Zero-Shot Protest Event Detection with Images and Large Multimodal Models

Cihang Xie*[1] and Zachary C. Steinert-Threlkeld†[2]

[1]University of California, Santa Cruz

[2]University of California, Los Angeles

June 28, 2025

*cixie@ucsc.edu

†zst@luskin.ucla.edu

Abstract

This paper evaluates the capabilities of large multimodal models (LMMs) in image analysis, focusing on the performance of Unum, LLaVA-7b, and LLaVA-13b. The analysis utilizes two datasets comprising 26,142 geolocated protest images and 1,000,000 random geolocated images from Twitter. The study assesses zero-shot classification abilities of these models through image captioning and direct image classification. Findings reveal that increased model size enhances narrative quality in captions and true negative identification but does not significantly improve the detection of true positives. Unum, despite its smaller scale, competes closely with larger models in identifying positive instances. The research highlights that while LLMs excel in generating detailed descriptions, their effectiveness in zero-shot image classification is limited. Further analysis using captions demonstrates more accurate classification, suggesting a potential approach for improved image analysis. Future research should explore ensemble coding techniques that leverage specific strengths of different models and investigate the trade-offs between model size and processing speed to optimize both efficiency and accuracy in practical applications.

Acknowledgments We thank Sharanya Majumder, Dilxat Muhtar, and Aryan Sunkersett for research assistance and Ethan Busby and Lisa Argyle for organizing the SPSA Generative LLMs conference within a conference.

1 Introduction

The emergence of large-language models (LLMs) has transformed computational analysis across multiple domains. Since the introduction of GPT-3 (Brown et al., 2020), these models have demonstrated great capabilities in understanding and generation of natural languages. The subsequent development of models such as GPT-4 (OpenAI, 2023) has established LLM as increasingly viable alternatives to traditional computational methods, particularly for tasks requiring contextual understanding. These advances have been especially notable in their few-shot and zero-shot learning capabilities (Wei et al., 2022), allowing them to adapt to new tasks with minimal or no additional training. Early transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) laid the groundwork for contextual understanding in natural language processing. However, the scale and capabilities of modern LLMs have expanded dramatically, as demonstrated by models such as PaLM-2 (Anil et al., 2023) and Claude (Anthropic, 2023). These models have shown remarkable abilities in understanding context, generating human-like responses, and adapting to specialized domains with minimal additional training.

This paper explores the capabilities of large multimodal models (LMMs) in processing and analyzing protest imagery and random images from social media platforms, comparing their performance with traditional methods. We examine a dataset comprising 26,142 geolocated protest images from Twitter and another set of 1,000,000 random geolocated images to assess the models' efficiency in zero-shot learning and image captioning. The models employed include variants of the Large Language and Vision Assistant (LLaVA) and Unum, each varying in complexity and parameter size. The analysis is conducted without any model fine-tuning to evaluate their out-of-the-box capabilities in a zero-shot scenario, focusing on the models' ability to generate descriptive captions and binary classifications for a variety of labeled images.

The results reveal that while LMMs show promising capabilities in recognizing and classifying elements within images accurately, there are notable challenges with specificity and

sensitivity, particularly when dealing with nuanced or less commonly depicted scenarios. In the protest image dataset, smaller models like Unum occasionally outperform larger counterparts by maintaining higher precision, though they struggle with recall. For random images, the LMMs frequently overgeneralize, suggesting a propensity for these models to 'see' events or elements that are not present, which is indicative of their current limitations in understanding context without specific training. The paper discusses these findings in detail, offering insights into the potential and limitations of using LMMs for automated image analysis in social and political research.

2 Background

The Use of LLMs in Text Analysis

Recent research has demonstrated how LLMs can be used in text analysis tasks. Gilardi, Alizadeh and Kubli (2023) conducted a study comparing ChatGPT's zero-shot performance against human crowd workers in text annotation tasks. Their findings revealed that ChatGPT significantly outperformed crowd-sourced human annotators on all tasks, achieving an average accuracy improvement of 25 percentage points in areas such as relevance assessment, position detection, topic classification and frame detection. This substantial improvement aligns with the findings of Wei et al. (2022), who demonstrated the effectiveness of LLM in few-shot learning scenarios in various NLP tasks.

Wang (2023) further advanced this field by demonstrating that fine-tuning pre-trained language models on small annotated datasets can exceed the performance of cross-domain classifiers trained on substantially larger datasets. Bisbee et al. (2024) introduced important caveats to these findings, as their research suggests that while LLMs like ChatGPT can effectively match human survey data at broader levels, they may exhibit significant discrepancies when analyzing specific subgroups and conditional relationships. Their observation that LLM responses often demonstrate less variation than human responses, particularly regarding racial and religious groups, might suggest that LLMs have a tendency to overgeneralize

and miss important nuances in responses across different demographic groups, potentially limiting their utility for fine-grained analysis of population subgroups and raising considerations about representation and bias in automated text analysis systems.

Applications of LLMs in Understanding Event Data

Research regarding the capabilities of LLMs in performing complex text analysis tasks has also expanded to include automated event data extraction and classification in an effort to understand how LLMs can be leveraged to categorize real-world events from textual descriptions with greater accuracy and nuance than traditional approaches. Halterman et al. (2023) demonstrated how LLMs surpass the limitations of dictionary-based approaches, which relied on exact pattern matching. Their implementation of a multi-stage LLM approach, utilizing different transformer models for distinct aspects of event extraction, proved particularly effective in handling complex cases that had previously challenged traditional systems. Their research demonstrated that LLMs excel at understanding contextual nuances that previous systems struggled with, such as distinguishing between phrases like "tanks rolled into town" indicating military action versus "aid convoys rolled into town" suggesting humanitarian assistance, without requiring explicit rules for each scenario. This ability to make human-like inferences about implicit information proved especially valuable for event detection – for instance, when a news article mentioned police dispersing protesters, their system could infer that police must have first mobilized to the scene, even if this was not explicitly stated in the text.

Du and Cardie (2021) introduced a framework that reconceptualized event extraction as a question-answering task. Their approach leverages BERT-based models to identify event triggers and extract arguments through natural language questions, eliminating the need for conventional entity recognition preprocessing. Their system achieved state-of-the-art results on the ACE 2005 dataset by implementing three increasingly sophisticated question generation strategies, ranging from simple argument role names to natural questions derived from annotation guidelines. This question-answering framework proved particularly effec-

tive at maintaining high precision while improving recall compared to baseline approaches, suggesting that this method helps with accurate argument boundary detection and enables transfer of knowledge between semantically related argument roles across different event types. Douglass et al. (2024) demonstrated that open-source LLMs running on consumer hardware could reliably extract detailed international crisis events from historical narratives. Using a variant of Meta's Llama 2 model, they achieved performance comparable to expert human coders on the complex International Crisis Behavior Events (ICBe) ontology without requiring fine-tuning or extensive prompt engineering. Their results demonstrate the feasibility of automating complex event coding tasks with current generation language models, indicating that sophisticated event extraction systems are now achievable without extensive computational resources.

