results are in **bold**, while the second best results are <u>underlined</u>. Abbreviations: Setup (learning setup), BK (input types), No (NoBK), Auto (AutoBK), Human (HumanBK), ROU (ROUGE-L), BERT (BERT-F1), Self (SelfCheckGPT-NLI).

5 Results and Discussion

5.1 Automatic Evaluation

Table 3 shows the automatic evaluation results on the generated intents of the two models across all learning setups and input types. Meanwhile, Table 4 reports the corresponding results for the generated BK in AutoBK settings by calculating similarity scores with the human-annotated background knowledge.

Input types (RQ1). The central observation from our experiment is that background knowledge is crucial to the performance of intent description generation. Specifically, for most settings, HumanBK input type gave the statistically highest performance across metrics. More interestingly, NoBK usually gave better performances than AutoBK, except in Llama few-shot, where there is no statistical significance. There are a few exceptions to this rule: SelfCheckGPT-NLI gave higher scores

[10] We attempted to fine-tune Llama model on the training data of MemeCap. However, the results turned out to be not as good as zero-shot and few-shot learning. Therefore, we omitted the result in this paper.

[11] For SelfCheckGPT-NLI, we assigned $score \leftarrow 1 - score$ to be consistent with the other metrics that the higher the score is, the closer the two pieces of text are.

Model	Setup	Metrics			
		BLEU	ROU.	BERT	Self.
Llama	zero-shot	0.003	0.073	0.827	0.331
	few-shot	**0.008**	**0.127**	**0.844**	0.294
Llava	zero-shot	0.003	0.086	0.83	0.384
	few-shot	0.002	0.072	0.821	**0.392**

Table 4: **Automatic Evaluation Results on Background Knowledge Generation in AutoBK setting.** For each metric, the overall best result is in **bold**, and the second best is underlined. Abbreviations: Setup (learning setup), ROU (ROUGE-L), BERT (BERT-F1), Self (SelfCheckGPT-NLI).

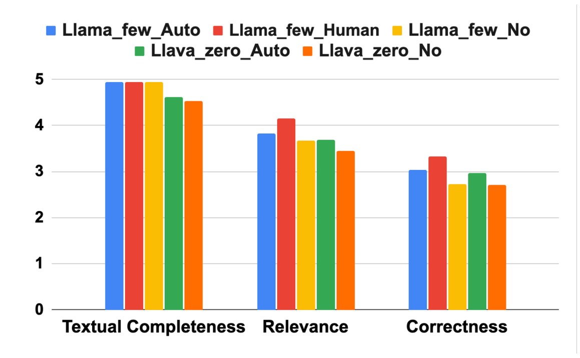


Figure 6: **Human Evaluation Results on Intent Description Generation.**

for AutoBK than NoBK where Llava was used, and NoBK sometimes outperformed HumanBK in Llama zero-shot experiments (on ROUGE-L and BERT-F1).

These results show that using human-annotated BK during the process produces better performance than no BK or auto-generated BK. We will further investigate these results via human evaluation (Section 5.2).

Performance on Background knowledge generation (RQ1). We take a closer look at the performance of models in BK generation. On BLEU, ROUGE-L, and BERT-F1, few-shot is better than zero-shot for Llama, and the opposite happens for Llava. However, SelfCheckGPT-NLI flips those results for both models.

We can connect these results with intent generation performance in AutoBK settings. In fact, among the AutoBK settings in Table 3, those with the best BK generation scores also score the highest on intent description generation. This suggests a correlation between the performance of BK generation and that of intent description generation across settings.

Models (RQ2). On BLEU, ROUGE-L, and BERT-F1, Llama (aided by Llava's image captions) outperformed Llava alone for most of the settings – across input types and learning setups. However, three over four metrics[12] gave a statistically higher score for Llava in experiments where zero-shot and NoBK input were used. Besides, SelfCheckGPT-NLI gave statistically higher scores for Llava when where AutoBK was used. Therefore, none of these models can entirely outperform the other across settings.

