Large Language Models Reveal Information Operation Goals, Tactics, and Narrative Frames

Keith Burghardt[1] **Kai Chen**[2] **Kristina Lerman** [1,2]

[1] USC Information Sciences Institute
[2] University of Southern California
{keithab,kchen,lerman}@isi.edu

Abstract

Adversarial information operations can destabilize societies by undermining fair elections, manipulating public opinions on policies, and promoting scams. Despite their widespread occurrence and potential impacts, our understanding of influence campaigns is limited by manual analysis of messages and subjective interpretation of their observable behavior. In this paper, we explore whether these limitations can be mitigated with large language models (LLMs), using GPT-3.5 as a case-study for coordinated campaign annotation. We first use GPT-3.5 to scrutinize 126 identified information operations spanning over a decade. We utilize a number of metrics to quantify the close (if imperfect) agreement between LLM and ground truth descriptions. We next extract coordinated campaigns from two large multilingual datasets from X (formerly Twitter) that respectively discuss the 2022 French election and 2023 Balikaran Philippine-U.S. military exercise in 2023. For each coordinated campaign, we use GPT-3.5 to analyze posts related to a specific concern and extract goals, tactics, and narrative frames, both before and after critical events (such as the date of an election). While the GPT-3.5 sometimes disagrees with subjective interpretation, its ability to summarize and interpret demonstrates LLMs' potential to extract higher-order indicators from text to provide a more complete picture of the information campaigns compared to previous methods.

Introduction

Information operations (info-ops) often utilize social media platforms to manipulate users at scale (Bradshaw and Howard 2019; Badawy, Ferrara, and Lerman 2018; Kim 2018; Burghardt et al. 2023). Efforts include, but are not limited to: manipulating authentic users on each side of a divisive issue (Ratkiewicz et al. 2021; Kim 2018; Stella, Ferrara, and Domenico 2018), reducing trust in democracies (Badawy et al. 2019), and promoting hate speech (Hickey et al. 2023), misinformation (Vosoughi, Roy, and Aral 2018), and financial scams (Pacheco et al. 2021).

Info-ops utilize a range of techniques to accomplish their goals, often though the coordination of several accounts to promote a particular message (Burghardt et al. 2023; Pacheco et al. 2021), while simultaneously appearing to act organically, as a human user, so as to build trust and evade detection by social media platforms (Ferrara 2017; Sayyadiharikandeh et al. 2020; Paper 2022). Despite substantial progress in detecting coordinated campaigns, whether via conventional methods (Luceri et al. 2023; Burghardt et al. 2023; Pacheco et al. 2021) or even LLMs (Luceri, Boniardi, and Ferrara 2023), there has so far been little work in utilizing AI to analyze campaigns at scale (Burghardt et al. 2023). Instead, analysis of campaigns has relied on manual review and annotation (Martin and Shapiro 2019; Carley 2020), which reduces our ability to detect campaigns at scale and understand their goals and tactics.

In this paper, we explore the utility of LLMs at annotating campaign goals, tactics, and narrative frames, using GPT-3.5 as a case study. We utilize an annotated dataset of campaigns (Martin and Shapiro 2019) as a benchmark to test the accuracy of zero-shot LLMs and find their outputs are substantially better than baselines, although they often make mistakes. We explore the utility of LLMs in previously understudied campaigns using two multi-lingual X datasets from the 2022 French election and 2023 Balikatan Philippine-U.S. military exercise in 2023, respectively. These data contain several coordinated campaigns and discuss a range of concerns, including the Russia-Ukraine war, domestic politics, and the economy. We first extract previously unknown coordinated campaigns using proxies of coordination (Burghardt et al. 2023; Luceri et al. 2023), a recent but well-studied method that is both interpretable and achieves reasonable precision and recall across a range of ground truth datasets (Luceri et al. 2023). This proxy links two accounts if they create posts with long strings of hashtags in the exact same order, thus generating a network of coordinated accounts where nodes are accounts and links are account pairs that are coordinated. We can then define campaigns as connected clusters of coordination. Within each coordinated campaign, we extract the concerns discussed in each post using a novel instruction-tuned LLama-2 model that achieves reasonable precision and recall. After filtering campaign posts by their respective concerns, we then apply GPT-3.5 to extract the validated features of prior campaigns (Martin and Shapiro 2019), such as the political goals, and countries responsible for the campaign. We further extract tactics via the BEND framework (Carley 2020), and frame narratives (Chong and Druckman 2007), which is a political science theory used to understand how politicians frame top-

ics to influence the populace. The LLM annotation demonstrates that framing offers a new understanding of information campaigns as well. To summarize, our contributions are:

- We develop a novel method to validate LLM info-ops capabilities based on a large dataset of information operations (Martin and Shapiro 2019).
- We create a new concern detection model and apply it to two datasets in order extract issues pushed by info-ops.
- We utilize the validated LLM on dynamic and concern-tailored info-ops in order to evaluate their answers to various well-grounded questions about goals, tactics, and narrative frames.

Overall, we demonstrate LLMs offer rich high-level indicators than previously possible (Burghardt et al. 2023), even if the tool is imperfect. While we do not expect LLMs to ever replace humans in analyzing information campaigns, we believe they can substantially speed up our understanding of each campaign. Our code, including prompts and outputs, are shown in the following repository: https://github.com/KeithBurghardt/LLM_Coordination.

Related Work

Information Operations Online manipulation (Tucker et al. 2018) is both widespread and covers a range of goals such as politics (e.g., the Brexit vote (Howard and Kollanyi 2016) and presidential elections (Burghardt et al. 2023; Ferrara 2017; Bessi and Ferrara 2016; Badawy, Ferrara, and Lerman 2018; Badawy et al. 2019; Kim 2018)), psychological warfare (Paper 2022), promoting hate (Hickey et al. 2023), scams (Pacheco et al. 2021), and pushing COVID-19-related messages (Graham et al. 2020; Piña-García and Espinoza 2022). We cannot easily assess the impact of these information operations (Bail et al. 2020), but their widespread use suggests they are useful for manipulation.

Notably, however, these analyses have been manual, either through exploring case studies (as above), or through a systematic categorization that have only been applied to a limited set of campaigns. For example, Carley developed a BEND framework (Carley 2020) to categorize the tactics, such as Excitement of a topic or Distortion of a narrative, these categories have not been applied at scale. Similarly, 126 campaigns have been systematically studied (Martin and Shapiro 2019) with their own set of categories, such as political goals, but it is unclear how well these methods extend to other campaigns. Finally, a completely different field, political science, has developed a parallel set of methods to understand narrative frames used to change the opinions of the populace, known as Framing Theory (Chong and Druckman 2007), which aim to categorize the way campaigns frame problems, causes, and remedies, as well as how they are summarized via slogans. This theory has been used to understand social media campaigns (Hon 2016; Shahid et al. 2020), but has been under-applied to understand information operations.