Extension of LLM Capabilities to Visual Analysis and Multimodal Understanding

While LLMs are primarily used to process and analyze textual information, LMMs, or large multimodal models, which consider both visual and textual inputs simultaneously, have the potential to understand and analyze complex relationships between text and images. LMMs represent a significant advancement in artificial intelligence by enabling integrated analysis of multiple forms of data, allowing for more comprehensive understanding of content that combines visual and textual elements.

Lyu et al. (2023) evaluated GPT-4V's, a multimodal model that enables users to instruct GPT-4 to analyze image inputs provided by the users, capabilities as a social media analysis engine, demonstrating its effectiveness in multimodal understanding of social media content. Their research revealed GPT-4V's sophisticated abilities in sentiment analysis, hate speech detection, and ideology detection, while exhibiting notable cultural and contextual awareness in interpreting image-text interactions. Through a series of experiments, the authors found that GPT-4V could grasp subtle nuances in image-text relationships, including understanding strategic visual framing. For instance, the model demonstrated an ability to detect

underlying sentiments and ideological messaging through visual content analysis, such as recognizing how the emphasis on particular subjects in images could be used for strategic framing of political narratives.

Wu et al. (2024) provided an evaluation of GPT-4's zero-shot visual capabilities, examining both linguistic and direct visual processing abilities. Their findings complement earlier work by Li et al. (2023) on multimodal foundation models and extend the understanding of vision-language models established by Alayrac et al. (2022) with Flamingo. The authors' experiments revealed that GPT-4V performs comparably to state-of-the-art vision-language models like EVA-CLIP's ViT-E, but they also identified important limitations, such as the model's difficulty with temporal relationships and certain technical constraints related to image resolution and processing.

Methodology

Data Collection and Preprocessing

This paper presents analysis of two datasets. The first is 26,142 geolocated protest images collected from Twitter, originally compiled by Joo and Steinert-Threlkeld (2022) for their study of protest events. This dataset contains tweets and their associated images that span multiple political events and geographic regions, including protests in Catalonia, Spain (secessionist protests), Hong Kong (2014 protests against electoral system changes), Russia (2017 anticorruption protests), South Korea (2016-2017 protests against President Park Geun-hye) and Venezuela (2014-2015 protests). The data was collected through geolocated tweets from these protest periods, with images downloaded from tweets when they were both protests and contained an image. Each image in our dataset is associated with its original tweet ID and metadata, enabling both contextual analysis and cross-reference with existing labeled datasets. In this study, Joo and Steinert-Threlkeld developed image classifiers based on convolutional neural networks (CNNs) to detect custom labels from images of protesters shared on Twitter, aiming to understand how protests are framed on social media. The

original dataset included specialized annotations for key protest characteristics such as state violence, protester violence, police presence, and crowd size, among other variables, providing a baseline deep learning benchmark against which to evaluate the performance of the LLMs. Building on this foundation, our work explores how LLMs perform in comparison, particularly in their ability to analyze and classify protest imagery.

The second dataset consists of 1,000,000 random geolocated images from Twitter from September 2013. This dataset was created to analyze how multimodal LLMs evaluate images not necessarily of protest, which will enable a comparison of their general performance to performance on specific tasks.

For both datasets, no fine-tuning is used. All results therefore reflect performance in what computer scientists call zero-shot inference, labeling without specific customization for the labeling task.

Models

Three large multimodal models of different sizes are used to generate captions and responses. Choosing models with a range of sizes allows for the comparison of performance as model complexity increases, an important consideration given the compute intensive nature of even small multimodal LMM models. The smallest is UForm-Gen2-Qwen-500m, a combination of Qwen 1.5, an LLM from the Chinese search company Alibaba, and CLIP-like ViT-H/14 for image recognition and classification (Bai et al., 2023, Ilharco et al., 2021); the model has 1.2 billion parameters, 500 million of which are from Qwen.[1] The next two are different implementations of Large Language and Vision Assistant 1.6 (LLaVA), a leading multimodal LMM (Liu et al., 2024). Versions of the model with 7 billion and 13 billion parameters are used. For the rest of the paper, the models are labeled "Unum", "LLaVA-7b", and "LLaVA-13b".

For the two sets of images, the three LMMs were given the prompt, "Describe this image." This prompt generates a long caption describing the image. For the 26,142 protest images,

[1]Unum is an Armenian research lab.

additional prompts of the structure "Does this image contain <label>?" where <label>is one of the 12 classes from Steinert-Threlkeld, Chan and Joo (2022) are given to each LLM, e.g. "Does this image contain police?" For the 1,000,000 images, prompts of the structure "Does this image (show, make) <label>" were given where <label>is one of the 20 event types from the CAMEO ontology (Schrodt et al., 2008). "show" or "make" are used depending on which is grammatically correct. A different set of prompts were chosen to investigate the extent to which zero-shot learning can produce political event data, as the CAMEO ontology includes events of increasing severity, from making public statements (1) to protests (14) to engaging in unconvential mass violence (20).

While the original intent was for the models to provide simple binary "yes" or "no" responses, the actual outputs were more verbose and required additional processing to convert them into binary categorical variables. Table 1 shows an example of the responses from each of the three models. We developed a systematic approach to binary coding that involved using text matching techniques, primarily grep commands, to transform the multimodal LMM responses into standardized binary labels.

To determine if the different verbs in the prompt, "(make, show)" versus "contain", explain why the LLaVA models do not produce "yes" or "no" answers, we replaced the former with "contain". The same LLaVA behavior holds. The different structure of the text output is therefore due ot the model architecture, not prompt difference. Prompting LLaVA to explicitly provide "yes" or "no" answers is the only way to make those models consistently generate binary answers, in contrast to Unum. Without that instruction, the model responds to each implicit "yes" or "no" question with a description of the image, and current resource constraints prevent regenerating answers with a new prompt.

Figures 1-4 each show an image and the caption Unum, LLaVA-7b, and LLaVA-13b generates.[2] Figure 1 is of a political rally, Figure 2 a protest, Figure 3 of nature, and Figure e4 is religious. Unum produces less rich descriptions and appears to be less accurate, as can

[2]The newline characters are often generated; though meaningless, it is left in here intentionally.