Learning setup (RQ2). In general, when Llama was used, few-shot was better than zero-shot. Conversely, the opposite happened when Llava was used. The superior performance of few-shot learning in Llama is aligned with the intuition that having demonstrations is useful. Meanwhile, Llava's inferior few-shot performance has been discussed by its authors that Llava was not explicitly trained to take multiple images as input[13].

The general trend above does have a few exceptions: SelfCheckGPT gave statistically higher scores for few-shot learning in Llava AutoBK, and zero-shot learning in Llava HumanBK.

5.2 Human Evaluation

For human evaluation, we evaluated the model outputs on 30 randomly chosen memes. Two annotators evaluated the outputs along three dimensions: **Textual Completeness** (i.e., How much do you agree that this sentence has complete English writing with good grammar?), **Relevance** (i.e., How relevant the sentence is to the topic of the meme?), and **Correctness** (i.e., How much do you agree that this is the author's intent?). Answers were recorded in the 5-point Likert scale, where higher scores indicate better quality.

To select settings for evaluation, we first focused on the effect of leveraging background knowledge to enhance the prediction of a meme's intent. Noticing the superior performance of Llama few-shot in most metrics, we selected its three settings – NoBK, AutoBK, and HumanBK – for human evaluation.

Next, the automatic evaluation results showed that in Llava NoBK and AutoBK settings, NoBK scored higher on BLEU, ROUGE, and BERT metrics; however, AutoBK scored higher on

[12] except BLEU which did not show statistical significance

[13] https://huggingface.co/docs/transformers/en/model_doc/llava#usage-tips

SelfCheckGPT-NLI. We know that NLI measures were used to assess the faithfulness of summarization, focusing on analyzing textual entailment between the context and the summary (Maynez et al., 2020). Given the contradiction between SelfCheckGPT-NLI and other metrics, we evaluated Llava zero-shot to determine whether SelfCheckGPT-NLI accurately captures the correctness between two sentences. We selected this setting since it demonstrated better scores among the two Llava settings.

Results are shown in Figure 6. Firstly, among the Llama few-shot settings, HumanBK significantly outperformed all other settings across all three metrics, which agrees with the automatic evaluation. Furthermore, while there was no statistical significance between AutoBK and NoBK in automatic evaluation, the human evaluation showed that AutoBK exhibits a higher performance than NoBK. These further demonstrate that a more sophisticated BK can influence the performance of intent generation.

Secondly, upon examining two outputs from the Llava model, it is observed that the performance of AutoBK surpasses that of NoBK across all three metrics in human evaluation. This is consistent with the SelfCheckGPT-NLI score, indicating that this metric effectively captures the correctness between the two sentences in our experiments.

6 Conclusion and Future Work

We examined the novel task of generating the description of intents in memes, specifically by (1) constructing MemeIntent, a benchmark of memes with intents, and background knowledge and (2) producing baseline results on our dataset against which future models can be compared. Our key findings suggest the importance of background knowledge treatments in intent description generation. To stimulate research on this task, we make our annotations publicly available.

Regarding future work, the experimental results w.r.t. the models and zero-shot vs. few-shot are inconclusive. Therefore, more experimentation is needed to get a clearer picture. Besides, we attempted to fine-tune Llama on the training set of MemeCap and test on MemeIntent, but the result was not good. This seems to be a failure to generalize from one meme interpretation dataset to another. Therefore, more efforts should be put into looking at the discrepancies between current datasets.

References

Ron Artstein. 2017. Inter-annotator Agreement. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 297–313. Springer Netherlands, Dordrecht.

Kent Bach and Robert M. Harnish. 1984. *Linguistic communication and speech acts*, 1.ed., 2. print edition. MIT Press, Cambridge, Mass.

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, NY, USA. Association for Computing Machinery.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021a. 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.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. 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.

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-snli-ve: Corrected visual-textual entailment with natural language explanations.

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.

Lezandra Grundlingh. 2018. Memes as speech acts. *Social Semiotics*, 28(2):147–168.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the Underlying Meaning of Multimodal Hateful Memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5995–6003, Macau, SAR China. International Joint Conferences on Artificial Intelligence Organization.

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.

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.

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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. 2023. Improved baselines with visual instruction tuning.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1.