Detecting Information Operations There have been a range of tools to analyze information operations, starting with bot-detection tools often on X (formerly Twitter) (Ferrara 2017; Stella, Ferrara, and Domenico 2018), and moving onto coordinated activity detection, namely whether accounts perform in concert to push a message (Starbird 2019; Burghardt et al. 2023). These tools have been applied across multiple platforms (e.g., Facebook (Giglietto et al. 2020b,a) or YouTube (Kirdemir, Adeliyi, and Agarwal 2022), as well as Twitter (Sharma et al. 2021; Nizzoli et al. 2021; Weber and Falzon 2021; Mazza, Cola, and Tesconi 2022; Cinelli et al. 2022; Burghardt et al. 2023; Luceri, Boniardi, and Ferrara 2023; Luceri et al. 2023)). While a range of methods to detect account coordination are used (Sharma et al. 2021; Weber and Falzon 2021; Schliebs et al. 2021; Pacheco et al. 2021; Schliebs et al. 2021; Kirdemir, Adeliyi, and Agarwal 2022; Giglietto et al. 2020a; Burghardt et al. 2023; Pacheco et al. 2021), we stick to an interpretable and well-respected metric, hashtag co-occurence (Burghardt et al. 2023; Luceri et al. 2023), which has been independently verified as a useful feature to detect information operations (Luceri et al. 2023). To find coordinated campaigns these data are converted into a network where accounts are nodes and links connect coordinated pairs (Pacheco et al. 2021; Burghardt et al. 2023). Coordinated campaigns are then connected components within this network.

AI Analysis of Information Operations While LLMs have proven remarkably useful across a range of fields (Katz et al. 2024; de Winter 2023) (including analysis of social media (Luceri, Boniardi, and Ferrara 2023; Liyanage, Gokani, and Mago 2023), this field is still quite nascent, with papers just beginning to understand how LLMs could, e.g., detect coordination (Luceri, Boniardi, and Ferrara 2023). While some previous work has used conventional methods to analyze stories within coordinated accounts (Ehrett et al. 2021), stances (Chen et al. 2021), or socio-linguistic indicators (Burghardt et al. 2023), there is little work on LLMs to analyze information operations.

Concern detection. Concerns, also known as wedge issues (Van de Wardt, De Vries, and Hobolt 2014; Heinkelmann-Wild et al. 2020), is crucial for understanding social media influence campaigns (Martínez 2023). Traditionally concern detection relies on open-world classification methods (Shu, Xu, and Liu 2018; Esmaeilpour et al. 2022; Bai et al. 2023), but these methods are limited by data size; moreover concern detection in social media is complicated and nuanced. Burghardt et al. (2023) proposed a method that extracts concern keywords from Wikipedia for training models such as BERT (Devlin et al. 2019). This method is effective but depends on the Wikipedia pages and the quality of keyword extraction.

Recent advances in LLMs have offered greater promise. For example, Chen et al. (2024a) utilize LLMs to enhance the extraction of keywords from Wikipedia pages. However, the high costs of powerful closed-source LLMs like GPT-4 (Achiam et al. 2023) force a turn towards knowledge distillation. This approach transfers capabilities from larger to smaller models efficiently (Gudibande et al. 2023; Chen et al. 2024b), an inspiration for our cost-effective concern detection framework, which does not utilize com-

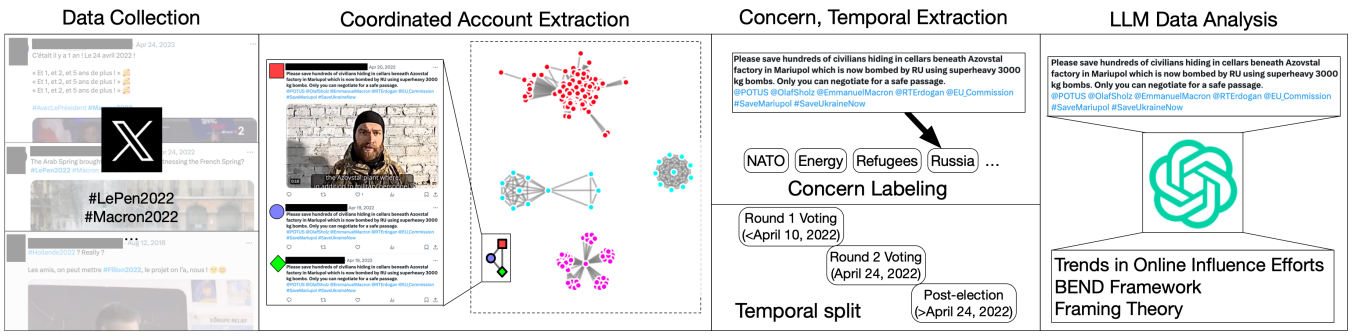

Figure 1: We develop an annotation technique as follows. First, we extract data from X matching a set of keywords related to the 2022 French election or the 2023 Balikatan U.S.-Philippines military exercise. We then collect coordinated campaigns from these data based on a well-accepted hashtag co-occurrence metric (Burghardt et al. 2023; Luceri et al. 2023). We then label concerns and break up posts to particularly important date ranges for each campaign (such as just before an election). Alternatively, when testing the campaigns extracted in (Martin and Shapiro 2019), we use LLMs to create posts from these campaigns (we lack concern or dynamic information from the concern data). Finally, we utilize GPT to extract features based on previous framework (Martin and Shapiro 2019), the BEND framework (Carley 2020), and Framing theory (Chong and Druckman 2007). We evaluate each output, to address the potential for halucinations, to understand the utility of the LLMs as methods to extract higher-order features of coordinated campaigns.

putationally costly LLMs and potentially expensive APIs. Finally, we utilize instruction tuning, which can enhance the performance of a tuned model, particularly in zero-shot tasks (Mishra et al. 2021; Sanh et al. 2021; Wei et al. 2021).

Methods

Figure 1 shows the overall diagram of the proposed system. The components of the system are described in detail below.

Data Collection

We use Twitter data pertaining to the discussions of the 2022 French election (Round 1 was April 10, 2022 and Round 2 occurred on April 24, 2022), and the "Balikatan" U.S.-Philippines military exercise from April 11 to April 28, 2023. All data have post ID and PII are removed prior to analysis to improve anonymity and collection methods were approved by the respective institutional review boards. Data were publicly collected in the U.S. and did not require consent.