Table 1: Frequency of Response Patterns for Police by Model Type

Variable	Response	LLaVA-13B	LLaVA-7B	Unum
police	No	0	0	2287
police	No, it does	0	0	9670
police	No, it doesn't	0	0	303
police	No, it doesn't.	0	0	23
police	No, the image	6307	3561	0
police	No, there are	142	0	792
police	No, this image	0	3	0
police	The image appears	12	47	0
police	The image contains	1	1	0
police	The image depicts	1	5	0
police	The image shows	1302	4543	0
police	The image you've	12	47	0
police	Yes	0	0	1167
police	Yes, the image	1601	1427	2
police	Yes, the man	0	0	1
police	Yes, the person	0	0	1
police	Yes, the scarf	0	0	1
police	Yes, there appears	2	0	0
police	Yes, there are	497	237	2187
police	Yes, there is	135	141	421
police	NA	16130	16130	9287

be seen in Figure 1. In the religion image, Figure 4, Unum accurately describes the scene and perceives spiritual overtones but misreads the scripture, whereas LLaVA-7b and LLaVA-13b correctly read the text, passage reference, and identify the passage as from the Bible. Across the four images, the difference in detail is greatest going from Unum to LLaVA-7b, but LLaVA-13b produces richer captions than LLaVA-7b. Overall, unum produces reasonably accurate descriptions. Those descriptions are less accurate and informative than the two LLaVA ones. LLaVA-7b is accurate, but LLaVA-13b pulls more detail from scenes and appears to understand intent better.

Figure 1: Example Political Rally Image and LMM Captions

Unum: A poster promoting a protest against the US government's actions in the Middle East. The poster features a photo of a man giving a speech at a podium, with the text "In the Middle East, a man is speaking to a crowd" and the image of a flag. The poster also mentions that the US government has been "abusing its power to bully and intimidate the people of the Middle East."

LLaVA-7b: The image is a graphic with a political message. It features a large portrait of a man with glasses, wearing a suit and tie, set against a backdrop that resembles a flag with red and white stripes. The man appears to be speaking or gesturing, as suggested by the position of his hands.
The text on the image is in Spanish and reads: "TENGAZA presente: Allegado a la moneda por la gente de Centro Derécha NO fue a votar ¡NO repetamos la historia!! El pueblo gileno cumplirá tu mandato." This translates to "Remember: The one accused of being on the currency by the people of Centro Derécha did not vote. Let's not repeat the history!! The people of Gileno will fulfill your mandate."
Below the portrait, there is a smaller image of a woman standing in front of a microphone, suggesting she is speaking at an event. The text below this image reads: "NO basta con escribir en Twitter o Facebook. Eve Lynn con nuestro voto defenderemos el futuro de nuestros hijos." This translates to "It's not enough to write on Twitter or Facebook. Eve Lynn with our vote will defend the future of our children."

LLaVA-13b: The image is a digital graphic with a political message. It features a photograph of a person standing in front of a large mural or banner with a portrait of a man in a suit and glasses, set against a red and white background that resembles the flag of the United States. The text on the image is in Spanish and includes a call to action, stating "¡NO repitamos la historia!" which translates to "Let's not repeat history!" The text also mentions a political figure by name and encourages people to vote for a specific candidate. The bottom of the image contains additional text that reads "NO basta con escribir en Twitter o Facebook" which translates to "It's not enough to write on Twitter or Facebook," and "CON NUESTRO VOTO DEFENDEREMOS EL FUTURO DE NUESTROS HIJOS" which translates to "With our vote, we will defend the future of our children." The image also includes a logo in the bottom left corner with the name "EVE LYN" and a slogan "MARÍLU." The style of the image is informative and appears to be designed for social media sharing, aiming to mobilize support for a political candidate.

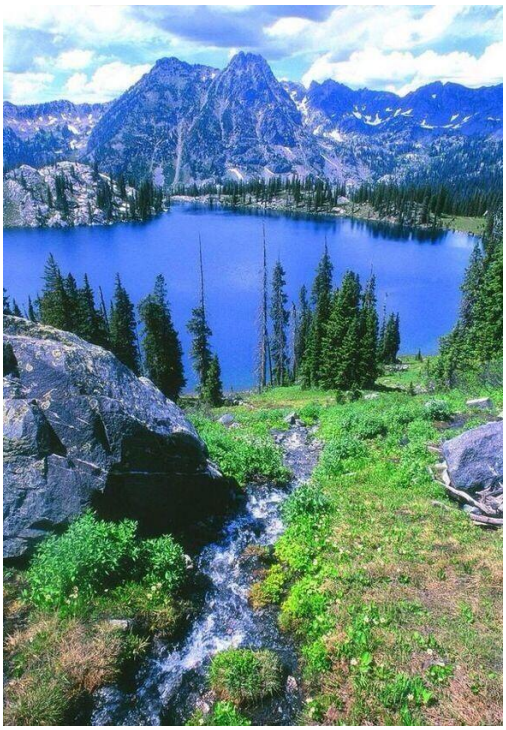
Figure 2: Example Protest Image and LMM Captions

Unum: The image captures a lively scene of a group of people, possibly a group of women, gathered on a truck. They are adorned in yellow shirts, and many of them are holding balloons. The balloons are scattered throughout the image, some closer to the truck and others further away. The truck itself is white and has a banner on it. The atmosphere appears to be joyful and celebratory, with people smiling and waving at the camera. The image does not contain any discernible text.

LLaVA-7b: The image captures a vibrant scene of a political rally. A group of people, dressed in yellow shirts and black pants, are seen celebrating and expressing their enthusiasm. They are standing on the back of a white truck, which is adorned with a large orange banner. The banner prominently displays the word V̈OTE; indicating the importance of the upcoming election. The people are holding balloons in shades of yellow and orange, adding to the festive atmosphere. The balloons are not just held, but they are being thrown into the air, creating a sense of movement and excitement.The backdrop of the image is a clear blue sky, suggesting that the rally is taking place outdoors. The overall mood of the image is one of joy and anticipation, as the people prepare to cast their votes and make their voices heard.

LLaVA-13b: The image captures a vibrant scene of a political rally. A group of people, predominantly women, are seen celebrating their cause. They are seated on the back of a white truck, which is adorned with a large yellow banner. The banner prominently displays the word V̈OTEïn bold black letters, indicating the rally's call to action. The women are dressed in bright yellow shirts, symbolizing their unity and support for their cause. They are holding yellow balloons, adding to the festive atmosphere of the rally. The truck is moving down a street, with buildings visible in the background, suggesting that the rally is taking place in an urban setting.The image is a snapshot of a moment of collective action and celebration, capturing the energy and enthusiasm of the participants. It's a testament to the power of peaceful protest and the importance of voting in democratic societies.

Figure 3: Example Nature Image and LMM Captions



Unum: The image captures a serene mountain landscape. Dominating the scene is a large blue lake, its surface reflecting the sky above. The lake is surrounded by a lush green forest, with trees of varying heights and colors. In the foreground, a small stream meanders through the grassy area, adding a touch of tranquility to the scene. The mountains in the background are covered in snow, their peaks reaching towards the sky. The sky is a clear blue with a few clouds scattered across it. The overall atmosphere is one of calm and natural beauty.

LLaVA-7b: The image captures the breathtaking view of Emerald Lake, a renowned landmark nestled in the heart of the Rocky Mountains. The lake, a vibrant shade of blue, is surrounded by a lush expanse of green trees and grass, creating a stark contrast with the clear blue sky above. The perspective of the image is from a high vantage point, offering a panoramic view of the lake and its surroundings. The majestic mountains in the background, blanketed in snow, add to the serene beauty of the scene. The image is a testament to the natural splendor of Emerald Lake, a popular destination for hikers and nature enthusiasts alike.