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.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ryan M. Milner. 2012. *The World Made Meme: Discourse and Identity in Participatory Media*. Ph.D. thesis, University of Kansas.

Philipp Mondorf and Barbara Plank. 2024. Beyond accuracy: Evaluating the reasoning behavior of large language models – a survey.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. 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, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. 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. 2021c. 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.

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. ISSN: 2640-3498.

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, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy

Chakraborty. 2023. What Do You MEME? Generating Explanations for Visual Semantic Role Labelling in Memes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9763–9771.

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).

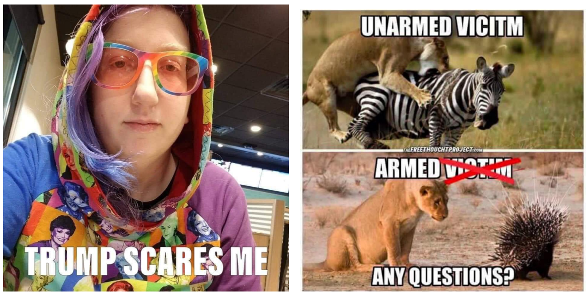
Sarah Wiegreffe and Ana Marasovic. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. 1.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. ArXiv:1904.09675 [cs].

A Challenges in Interpretation

We conduct a manual analysis of two memes taken from our dataset, with the goal of understanding the challenge of interpreting memes.

In Figure 7a, one first sees a person in a colorful outfit (via the jacket, the glasses, the hair). Then they may infer that this is an LGBT person. After that, they read the text saying '*Trump scares me*'. To connect that with the image, they further interpret the emotion of the LGBT person and recognize that they are apparently nonchalant. This is a word-face contrast, which suggests there is something wrong with one of the two. If the facial expression is 'wrong', one knows that the LGBT person may have a problem expressing fear, and the intent is to criticize Trump for being a scary person. But the other interpretation is more probable – that the words are wrong. Then, the LGBT person is actually not scared, thus being over-sensitive. This interpretation path might be triggered by the unconventional outlook of the person, which typically *scares* people, thus making them think the person in the meme is a bad person. Finally, because the liberals in the US support LGBT rights, this line of interpreting leads to a more significant intent, that is *mocking the liberals as over-sensitive and scary*.

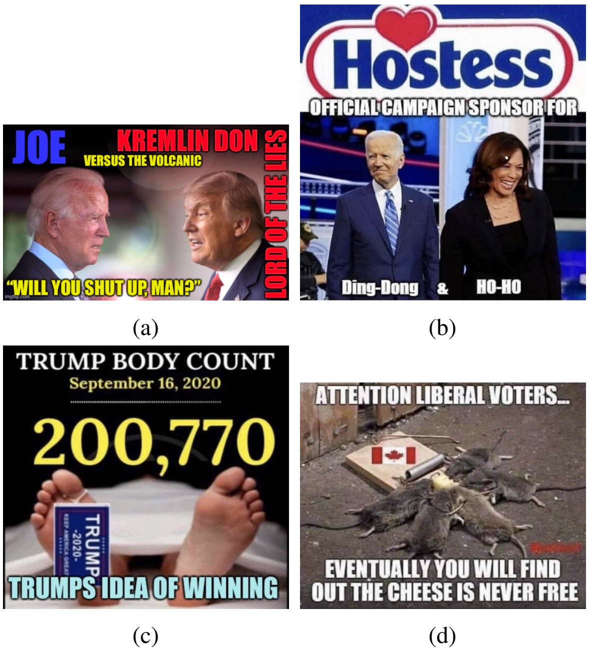


(a) (b)

Figure 7: Some example memes from the SemEval-2021 Task 6 dataset.