The 2022 French election dataset contains 5.9M posts (4.3M reposts, 1.6M replies, 18K original posts; 678K authors total) spanning February 15 to the end of June 30, 2022. Data were collected by a third party via keywords related to the French election, such as "election" or "élection" and presidential candidate accounts and their parties (@EmmanuelMacron, @ZemmourEric, or @MLP_officiel). Within this dataset, 95% of posts were in French and the rest (5%) were in English based on X's language feature.

In contrast, the U.S.-Philippines dataset contains 4.7M posts (3.2M reposts, 852 replies, 605K original posts, 130K quote posts; 1.9M authors) spanning January 1, 2023 to June 28, 2023. Data were similarly collected from a third party via keywords. Within this dataset, 94% of posts were in English and the rest (6%) were in Tagolog.

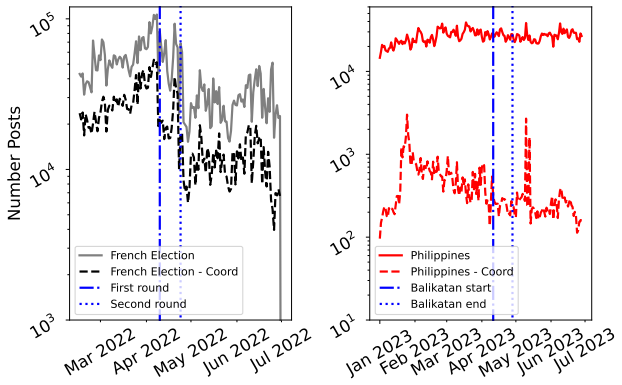

Figure 2: Number of posts over time from coordinated and non-coordinated accounts. (a) 2022 French election, with Round 1 and Round 2 elections labeled, and (b) 2023 Balikatan U.S.-Philippines military exercises that took place between April 11 and April 28, 2023.

The LLM is therefore tested multilingual datasets, broadly in either French or English. We see posts over time in Fig. 2, which shows the daily post count for coordinated and coordinated accounts (note the y-axis is log-scaled).

Coordinated Account Extraction

Inspired by (Burghardt et al. 2023; Luceri et al. 2023; Pacheco et al. 2021), we use the co-occurrence of hashtags in a given post as indicators of coordination, which has been well-validated in ground truth datasets (Luceri et al. 2023). To this end, we detect whether two accounts are coordinated if they each contain posts with at least three hashtags, all of them the same and in the same order. This simple heuristic captures the notion that individual text may vary but if the

text is semantically the same, usually the hashtags will be in the same order. This is independently validated on ground truth data and performs well with high precision and recall in many datasets (Luceri et al. 2023). This indicator of coordination creates networks whose accounts are linked if the indicator flags these accounts as coordinated. This in turn means we can define different coordination campaigns as separate connected components, which has been validated in previous work (Burghardt et al. 2023). For coordination detection, we used Intel(R) Xeon(R) CPUs in an internal cluster.

Concern Labeling

Detecting concerns is a crucial component of identifying influence campaigns. A *concern* refers to an issue characterized by multiple, competing, and often incompatible positions, commonly known as "wedge issues" (Van de Wardt, De Vries, and Hobolt 2014). For instance, crime or environmental issues could be significant concerns to the Philippines' population. These concerns are rooted in deeper moral or sacred values, which play a pivotal role in shaping social identity and beliefs. By exploiting these concerns, adversarial actors can effectively widen social divisions or manipulate elections, leveraging the profound impact of moral and sacred values on societal dynamics (Ayala 2010).

Concern detection in social media is a complex and challenging task. Social media discussions touch on a broad array of concerns, many of which are dynamic and emerge from recent events. This variability requires methods that have a strong general ability to respond to new issues. Social media datasets are also massive and often lack sufficient context to accurately interpret and categorize concerns. This lack of context hampers the ability of models to make informed classifications. The boundaries between different concerns are often blurred, making it challenging to clearly delineate and classify them. This ambiguity complicates the task of concern detection and underscores the challenges of concern detection in social media.

To address these challenges, we propose a novel framework that leverages LLMs to automate concern detection. The overview of the framework is shown in Figure 3, and it is composed of the following steps:

- **Sample classification.** Our first step is to annotate a sample of X posts. Given a set of concerns relevant to the domain identified by subject matter experts, we first create an *expert model* for concern detection. The categories were wide-ranging, such as Russia, the economy, domestic politics, international relations, and the US military. The concerns are explained in more detail in the Results section. While there are other potential categories that could be identified, these nonetheless act as a proof-of-concept of our methodology. We use GPT-4 (Achiam et al. 2023) as the expert or teacher model as it has demonstrated remarkable ability to classify text in a zero-shot setting but is expensive to employ for the classification of 1M+ post datasets. More specifically, we use GPT-4 to classify 20K posts for each dataset (40K posts total) and 3.6-10K posts within each dataset are annotated by multiple subject matter experts for validation (10K posts for the 2022 French Election dataset, and 3.6K posts for the Balikatan datast; see Fig. 4 for the frequency of each concern, where each posts may have multiple concerns). The prompt for concern classification using GPT-4 is shown in Appendix (we allow GPT-4 to annotate multiple concerns for each post). While we do not validate GPT-4 itself, we believe it's results are valid due to its ability to generate training data for an accurate concern prediction model, as described below.
- **Instruction tuning.** We use a distilled student model (Llama-2-7B (Touvron et al. 2023)) to annotate millions of posts, which is trained via instruction tuning, a commonly used method to fine-tune LLM parameters to a specific task (Zhang et al. 2023). The specific format of the instruction pairs used for Llama-2 is shown in the Appendix . For training, we use two NVIDIA A100 (80GB) GPUs for 3 epochs, with batch size 1, the gradient accumulation steps 16, the learning rate 2×10^{-5} (other details of the training can be seen in the link to our code).

We measure the model performance using precision, F1, and ROC-AUC.

LLM Information Campaign Annotation

We use three frameworks for LLM annotation with GPT-3.5 (using the version as of March 12-21, 2024). While other models (including GPT-4) might be more accurate, they are too expensive for the broad range of questions we ask and tweets we have to annotate (the GPT-3.5 annotations in this paper cost $50 and GPT-4 would, at present, swell that number 20-fold). There are also ways to further improve the prompts via chain-of-thought prompting or few-shot learning, but we use zero-shot prompting as a starting point.