LLaVA-13b: The image captures the breathtaking view of a serene mountain lake, nestled amidst a lush forest. The lake, a deep shade of blue, is the centerpiece of the image, its calm waters reflecting the surrounding greenery. The forest, a mix of coniferous and deciduous trees, blankets the landscape, their verdant hues contrasting beautifully with the blue of the lake.A small stream, its waters a light blue, meanders through the foreground, adding a dynamic element to the otherwise tranquil scene. The stream's path is marked by rocks and boulders, hinting at the ruggedness of the terrain.In the distance, majestic mountains rise against the horizon, their peaks dusted with snow. The mountains provide a stunning backdrop to the lake and forest, their snow-capped peaks adding a touch of grandeur to the scene.Above it all, the sky is a clear blue, dotted with fluffy white clouds. The sky's vastness and the clouds' softness add depth to the image, creating a sense of openness and freedom.Overall, this image is a testament to the beauty of nature, capturing a moment of peace and tranquility in a stunning mountainous landscape.

Figure 4: Example Religion Image and LMM Captions

Unum: The image features a serene sunset over a body of water. The sky is painted in hues of orange and pink, with clouds scattered across it. The sun is setting, casting a warm glow over the scene. The text ""One Day My Life Is Over"" is prominently displayed in white text, with the phrase ""One Day My Life Is Over"" written in smaller, white letters. The image is taken from a high angle, giving a panoramic view of the sunset and the water. The overall composition of the image suggests a contemplative or spiritual theme.

LLaVA-7b: The image is a digital graphic with a landscape photograph as its background. The photograph depicts a sunset or sunrise with a gradient of colors ranging from warm oranges and yellows near the horizon to cooler blues and purples above. The sky is filled with clouds that are illuminated by the sun, creating a dramatic effect. Below the sky, there is a body of water reflecting the colors of the sky.Overlaid on the image is text in a casual, handwritten font. The text reads, Öne day I will wipe away every tear from your eyes. Revelation 21:3-4.T̈he text is in white, which stands out against the darker background of the image. The overall style of the image suggests it may be intended for inspirational or religious purposes, as indicated by the reference to the biblical passage from Revelation.

LLaVA-13b: The image is a photograph with a serene landscape and overlaid text. The landscape features a body of water, possibly a lake or sea, with a horizon where the sky meets the water. The sky is filled with clouds that are illuminated by the sun, creating a warm glow. The sun is partially obscured by the horizon, casting a golden light on the water's surface.Overlaid on the image is a quote in white text that reads, Öne day I will wipe away every tear from your eyes.T̈his is a reference to the Bible, specifically Revelation 21:3-4. The text is in a casual, handwritten font style, which contrasts with the natural and tranquil background. The overall effect is one of hope and comfort, suggesting a promise of a future free from sorrow. The image appears to be designed to evoke emotions of peace and reassurance.

3 Results

Two approaches to zero-shot learning are evaluated. First, the LMMs are asked to provide binary labels in response to a series of prompts, the specifics of which vary depending on the dataset. Second, the LMMs are asked to generate captions (descriptions) of each image and natural language processing is applied to those captions to classify the image. This caption based detection is explored for political and non-political content.

Binary Labels

For the 26,142 protest images, each model's caption response is converted to a binary measure as described in the previous section. Each image also already contains a binary label based on the supervised convolutional neural network trained for (Steinert-Threlkeld, Chan and Joo, 2022). These labels are one of {chidren, fire, flag, large group, small group, night, photo, police, protest, protester violence, shouting, sign, state violence, any violence}. Since those labels are based on human-annotated training data, they are taken as ground truth data against which to compare the LLM predictions.

Table 2 shows the confusion matrices for the children (left column) and shouting (right column) labels. Each model provides too many false positives and negatives. Unum consistently provides the most false positives but the fewest false negatives whereas the LLaVA models have an abundance of false negatives.

To explore these patterns for all labels, Figure 5 visualizes the F1, precision, and recall are then calculated for each label. Several interesting differences emerge. The most surprising is that Unum, the smallest model, consistently has larger F1 scores than the larger models. Only for detecting children and fire do the LLaVA models outperfrom Unum, but they do so with an extreme over eagerness: their recall is high but their precision low. In contrast, Unum has much better precision in those categories. The F1 scores for violence are similar for all models, but again Unum's precision is much better. The LLaVA models also perform poorly when detecting more specific measures of violence such as protester violence, state