Consider another example in Figure 7b. A reader first sees a lion biting the zebra, about to kill it; then a lion looking at a hedgehog with upright quills, not sure if it is attacking or not. Then they read the first text saying '*unarmed victim*'. *Victim* refers to the zebra, and *unarmed* is a word for humans, so this is a metaphor for unarmed people being attacked. This line of thought triggers the reader's knowledge about the constant debate over gun control policies in the US. Next, the phrase '*armed victim*' with the word *victim* crossed suggests that the hedgehog, or metaphorically the gun owner, is safe. Finally, the rhetorical question '*Any questions?*' conveys that this is clear evidence so that *gun use should be allowed with no doubt*. Along this line of reasoning, the fact that the zebra is violently bitten provokes fear in the reader, which urges them to become the hedgehog and get a gun for self-defense.

In both of these examples, sophisticated logical (*logos*) and emotional (*pathos*) processes have triggered each other, forming the most probable interpretation path that leads to the recognition of the intent. This is a tricky task that only humans with appropriate knowledge and experience can perform. In fact, logical reasoning requires sufficient *background knowledge* to have the right facts to start with (e.g., how an LGBT person usually looks like, that liberals support LGBT rights, gun control is debated). Moreover, humans are also easily triggered by emotional stimuli (e.g., a strange look is scary, and safety is important). Those emotions are two-fold – they 'disambiguate' multiple possible



(a) (b)

(c) (d)

Figure 8: Example memes used in our annotation guidelines

interpretations via psychological biases, while also reinforcing the intent through pathos. This orchestrated effect can form a logical fallacy in disguise to achieve the final intent.

B Annotation Details

This section shows details about our annotation procedure.

B.1 Guidelines For Intents

This task introduces the notion of *intent*. An intent of a meme is what the meme author ultimately wants to do with the meme, perceived by the audience.

For example, the meme in Figure 8a has one final intent, which is: [The meme] praises Biden for being a better leader than Trump.

B.1.1 Frequently Asked Questions

Can I write new Intents? Yes, you should! If you think some intent is missing, add it.

How to write the Intent? Write an intent in format "[<verb1> <target1> <etc.>] x n", so that this sentence when being prefixed by “The meme” will form a grammatically correct sentence. For example, write "insults Trudeau for lying and insults anyone who believes in him as stupid".

What to do with multiple intents? Rank them by decreasing order of preference (i.e., from what you believe the most to the least).

B.2 Guidelines For Background Knowledge

For BK, summarize the key background information (skipping trivial knowledge). Also, write down (1) what you don’t know but seem important, and (2) what you are not sure if it’s right. Use question marks for those, e.g., "The place is in Cuba?".

For example:

- (Figure 8b) Hostess is a company that has products named Ding-Dong and Ho-Ho. Ding-Dong is also used to refer to someone who is slow. Ho-ho is also used to referred to someone who is lustful.
- (Figure 8c) A lot of Americans died during Covid 19, when Trump was presiding over the United States.
- (Figure 8d) This is about the Canadian election. Canada has 2 parties, one of which is the Liberal party.

C Ethics Statement

Broader implications. As mentioned before, the solution to the intent description generation task is of practical significance. From a practical perspective, knowledge of the message being conveyed in a meme could be useful for other meme-related processing tasks. For instance, knowing what the message is could facilitate the determination of whether a meme contains harmful content. Theoretically speaking, being able to generate messages like humans requires that a machine read between the lines and achieve a deeper level of understanding of perceptual input, enabling machine perception to get one step closer to human perception.

Ethical considerations. Having said that, we are all aware that some memes contain harmful content, so when our models are applied to these harmful memes, they will make an intent that is harmful explicitly. The resulting message could have a negative psychological impact on the users, especially if they are the target of the harmful content. Therefore, as with many other AI/NLP technologies, our models should be used with care. We should emphasize that our intent is to build models for generating the messages conveyed in memes, hoping that readers of memes will be less likely to be manipulated after understanding the messages being conveyed.

Human annotator information. All annotators were hired during Fall 2023 - Spring 2024 as student workers (15-20 hours/week) with full consent. All of the annotators were undergraduate and graduate students in computer science aged around 18-24. The group comprised both male and female students with members from Asian ethnicity, with fluent to native English level.

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 would 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.

Terms of use. This dataset is consistent with the terms of use and the intellectual property and privacy rights of people. There is nothing about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses.

Data distribution. We have open-sourced the data produced from this work. It is released on a GitHub repository with the MIT license.