We first annotated campaigns based on categories developed from (Martin and Shapiro 2019):

- Targeted country
- Attacking country
- Political goal (category and description)
- Information operation description

Because we have a set of 126 annotated campaigns (Martin and Shapiro 2019), we can directly compare LLM annotations to ground truth descriptions in order to validate our zero-shot prompts.

Next, we create prompts from the BEND framework (Carley 2020). Because "B" and "N" objectives (e.g., actions that decrease the importance of an opinion leader) represent actions that cannot be measured from posts alone, we focus instead on the "E" (positive) and "D" (negative) objectives:

- Engage: bring up related/relevant topics
- Explain: provide details on or elaborate the topic
- Excite: elicit positive emotions
- Enhance: encourage messages on the topic
- Dismiss: explain why topic is unimportant
- Distort: alter the main message of a topic
- Dismay: elicit negative emotions
- Distract: discuss a different, irrelevant topic

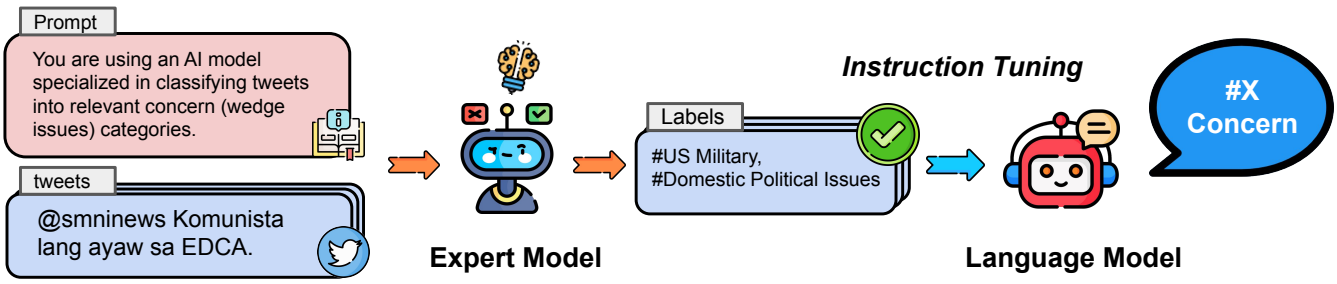

Figure 3: The framework of concern detection. Initially, we sample a small amount of data and label them using an expert model. Subsequently, we train a language model on this labeled dataset through instruction tuning.

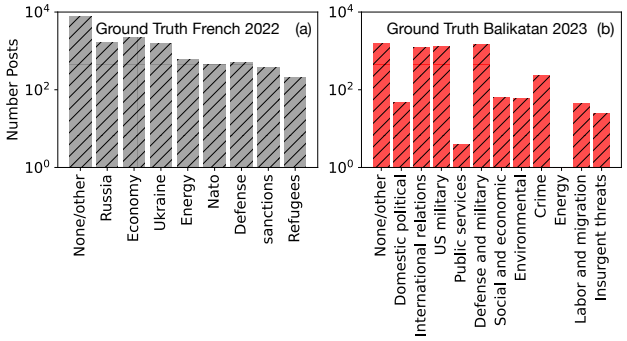

Figure 4: The frequency of concern posts in human annotated validation data. (a) 2022 French election and (b) 2023 Balikatan U.S.-Philippines military exercises.

Third, we borrow from Framing Theory (Chong and Druckman 2007; Entman 1993), a well-studied problem in political science, to understand how issues are framed to influence the public. While typically applied to news media or politicians, we show these features can help us understand information operations as well. To this end, we prompt the LLM to provide:

- The problem the information operation is describing
- The cause of the problem
- The remedy for this cause
- Metaphors used in this frame
- The catchphrases and slogans
- Frame examples for the problem, cause, and remedy (for qualitative validation)

In addition, we asked the LLM to provide:

- Cultural cues or in-group language
- Textual motifs (which was discarded due to irrelevant responses)
- Intentional rhetorical fallacies

For the last bullet point, we included a list of intentional rhetorical fallacies in the prompt (e.g., Ad Hominem, Ad Baculum, etc.) from the Internet Encyclopedia of Philosophy (https://iep.utm.edu/fallacy/). While we cannot verify the accuracy of the LLM for the latter two sets of tasks (as we do not have validation data) the qualitative evaluation offers an important exploration of LLM annotation quality to guide future understanding of coordinated accounts.

Evaluating LLM Coordination Annotation

LLMs often suffer from hallucinations (Yao et al. 2023), thus we want to evaluate how well LLMs can accurately explain the goals of information campaigns. To this end, we extract from prior work (Martin and Shapiro 2019) descriptions of 126 information operations. Our goal is to check if LLMs could create semantically equivalent statements about these operations across various categories.

We therefore fed into GPT-3.5 a prompt consisting of all information about the info-ops provided by the prior dataset (Martin and Shapiro 2019), namely the targeted country, the attacking country, the political goal category, the political goal, and the event description. We then asked that the LLM generate 10 posts from X. After creating these posts, we then create a prompt for the LLM to determine the targeted country, the attacking country, political goal (category and description) and a description of the event. We finally compare the output of this model against the ground truth description. Because we use the ground truth to generate the data (we lack the actual posts from the information campaigns described in (Martin and Shapiro 2019)), there is some data leakage, thus our results act as an upper bounds to zero-shot GPT-3.5 performance. That said, the results are likely ecologically valid as we expect LLMs are being increasingly used for information operations. Example posts are in the Appendix Table 3. While some of these X-like posts seem too on-the-nose (e.g. #DiscreditKurz), the posts also qualitatively resemble posts from real coordinated accounts (in part because LLMs may be used in many campaigns).

LLM outputs can often be evaluated via simple metrics, such as F1 or ROC-AUC, as seen in the concern detection LLM model above, among others (Chang et al. 2023). Other metrics might inclue BLEU or METEOR scores when evaluating LLMs' ability to translate text (Banerjee and Lavie 2005), but these score n-gram similarity rather than semantic similarity, which we are interested in. We therefore utilize NLI techniques (Chang et al. 2023) to determine if the LLM output and the ground truth description are equivalent. We use BART (Lewis et al. 2019) fine-tuned for natural language inference (https://huggingface.

co/facebook/bart-large-mnli), GPT-3.5, and GPT-4 to determine whether the GPT-3.5 LLM output and the ground truth text are semantically equivalent. BART is given a hypothesis (the ground truth campaign description) and determines the confidence that this falls out from the premise, the LLM output. GPT-3.5 and GPT-4 are each prompted to determine whether the ground truth description and LLM output are equivalent (see link to our code for prompr details). While each metric may have its flaws (Tao et al. 2023), the consistency of all metrics improves the robustness findings (the prompts are in our code listed in the Introduction).