Table 2: Confusion Matrices for Children and Shouting

Children

Unum

		Predicted	
		0	1
Ground Truth	0	15,621	10,046
	1	273	202

Shouting

Unum

		Predicted	
		0	1
Ground Truth	0	11,018	14,641
	1	204	279

LLaVA-7b

		Predicted	
		0	1
Ground Truth	0	20,033	5,634
	1	410	65

LLaVA-7b

		Predicted	
		0	1
Ground Truth	0	23,382	2,277
	1	330	153

LLaVA-13b

		Predicted	
		0	1
Ground Truth	0	20,562	5,105
	1	401	74

LLaVA-13b

		Predicted	
		0	1
Ground Truth	0	23,695	1,964
	1	361	122

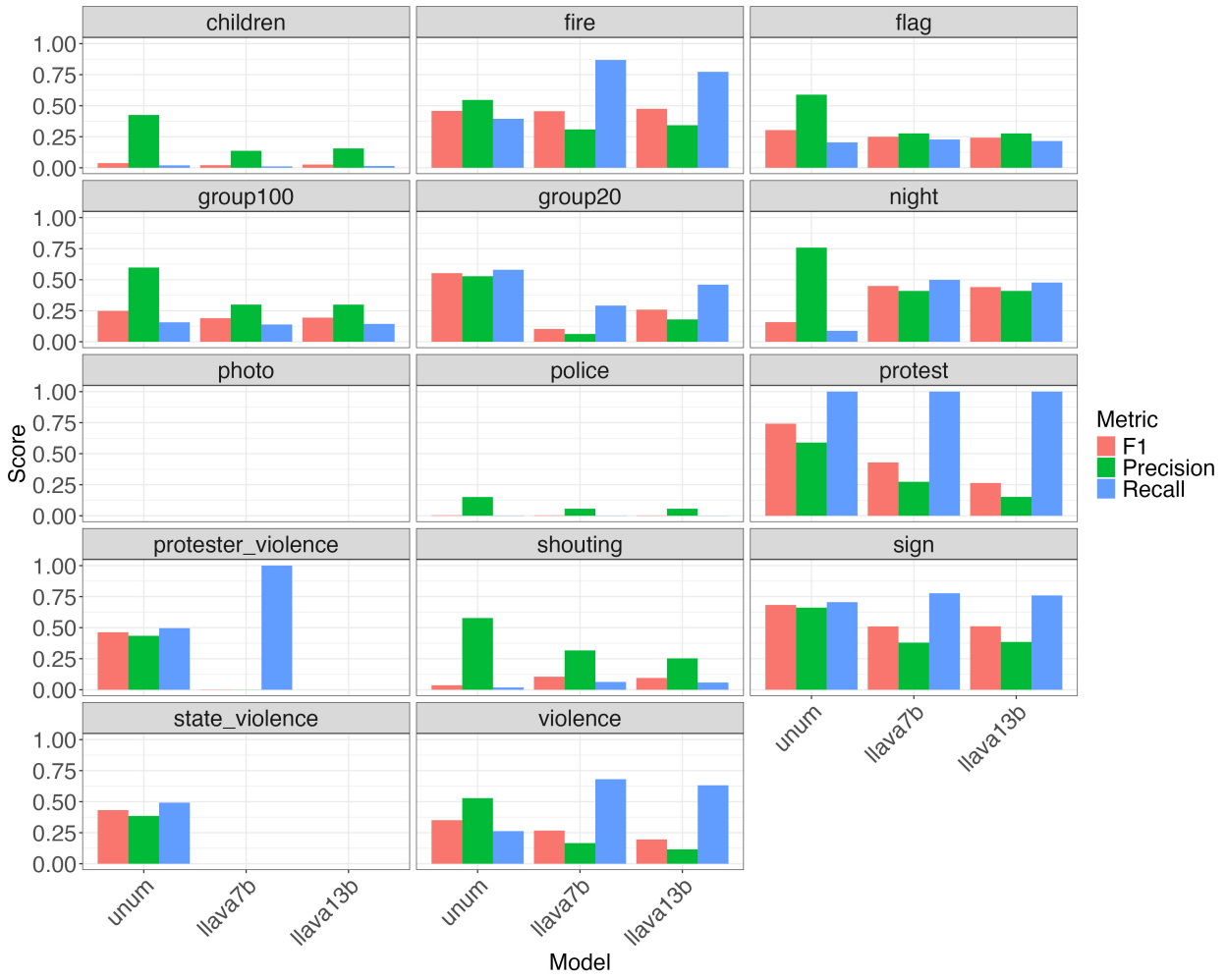
violence, and police.

Overall, the large multimodal models are impressive for their ability to classify unseen images, but they generate too many false negatives and positives to be used without fine-tuning.

For binary labels for the 1,000,000 images, 20 prompts were provided to each LMMs, one prompt per CAMEO event type.[3] Since the protest image dataset only contains protest events by construction, this analysis is performed only on the 1,000,000 random images. For all event types, the models overwhelmingly answer "yes", generating event proportions that are clearly not realistic. Figure 6 shows the results for a subset of event types for the Unum model. According to the model, 99.96% of images demonstrate aid provision, about 36.44% contain a protest, and 33.82% document unconventional massm violence. Such prevalence is obviously incorrect: LMMs cannot replicate the CAMEO coding scheme, and very liekly

[3]The event types are "make public statement", "appeal", "express intent to cooperate", "consult", "engage in diplomatic cooperation", "engage in material cooperation", "provide aid", "yield", "investigate", "demand", "disapprove", "reject", "threaten", "protest", "exhibit military posture", "reduce relations", "coerce", "assault", "fight", and "engage in unconventional mass violence".

Figure 5: Comparing Multimodal LLM Fit by Label, Protest Images



any other scheme, using zero-shot learning.

The same analysis was performed for LLaVA-7b and LLaVA-13b. Because of the difficulty with prompt engineering described in Section 2, fewer than 1% of images are determined to be of each event type. Those bars are therefore not plotted because they would barely appear.

Caption Classification

The next set of analysis looks at the possibility of using LMMs for analyzing the content of an image based on the caption a LMM generates to describe it. This approach is likely to produce more accurate estimates than the binary label prompts because the LMM is not

forced to assign a label to an image. In addition, the use of captions does not prejudge the labels that will be generated. When the content of an image is political, such as of a rally or protest, the LMM becomes an event detector. This possibility is especially powerful because existing event detection datasets are created from complex data processing pipelines and bespoke classifiers. If an off-the-shelf LMM can detect political events without customization, then many more event datasets could be created.

Two sets of labels are estimated from the captions. The first is the 14 from (Steinert-Threlkeld, Chan and Joo, 2022). The second is ten labels generated for this project designed to encompass a wide range of event types. These labels are protest, political rallies, sports, nature scenes, religious scenes, music, pornography, video games, and pop culture. For each set of labels, bags of words are applied to the captions. An image is assigned a label if any word matches, and topics are not exclusive.

Figure 7 shows the result for the 14 protest labels estimated from the captions for the 1,000,000 random images (top) and (Steinert-Threlkeld, Chan and Joo, 2022) images (bottom). In the protest image dataset, each LMM identifies protest as the most common content, excluding LLaVa-7b's anomalous identification of children. Though protest is the most common label, no model identifies more than 30% of the images as about protest, though the percentage should be 82.5 (the precision of the classifier from (Steinert-Threlkeld, Chan and Joo, 2022) used to identify protest images) or 100 (every image is labeled by (Steinert-Threlkeld, Chan and Joo, 2022) as being of protest). Not every caption contains specific words such as "protest" or "demonstration", so the underperformance is primarily due to bag of words model and not the captions themselves. Flags, a common feature of protests, are frequently recognized as is fire. unum consistently recognizes a higher percentage of events than LLaVa-13b, which in turn recognizes more than LLaVa-7b. For the 1,000,000 random images, sign is the most frequent category, then children, night, flag, and fire. For all labels, the models record similar percentages.

Figure 8 shows the results for the ten labels that include non-protest content. The results

for the 1,000,000 random images are shown in the top figure, the (Steinert-Threlkeld, Chan and Joo, 2022) in the bottom. For the protest images, a slightly higher percentage are identified as containing protest, and almost as many as containing a political rally. As with the 14 protest specific labels, the rank order of frequency is unum, LLaVa-13b, and LLaVa-7b. For the 1,000,000 images, the LLaVA models generate very similar label distributions while unum records fewer in all categories. Since unum is a smaller model than the LLaVa ones, it generates shorter captions: this behavior is expected.

