Introduction:

The report we read and based our project on was on harmful and harmless meme detection. The report used new techniques and strategies to understand memes and the impact that memes have on the reader. The reporters used an LLM that would have a debate on whether the meme was harmful or harmless, then the LLM would judge the debate and take notes. Once the judge created notes a smaller LLM would read the notes and judge the outcome of the debate and that would be the final verdict. Omar, Luke, and I chose to focus on image processing, debate, and incorporate a multilingual aspect to the LLM. The report did not specify which language their new LLM knows and if it has a multilingual aspect. We as a group know that memes are not only in English, and we included to see if LLMs can do the same to Spanish memes.

Methodology/Approach:

Our first step was to collect memes. We collected a total of 100 memes. 25 harmless memes in English, 25 harmful memes in English, 25 harmless memes in Spanish, and 25 harmful memes in Spanish. To see how well the LLM can read images and determine if the memes are harmful or harmless, we sent the memes to native English and Spanish speakers. The native speakers saw the memes and wrote down if the memes are harmful or harmless and their reasoning. This gave us a basis to compare the LLM responses to human evaluations. This process also contributed to helping eliminate any bias we as a group would have to help the LLM scores.

Our next step is to accurately choose and use an LLM. After going through multiple LLMs and multiple versions we landed on Blip2. With Blip2 we ran through the memes and collected the data. We also ran the memes through ChatGPT-4. We want to compare the Human evaluations, Blip2 code, and one of the leaders of Artificial Intelligence today, ChatGPT-4. To compare Blip2 and ChatGPT-4 we asked the same following 10 questions:

1. Describe what is happening in the image?
2. Is there any weapons present in the picture?
3. Is there any sexual acts present in the picture?
4. Is there any hate symbols present in the picture?
5. Is there any curse words present in the picture?
6. Make a debate why the image is harmful?
7. Make a debate why the image is harmless?
8. Based on the debates in prompt6 and prompt7 is this image harmful or harmless?
9. From a scale of one to ten how harmful is the image?
10. Should this meme be wide spread?

The last step is to compare the evaluations. We are evaluating the following:

1. Blip2 English and Spanish memes evaluations.
2. ChatGPT-4 English and Spanish memes evaluations.

3. Blip2, ChatGPT-4, and human evaluations for English and Spanish memes.

Findings/Evaluation:

Blip2 English and Spanish meme evaluations:

Prompts:	English	Spanish
Describe what is happening in the image?	50/50	0/50
Is there any weapons present in the picture?	49/50	0/50
Is there any sexual acts present in the picture?	49/50	0/50
Is there any hate symbols present in the picture?	50/50	0/50
Is there any curse words present in the picture?	45/50	0/50
Make a debate why the image is harmful?	20/50	0/50
Make a debate why the image is harmless?	10/50	0/50
Based on the debates in prompt6 and prompt7 is this image harmful or harmless?	50/50	0/50
From a scale of one to ten how harmful is the image?	48/50	0/50
Should this meme be wide spread?	50/50	0/50

ChatGPT-4 English and Spanish memes evaluations:

Prompts:	English	Spanish
Describe what is happening in the image?	50/50	50/50
Is there any weapons present in the picture?	50/50	47/50
Is there any sexual acts present in the picture?	50/50	50/50

Is there any hate symbols present in the picture?	50/50	50/50
Is there any curse words present in the picture?	50/50	50/50
Make a debate why the image is harmful?	50/50	50/50
Make a debate why the image is harmless?	50/50	50/50
Based on the debates in prompt6 and prompt7 is this image harmful or harmless?	50/50	50/50
From a scale of one to ten how harmful is the image?	50/50	33/50
Should this meme be wide spread?	50/50	50/50

Blip2, ChatGPT-4, and human evaluations for English and Spanish memes

Criteria	Blip2 English memes	ChatGPT-4 English memes	Human Evaluation for English memes	Blip2 Spanish memes	ChatGPT-4 Spanish memes	Human Evaluation for Spanish memes
Describe what is happening in the image?	50/50	50/50	50/50	0/50	50/50	50/50
Able to make debates:	30/50	50/50	50/50	0/50	50/50	50/50
Able to make a harmful/harmless judgement	50/50	50/50	50/50	0/50	50/50	50/50

Artifact Discussion:

With the Blip2 English and Spanish meme evaluations table we can see the Blip2 model could not comprehend and determine all 100 memes. We also see the more complex the question is the more likely Blip2 fails to answer the question. The model failed completely when it came to the Spanish memes and could not complete any tasks. For the English memes Blip2 was the best at describing the images, determining the presence of hate symbols, if the meme is harmful or harmless prompt and answering the mass distribution prompt all at 50/50. An interesting outcome is prompt8 that says: “Based on the debates in prompt6 and prompt7 is this image harmful or harmless?”, but prompt6 is 20/50 and prompt7 is 10/50. This could be as Blip2 is not completely understanding the question and referencing the other prompts. If this was so we would see prompt8 with a much lower score.

In the ChatGPT-4 English and Spanish memes evaluations table we see that ChatGPT-4 is spot on with the English memes but has high scores but not as good on: “Is there any weapons present in the picture?” and “From a scale of one to ten how harmful is the image?”. For the prompt on the images, it is due to some weapons being cut off of the image, but a person who sees the memes can tell what the weapon is. The 1-10 prompt is a 33 as a lot of ChatGPT-4 determined hateful memes will get a score of under 5 just like the harmless memes. This could be as ChatGPT-4 does not know how culturally Significant the Spanish memes are as it knows for the English memes. ChatGPT-4 will read the meme and sometimes translate the text or even understand the meme well enough that it knows how to implement the Spanish texts in the explanation.

For the Blip2, ChatGPT-4, and human evaluations for English and Spanish memes table we see the comparison with Blip2, ChatGPT-4, and human evaluation. We see that Blip2 does a decent job with reading the images and giving a judgment but not in creating a debate. This only applies to English memes. Blip2 could not comprehend and give judgment on Spanish memes, receiving 0s in all categories. We see ChatGPT-4 is spot on and completes all categories with 50/50. ChatGPT-4 is much closer to human evaluation than Blip2, setting it as the more superior AI.

Future Work and Conclusion:

In conclusion we see ChatGPT-4 is a better LLM than Blip2 as it is closer to human evaluations, is multilingual, and can do the same functionalities as Blip2 at a better rate. ChatGPT2 can still use some more work when it comes to understanding other languages, different countries' cultures, and significant figures for different cultures. Blip2 has a long way to catch up as it is not multilingual and not as comprehensive to instructions/questions that rely on the previous answers.

For future works artificial intelligence needs to be more globalized and learn about other parts of the world. Both LLMs are proficient in United States culture, but not as proficient in Latino culture which negatively affects the accuracy of the answers. Making LLMs multilingual and multicultural memes from foreign countries can be categorized as harmless or harmful better. Another potential issue that would need to be looked into is languages that are not Latin based like Arabic, Mandarin, Urdu, and etc. As the letters are different and do not follow the Latin rules like English, Spanish, French.