Results

Evaluating Concern Labels

To validate the reliability and practicality of concern detection, we conduct two validation experiments focusing on posts collected for distinct geopolitical domains: the 2022 French elections and the Philippines. Both datasets are multilingual. In the context of the 2022 French elections, the analysis is structured around nine predefined concern categories: *Ukraine, Russia, NATO, refugees, defense, economy, economic sanctions, energy*, and a residual category designated as *none or other* to encapsulate unclassified concerns. The Philippines Balikatan data involves twelve concerns: *crime, defense and military, domestic political issues, international relations, labor and migration, public services, social and economic issues, US military, energy*, and a similarly defined *none or other* category. These categories were selected to cover a broad range of social and political concerns, allowing for an evaluation of the framework's capacity to navigate complex thematic content across disparate linguistic and cultural narratives.

Table 1 reports performance of the concern detection model on the French Elections dataset, where standard deviations reported are from bootstrapping the prediction and ground truth pairs 1000 times (due to the long training time of several hours, bootstrapping the training data and retraining the models was not feasible). The baselines were predicting all the same label (e.g., all posts are "economy", "none or other", etc. The model has relatively high average precision and F1 score, with especially strong performance for concerns like *Ukraine*, *Russia*, and *NATO*. This suggests the model is effectively identifying relevant concerns with few false positives. However, the *Refugees* and *Defense* categories exhibit weaker performance, indicating potential difficulties in capturing the nuance within these concerns due in part to their low frequency in posts. Table 2 reports performance of the model on the Balikatan dataset. As before, we use the same baseline and standard deviations reported are from bootstrapping the prediction and ground truth pairs 1000 times. The *Defence and military* and *Public services* concerns demonstrate lower performance in part due to the low frequency of these events. It is less challenging for the model to recognize concerns like *Crime* and *Labor and migration*. Energy has only 1 annotated post, so we do not validate it.

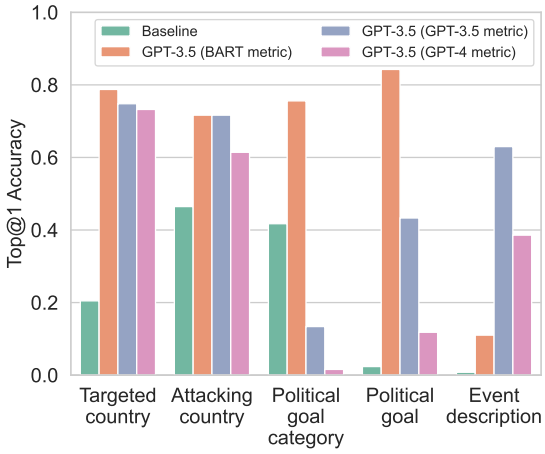

Figure 5: LLM metrics comparing GPT-3.5 annotations and ground truth from (Martin and Shapiro 2019). We use BART, GPT-3.5, and GPT-4 to evaluate whether GPT-3.5-based descriptions of posts agree with ground-truth data. We find across all metrics typically strong agreement with ground-truth, except for zero-shot political category.

LLM Evaluation

Our evaluation is shown in Fig. 5. We find that the accuracy (how many descriptions are equivalent) far exceed baselines, especially for event descriptions, political goals, and targeted countries. Therefore while the LLMs are far from perfect, they can reasonably accurately infer many subtle details of a campaign. The performance is poorest, however, for the political goal category. We decided for consistency we would not present the set of possible categories to the LLM, which means the performance is necessarily poor when the LLM does not know the categories a priori. In contrast, not providing the "country" categories in advance did not substantially harm the LLMs ability to detect the targeted and attacking country.

Discovering Coordinated Accounts

We use coordination indicators to identify clusters of coordinated accounts in our French elections and Balikatan datasets. We show the number of coordinated accounts in each cluster in Fig. 6. In the plots, we limit to clusters with at least 10 accounts each. In total, there were 29K coordinated accounts in the 2022 French election dataset, and 4.7K coordinated accounts in the 2023 Balikatan dataset. This corresponds to 2.7M posts in the 2022 French election (an astonishing 45.5% of all posts) and 79K in the Balikatan dataset (1.7% of all posts). While the former figure sounds surprising, they were written by merely 11.6% of all accounts (0.2% of accounts were coordinated in the Balikatan dataset). That said, 11.6% is still a somewhat high figure and may reflect the keywords and accounts captured in the data sample. Another peculiarity of the French election dataset is that there are just 4 clusters with more than 10 accounts. The largest cluster corresponds to the largest proportion of all ac-

Table 1: 2022 French Elections: Performance of Concern Labeling

Performance metric	Precision		F1		ROC-AUC	
	Model	Baseline	Model	Baseline	Model	Baseline
None or Other	0.828±0.005	**0.878±0.003**	**0.806±0.004**	0.782±0.004	**0.789±0.005**	0.5
Russia	**0.906±0.004**	0.289±0.006	**0.846±0.004**	0.169±0.004	**0.808±0.005**	0.5
Economy	**0.829±0.006**	0.356±0.006	**0.689±0.006**	0.217±0.004	**0.656±0.005**	0.5
Ukraine	**0.944±0.003**	0.289±0.006	**0.830±0.005**	0.169±0.004	**0.773±0.006**	0.5
Energy	**0.88±0.01**	0.117±0.004	**0.79±0.01**	0.062±0.003	**0.73±0.01**	0.5
NATO	**0.87±0.01**	0.084±0.004	**0.74±0.01**	0.044±0.002	**0.68±0.01**	0.5
Defense	**0.65±0.02**	0.096±0.004	**0.59±0.01**	0.051±0.002	**0.567±0.008**	0.5
Economic Sanctions	**0.92±0.03**	0.074±0.004	**0.55±0.01**	0.039±0.002	**0.530±0.006**	0.5
Refugees	**0.72±0.03**	0.04±0.003	**0.61±0.02**	0.02±0.001	**0.57±0.01**	0.5