Figure 6: Unum: Over 300,000 Instances of Protest and Unconventional Mass Violence

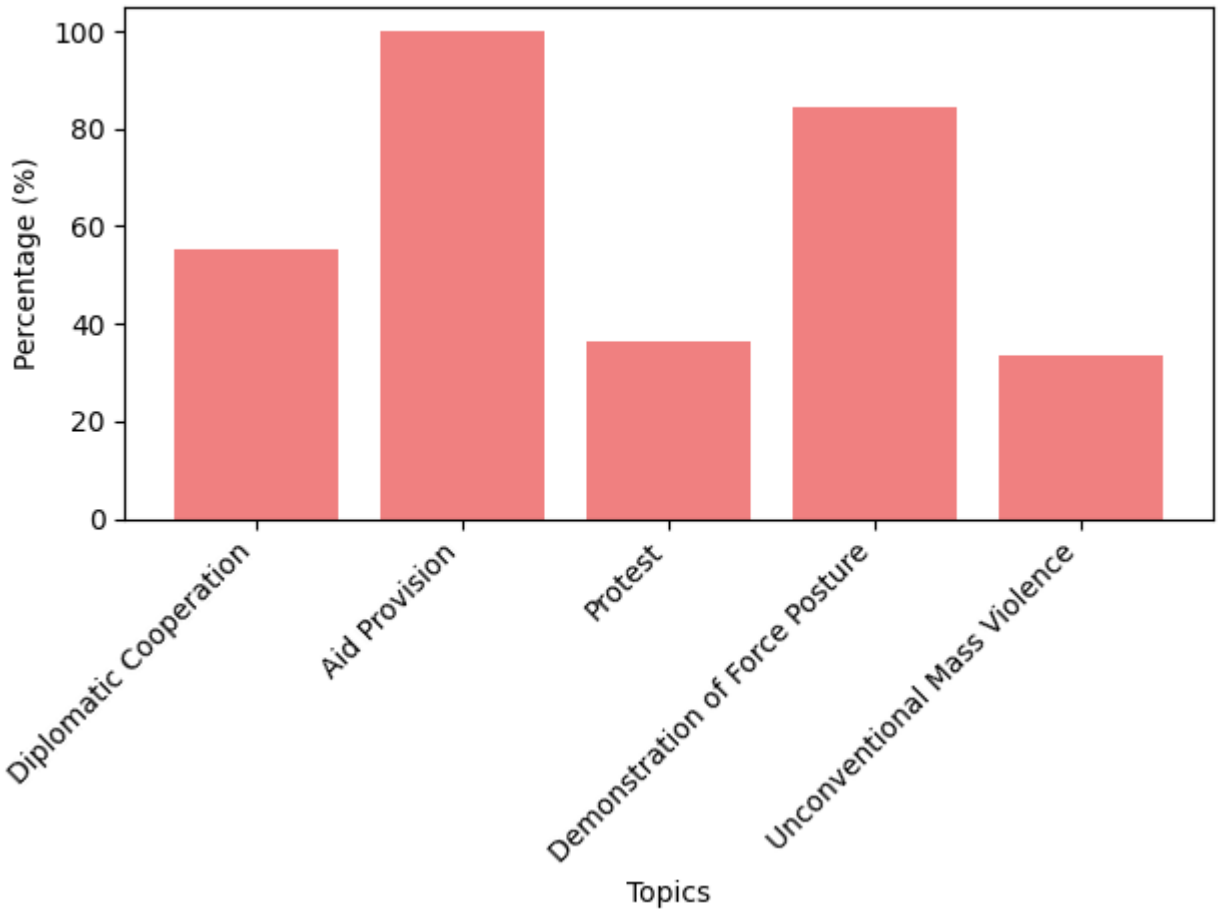


Figure 7: Protest Caption Topic Prevalence Across LMMs, Random Sample and Protest Images