Table 2: 2023 Balakitan Military Exercise: Performance of Concern Labeling

Performance metric	Precision		F1		ROC-AUC	
	Model	Baseline	Model	Baseline	Model	Baseline
None or other	**0.742±0.02**	0.533±0.016	**0.72±0.016**	0.696±0.014	**0.71±0.014**	0.5
Domestic political issues	**0.163±0.027**	0.048±0.007	**0.259±0.037**	0.092±0.013	**0.739±0.035**	0.5
International relations	**0.305±0.031**	0.106±0.01	**0.414±0.034**	0.191±0.016	**0.735±0.025**	0.5
US military	**0.448±0.026**	0.216±0.013	**0.565±0.025**	0.355±0.017	**0.754±0.016**	0.5
Public services	0.0±0.0	**0.004±0.002**	0.0±0.0	**0.008±0.004**	0.492±0.002	**0.5**
Defense and military	**0.637±0.052**	0.252±0.014	0.329±0.033	**0.402±0.018**	**0.59±0.014**	0.5
Social and economic issues	**0.285±0.081**	0.033±0.005	**0.277±0.073**	0.064±0.01	**0.626±0.04**	0.5
Environmental issues	**0.794±0.097**	0.04±0.006	**0.506±0.081**	0.077±0.012	**0.686±0.039**	0.5
Crime	**0.746±0.09**	0.021±0.005	**0.795±0.069**	0.041±0.009	**0.926±0.038**	0.5
Energy	—	—	—	—	—	—
Labor and migration	**0.604±0.102**	0.025±0.005	**0.573±0.088**	0.049±0.009	**0.773±0.052**	0.5
Insurgent threats	**0.105±0.107**	0.02±0.004	**0.063±0.062**	0.039±0.008	**0.52±0.024**	0.5

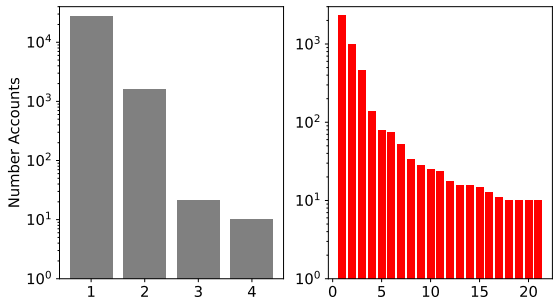

Figure 6: The number of coordinated accounts in each connected component (ranked by size). Accounts link with each other if they share the same long string of hashtags in each post. (a) 2022 French election and (b) 2023 Balikatan U.S.-Philippines military exercises.

counts (26.9K or 94% of all coordinated accounts). Examples of the coordinated clusters are shown in Fig. 7. In this figure, which shows the second largest cluster in the 2022 French election (top) and the largest cluster in the 2023 Balikatan dataset, we see examples of distinct accounts sharing near-identical posts, which are strong indicators of inauthentic online activity.

Interestingly, while the coordinated accounts from the 2022 French election were posts apparently directed towards Macron and the French public, we see many coordinated accounts unrelated to Balikatan within the latter dataset. This may represent a lack of information operations targeting the Balikatan event or that information operations were using language we did not capture in the keywords used to collect the posts, which is also represented by the relative lack of posts reposted or written by coordinated accounts.

We also analyze the concerns extracted from these coordinated accounts in Fig. 8. This figure shows the frequency of posts related to each concern we capture (note the log-scale of the y-axis); Russia is the most often mentioned concern. Meanwhile, the Balikatan dataset most often discusses domestic political issues and international relations. Compared to ground truth data (Fig 4), domestic political and public service concerns are over-represented.

But these concerns are highly selective in where they appear. We show examples of the percent of posts discussing two example concerns within each coordinated cluster in Fig. 9. These figures show the overabundance of the Ukraine concern within two clusters of the French election (Fig. 9a), which, as seen in Fig. 7, corresponds to pro-Ukraine messages. In contrast, Fig. 9b shows an overabundance of the domestic political concern, especially for the second largest coordinated cluster. This cluster contains an anti-Conservative UK party information campaign (i.e., the posts attack Tories).

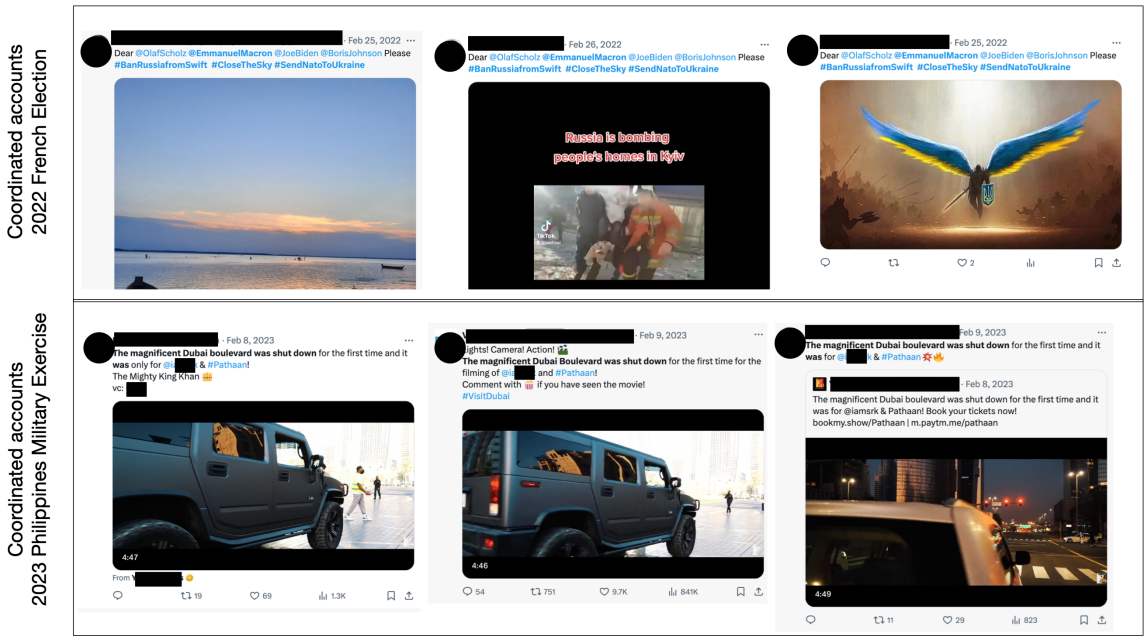

Figure 7: Example coordinated account posts for (top) the 2022 French election and (bottom) the 2023 Philippines Balikatan dataset.

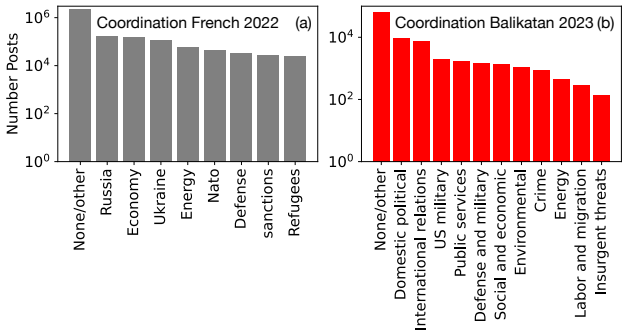

Figure 8: The frequency of concern posts in coordinated accounts. (a) 2022 French election and (b) 2023 Balikatan U.S.-Philippines military exercises.

Applying LLMs to Coordinated Account Posts

We will look at two coordinated campaigns as case studies. Due to space limitations we only briefly summarize how they change for various concerns, but we have all results in the link at the top of the paper. We explore two ways that these data are subdivided to better understand information operations: splitting these data by events (e.g., before versus after an election) and splitting these data by the concern the posts share. We qualitatively observe strong differences between what coordinated campaigns discuss before and after events, thus motivating the dynamic nature of our exploration. We also notice that concerns offer more nuance to the claims (e.g., why they are sharing pro-Ukraine concerns) compared to data that is not split by concerns.

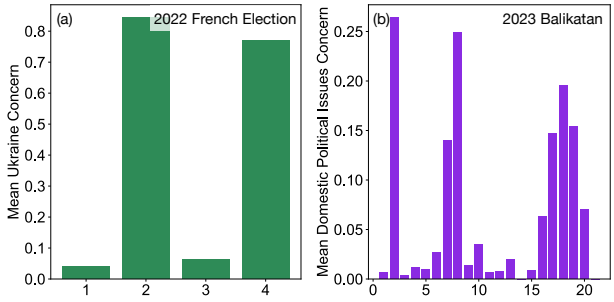

Figure 9: Example concerns for different clusters in the 2022 French election and Balikatan datasets. The x-axis are the cluster IDs, ranked from largest to smallest, where sizes are shown in Fig. 6. (a) Ukraine concern in the 2022 French election and (b) domestic political concern in the 2023 Philippines Balikatan dataset.

For the 2022 French election, we sampled 20 random posts before the election, between rounds 1 and 2, and after round 2. We specifically focus on accounts sharing the *Ukraine concern* to reduce noise in the captured posts (campaign we study in detail, shown in Fig 7 is heavily pro-Ukraine and requests assistance from France).

When studying the attributes from (Martin and Shapiro 2019), we notice, however, country attribution is imperfect, where before, during, and after the round 1 and 2 elections, the targeted country is Ukraine rather than Macron or France. The country responsible for the information campaign is either Russia (before the campaign) or France (between rounds 1 and 2, and after round 2). This suggests the LLM confuses attribution (claiming Russia is responsible for the campaign given it is mentioned in the posts). It narrows down the list of potential candidate countries, but the model nonetheless misunderstands the posts. That being said, the political goal is consistent across time (e.g., before round 1, the LLM states): "Influence international actors to take action against Russian aggression in Ukraine" which agrees with expectations.

When studying the BEND framework, we come across surprising utility but also limits of these zero-shot prompts. Namely, when asking for Explain tactic posts, we find the model appears to be spot on: after listing example posts, it says "These posts provide specific examples of the destruction caused by Russian aggression in Ukraine and call for action to protect Ukrainian civilians. They elaborate on the situation and highlight the urgency of the crisis.". In contrast, when asking for the Excite tactic prior to the first round, an example output is:

"...2. @EmmanuelMacron NOW!!! RUSSIA FIRES ON NUCLEAR POWER PLANT IN #UKRAINE!!! STREAM: https://XXXX WAKE UP, WORLD! #StopPutin #StopRussia CLOSE THE SKY OVER UKRAINE!!!"

"This tweet uses capital letters and exclamation marks to convey a sense of urgency and excitement in calling for action to stop Russian aggression against Ukraine."

This shows how the LLM mistakes activation for excitement and misunderstands whether positive sentiment exists in the message. These are seen in other examples, such as the Distract tactic, where the LLM states "These tweets distract from the urgent situation in Ukraine by discussing EU candidate status for Ukraine, calling out a specific individual (Putin), and expressing gratitude for support without any further context related to the conflict. By introducing irrelevant topics, they divert attention away from the main issue at hand which is the ongoing crisis in Ukraine." The misunderstanding of distraction may be due to a lack of "typical" (non-coordinated) posts listed as counter-examples. In addition, few-shot and Chain-of-Thought prompting might further improve the output quality (Zhang et al. 2022).

Finally, when analyzing the framing narrative prompts, we find the LLM is particularly well-suited. The problem is stated as "Russian aggression against Ukraine..." before the round 1 election, and similarly in other time periods. The cause is "...Russian aggression against Ukraine" which may be due to a lack of context, but it's also unclear what a human might say. Finally, the remedy is: "1. Stop Russian aggression against Ukraine. 2. Close the sky over Ukraine. 3. Exclude Russia from the UN Security Council. 4. Ban Russia from SWIFT. 5. Protect the Ukrainian sky. ..." etc., which is a good summary of the results. Catchphrases and slogans include "STOP RUSSIAN AGGRESSION AGAINST #UKRAINE, CLOSE THE SKY OVER UKRAINE, EXCLUDE RUSSIA FROM THE @UN SECURITY COUNCIL, #StopPutin" etc. (taken directly from the posts fed into the model), which shows slogans in the posts double as remedies.

Due to limited space, we will briefly mention that later stages, such as after round 2, we see more plausible and concrete remedies as the war progressed such as, "1. Urging world leaders, particularly Emmanuel Macron, to negotiate a safe passage for the civilians trapped in Mariupol to evacuate and receive humanitarian aid. 2. Providing weapons and support to Ukraine to help defend against Russian aggression and protect the civilians." etc.,

For the 2023 Balikatan dataset, we analyze posts from the largest coordinated cluster before and after the Balikatan military exercise. This campaign promotes Pathaan, a Bollywood movie. While this is a coordinated campaign, it lies outside of typical campaigns studied by (Martin and Shapiro 2019) and is more akin to spam. We study the None or Other concern to focus on the most relevant posts for this set of coordinated accounts (which do not discuss otherwise common concerns, such as Crime or US Military). GPT-3.5 states that the country targeted (and the country responsible for the information operation) is consistently India. While the government of India is likely not responsible, the audience and potential coordinators of the information campaign are likely within India due to the genre of the movie. Before April 11, 2023, the political goal is stated as "Promoting the Bollywood movie 'Pathaan"' which hints at another utility of the GPT-3.5 outputs: with one prompt, the political goal, we can immediately determine that those accounts were not related to "important" coordinated accounts (i.e., those about Balikatan or targeting democracies). We can also easily understand the underlying role of these spam accounts because after April 28, 2023, they were pushing messages about "Adipurush", a separate Indian movie. Moreover, there appears little overlap in the directors or producers to each film allowing us to speculate that this set of X accounts is run by a group that offers their services to a number of different organizations. We also find the LLM extracts relevant BEND tactics, such as posts promoting excitement (e.g., "'...book now!"). Finally, framing theory prompts reveal: "The cause of the problem mentioned in these tweets is the high demand and anticipation for the movie 'Pathaan.'..." and "The remedy to the problem mentioned in the tweets is for individuals to book their tickets in advance for the movie 'Pathaan"' which shows the accounts are trying to drive ticket sales.