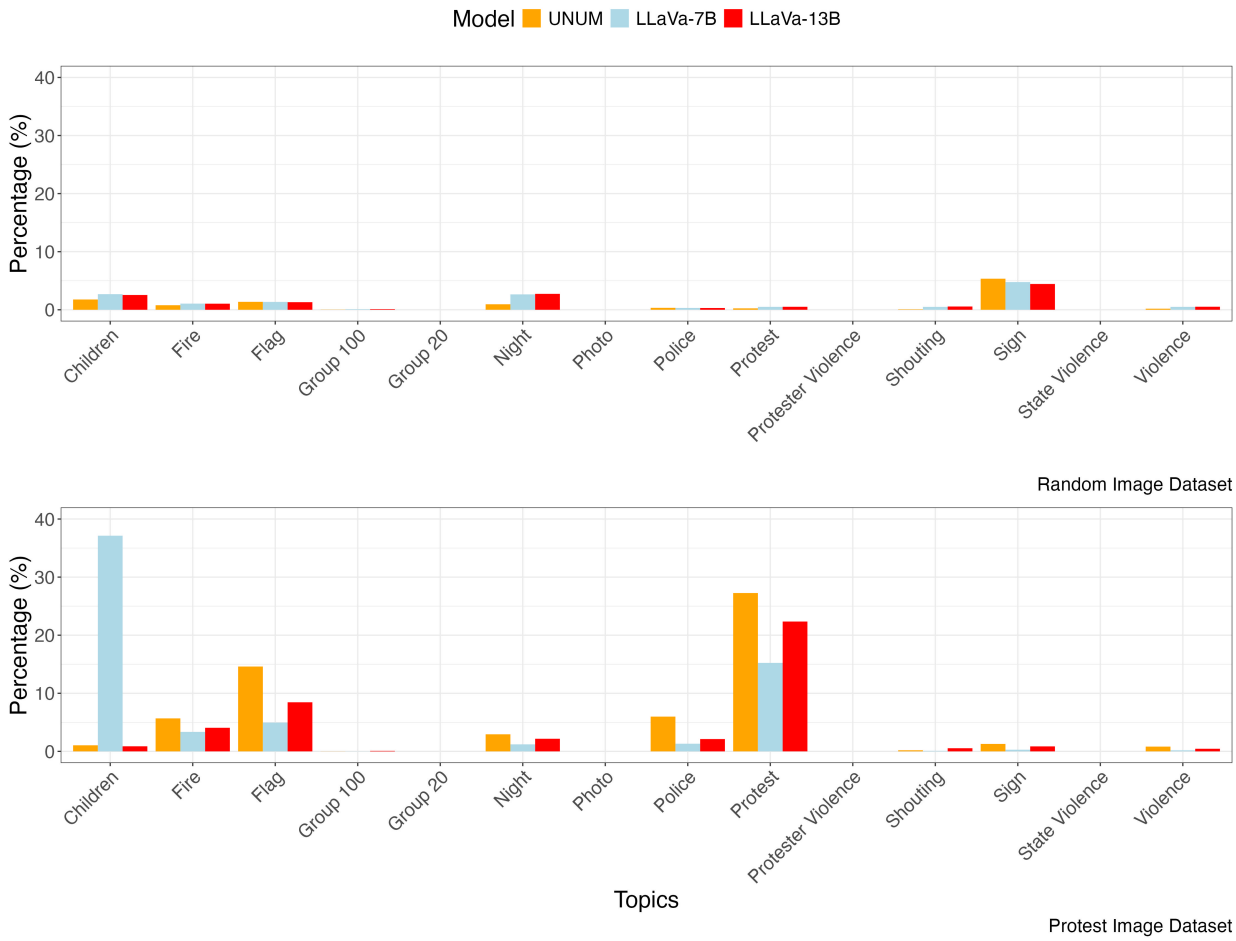
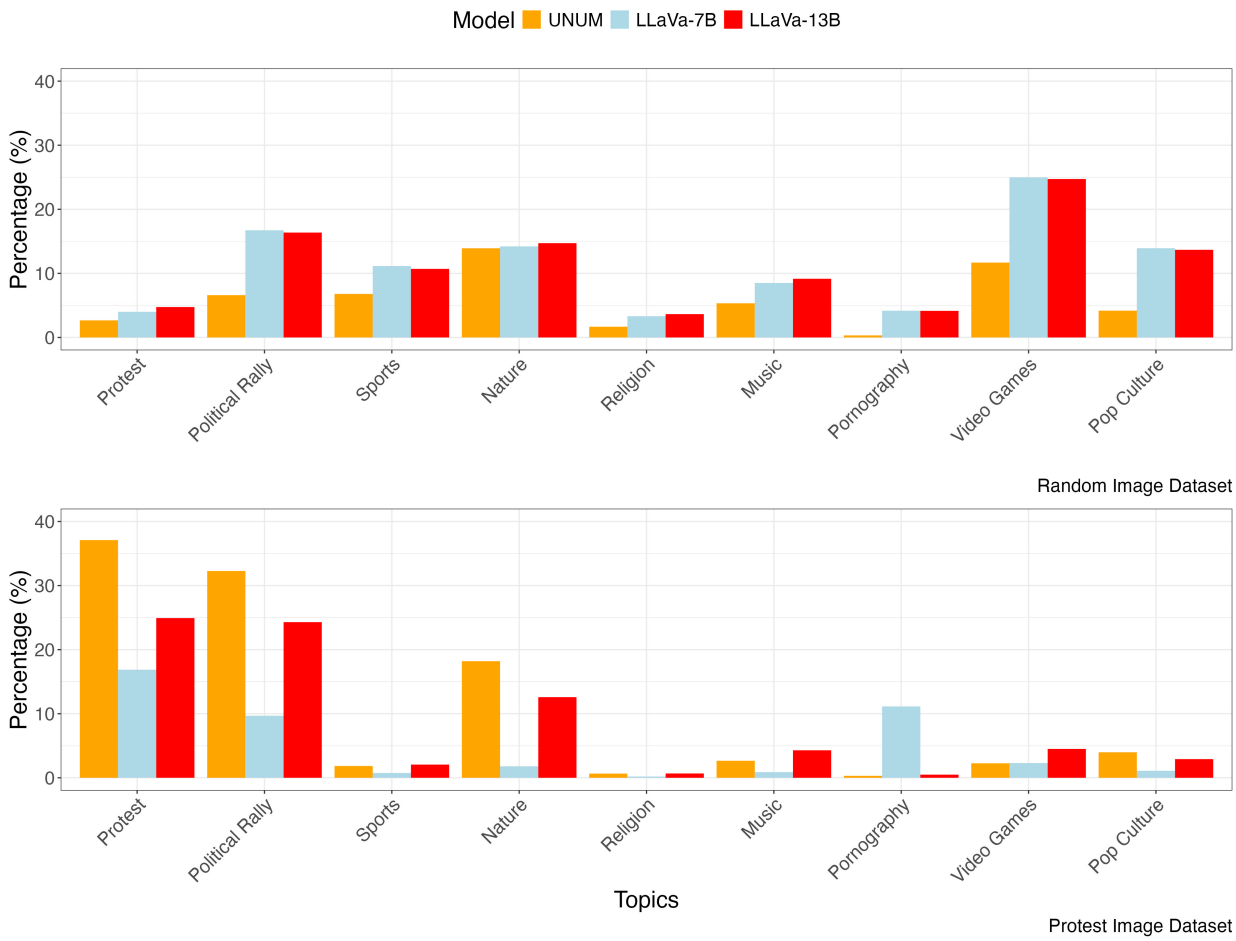


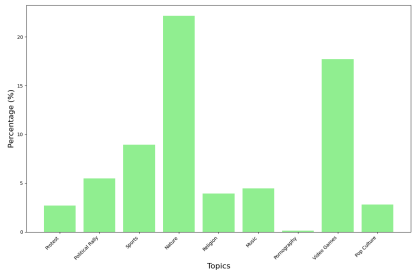
Figure 8: General Topic Prevalence Across LMMs, Random Sample and Protest Images



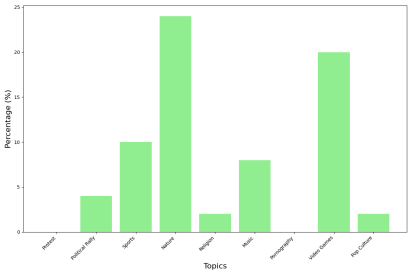
Spatial Heterogeneity

Because the images are from tweets with location information, topics can be mapped to countries to see how topic prevalence varies across countries. The top row of Figure 9 shows the topic prevalence for Japan, Puerto Rico, and the United States, respectively. Twitter assigns codes to 230 unique countries and territories, so the bottom row shows violin plots by topic and LMM. The distribution of topics varies across countries, though several consistencies emerge. (The outliers are likely from very small territories.) Unum identifies the fewest number of images per event type while the LLaVA models generate similar results. Despite these differences, the relative frequencies of topics are similar across the models. The distribution across topics also suggests what quickly becomes obvious to anyone studying social media data: political topics are not common, people prefer talking about almost anything else.

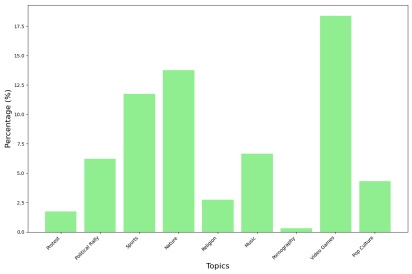
Figure 9: Prevalence of Topics by Country and LMMs



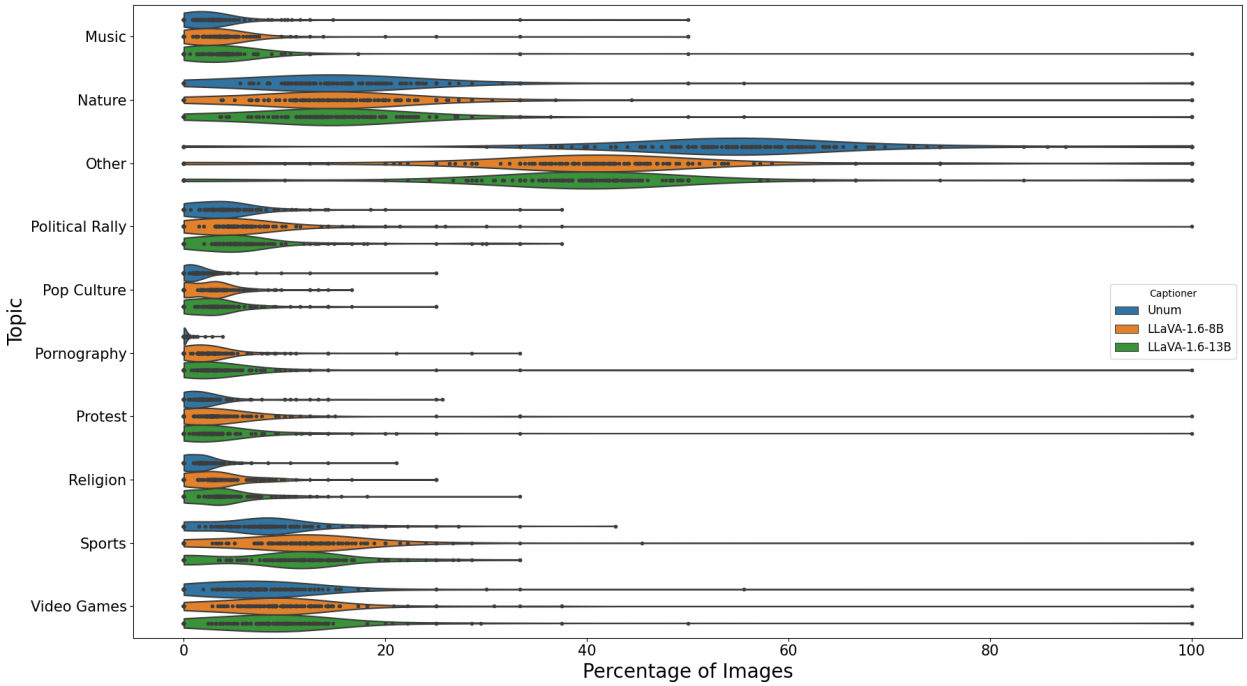
(a) Japan



(b) Puerto Rico



(c) United States



(d) Prevalence of Topics Across Countries by Model

4 Conclusion

This paper demonstrates the capabilities and limitations of large multimodal models (LLMs) in image analysis. The models examined, Unum, LLaVA-7b, and LLaVA-13b, reveal that as model size increases, the quality of image captioning improves, yet the performance in accurately identifying positive instances does not. Unum, the smallest model, while less comprehensive in caption detail, manages to maintain a competitive edge in identifying true positives compared to its larger counterparts, LLaVA-7b and LLaVA-13b. These larger models are better at identifying true negatives. Additionally, the results highlight the challenges of zero-shot image classification, which shows inconsistent performance across various scenarios. In contrast, zero-shot classification via the analysis of generated captions from these models displays more promise, suggesting a more accurate approach for image analysis.

Future work can build on these results in several ways. First, any prompt expecting a "yes" or "no" response should explicitly state that. Second, prompt engineering can lead to efficiency gains. For example, the CAMEO multiclass task can be reduced from 20 prompts to 1 where the model is forced to choose one label. Third, zero-shot learning needs to be validated with human annotators, especially for the labels not from the protest image dataset. Fourth, more robust natural language models to classify captions should be developed to improve the recognition of events from more than only keywords.

Large multimodal models are like new research assistants: even the lower performing ones can describe an image well, but they are all too eager to say yes when asked a question, generating an unacceptable number of false positives and negatives. Future research needs to study how to make them like excellent research assistants. Exploring ensemble coding techniques could be beneficial, potentially combining models that excel at identifying the presence of specific features with those better at confirming their absence. Moreover, it is crucial to explore the trade-offs between the operational speed of smaller models and their reduced caption accuracy. While larger models yield more detailed and accurate descriptions, their computational demands and slower processing speeds may not always be justifiable. In-

vestigating these trade-offs will be essential for practical applications, especially in scenarios requiring rapid image processing.

References

Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman and Karen Simonyan. 2022. "Flamingo: a Visual Language Model for Few-Shot Learning.". arXiv:2204.14198 [cs].

Anil, Rohan, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar

Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov and Yonghui Wu. 2023. "PaLM 2 Technical Report.". arXiv:2305.10403 [cs].

Anthropic. 2023. "The Claude 3 Model Family: Opus, Sonnet, Haiku.".
URL: *https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claud*

Bai, Jinze, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou and Tianhang Zhu. 2023. "Qwen Technical Report." *arXiv preprint arXiv:2309.16609* .

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis* 32(4):401–416.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei. 2020. "Language Models are Few-Shot Learners.". arXiv:2005.14165 [cs].

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.". arXiv:1810.04805 [cs].

Douglass, Rex W., Thomas Leo Scherer, J. Andrés Gannon and Erik Gartzke. 2024. "ICBeLLM: High Quality International Events Data with Open Source Large Language Models on Consumer Hardware.". arXiv:2401.10558 [stat].

Du, Xinya and Claire Cardie. 2021. "Event Extraction by Answering (Almost) Natural Questions.". arXiv:2004.13625 [cs].

Gilardi, Fabrizio, Meysam Alizadeh and Maël Kubli. 2023. "ChatGPT outperforms crowd workers for text-annotation tasks." *Proceedings of the National Academy of Sciences* 120(30):e2305016120.

Halterman, Andrew, Philip A. Schrodt, Andreas Beger, Benjamin E. Bagozzi and Grace I. Scarborough. 2023. "Creating Custom Event Data Without Dictionaries: A Bag-of-Tricks.". arXiv:2304.01331 [cs].

Ilharco, Gabriel, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi and Ludwig Schmidt. 2021. "OpenCLIP.". If you use this software, please cite it as below.
URL: *https://doi.org/10.5281/zenodo.5143773*

Li, Chunyuan, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang and Jianfeng Gao. 2023. "Multimodal Foundation Models: From Specialists to General-Purpose Assistants.". arXiv:2309.10020 [cs].

Liu, Haotian, Chunyuan Li, Qingyang Wu and Yong Jae Lee. 2024. "Visual instruction tuning." *Advances in neural information processing systems* 36.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach.". arXiv:1907.11692 [cs].

Lyu, Hanjia, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei and Jiebo Luo. 2023. "GPT-4V(ision) as A Social Media Analysis Engine.". arXiv:2311.07547 [cs].

Schrodt, Philip A, Omür Yilmaz, Deborah J Gerner and Dennis Hermreck. 2008. The CAMEO (conflict and mediation event observations) actor coding framework. In *2008 Annual Meeting of the International Studies Association*.

Steinert-Threlkeld, Zachary, Alexander Chan and Jungseock Joo. 2022. "How State and Protester Violence Affect Protest Dynamics." *The Journal of Politics* 84(2):798–813.

Wang, Yu. 2023. "Topic Classification for Political Texts with Pretrained Language Models." *Political Analysis* 31(4):662–668.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le and Denny Zhou. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35:24824–24837.

Wu, Wenhao, Huanjin Yao, Mengxi Zhang, Yuxin Song, Wanli Ouyang and Jingdong Wang. 2024. "GPT4Vis: What Can GPT-4 Do for Zero-shot Visual Recognition?". arXiv:2311.15732 [cs].