Discussion and Conclusions

To summarize, we worked with domain experts to curate a list of concerns associated with two distinct datasets, a 2022 French Election dataset and a 2023 Balikatan dataset. We then developed a framework to detect these concerns using an instruction-tuned Llama-2 LLM. We find that the

Llama-based concern detection model achieves high accuracy, making it suitable for our analysis. We then applied GPT-3.5 to annotate coordinated campaigns that contained each of these concerns. Namely, we use GPT-3.5 to extract the goals, tactics, and frames that are based on information operation descriptions from an information operations dataset (Martin and Shapiro 2019), the BEND framework (Carley 2020), and Framing Theory (Chong and Druckman 2007). GPT-3.5 was first evaluated against 126 known campaigns and achieved reasonable accuracy. We then detected coordinated campaigns within the two datasets via a standard hashtag-based method (Burghardt et al. 2023), and annotated campaigns via the GPT-3.5 LLM. These posts were first filtered by the concerns they shared and when these posts were made to help distill the main properties of dynamic information operations. When GPT-3.5 was applied to these campaigns, we found the LLM was effective at extracting the goals, tactics, and narrative frames, although it made notable errors. While GPT-3.5 is not always correct, the results are still informative for large sets of coordinated accounts. LLMs appear likely to help stake-holders find and understand the behavior of various information operations at scale and find common patterns across information operations, which could otherwise be missed from purely manual annotation. While LLMs still require a human in the loop to interpret these results, we find in practice that verifying LLM annotations is much faster than humans creating their own annotations, much like how LLMs can speed up letter-writing, coding, and other tasks that still require a human in the loop.

Limitations

There are a few important limitations of our work. First, LLMs can hallucinate. Detecting hallucinations is hard to automate at present, therefore we need a human in the loop to verify any output. Efforts to detect or address hallucinations, especially to annotate coordinated campaigns, needs to be analyzed. Moreover, the accuracy of a LLM depends strongly on carefully constructed prompts. The degree to which semantically equivalent prompts may affect the output of the model needs to be further explored.

In addition, the low performance of the concern detection model on some concerns such as *refugees* and *defense* suggests that Llama or GPT-4 (whose training data we rely on) may fail to understand the nuance of these concerns. Finally, the reproducibility of results depends strongly on the model version and the stochasticity (e.g., temperature) of the output. These are challenges for any replication study, especially for LLMs that are regularly updated.

Ethical Considerations

Our work utilizes public data on information campaigns and anonymized social media data, therefore we believe the harm of this research on any human subjects is minimal. The potential harm of this research, however, is that we might provide ways for information campaigns to better hide their underlying goals, or to scrutinize other coordinated campaigns to improve their influence operation. We believe, however, that these potential harms are minimal as the information operations already try to hide their behavior to avoid their accounts being suspended, and the campaign operators could already scrutinize known campaigns that have been publicly reported. In addition, there might be costs to misclassifying accounts as coordinated or LLMs misunderstanding a coordinated campaign by flagging or misunderstanding text from authentic accounts. Care must therefore be taken to assume individual accounts are authentic until proven otherwise.

Future work

There are a range of ways to improve the concern detection and LLM annotation models.

The concern detection model could be improved with additional human-annotated training data. This can be applied as gold standard instruction-tuning training data, for GPT-4 few-shot learning prompts in order to improve the synthetic training data, or to up-sample concerns with very few annotated post (such as public services or energy concerns in the Balikatan dataset) in order to better validate any future model. Newer models, such as Llama-3 (Huang et al. 2024) or larger models, such as Llama-2-70B could also improve the concern detection accuracy and should be explored in the future.

There are also a number of ways to improve coordination campaign annotation. First, we can provide few-shot or chain-of-thought (Zhang et al. 2022) prompting to improve the accuracy of the output. In addition, a collection of domain experts should independently verify whether the model can, for example, approximately reconstruct the major features of known coordinated campaigns (Martin and Shapiro 2019), as well as unknown campaigns. This will improve the validation of our observations. Finally, we can use larger models, such as GPT-4 (Achiam et al. 2023), to improve the quality of the annotations.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ayala, F. J. 2010. The difference of being human: Morality. *PNAS*, 107(supplement_2): 9015–9022.

Badawy, A.; Addawood, A.; Lerman, K.; and Ferrara, E. 2019. Characterizing the 2016 Russian IRA influence campaign. *Social Network Analysis and Mining*, 9(1): 1–11.

Badawy, A.; Ferrara, E.; and Lerman, K. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In *ASONAM*, 258–265.

Bai, K.; Wang, G.; Li, J.; Park, S.; Lee, S.; Xu, P.; Henao, R.; and Carin, L. 2023. Open world classification with adaptive negative samples. *arXiv preprint arXiv:2303.05581*.

Bail, C. A.; Guay, B.; Maloney, E.; Combs, A.; Hillygus, D. S.; Merhout, F.; Freelon, D.; and Volfovsky, A. 2020. Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *PNAS*, 117(1): 243–250.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Bessi, A.; and Ferrara, E. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday*, 21(11-7).

Bradshaw, S.; and Howard, P. N. 2019. The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation.

Burghardt, K.; Rao, A.; Guo, S.; He, Z.; Chochlakis, G.; Sabyasachee, B.; Rojecki, A.; Narayanan, S.; and Lerman, K. 2023. Socio-Linguistic Characteristics of Coordinated Inauthentic Accounts. *arXiv preprint arXiv:2305.11867*.

Carley, K. M. 2020. Social cybersecurity: an emerging science. *Computational and Mathematical Organization Theory*, 26(4): 365–381.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Chen, C.-F.; Shi, W.; Yang, J.; and Fu, H.-H. 2021. Social bots' role in climate change discussion on Twitter: Measuring standpoints, topics, and interaction strategies. *Advances in Climate Change Research*, 12(6): 913–923.

Chen, K.; He, Z.; Burghardt, K.; Zhang, J.; and Lerman, K. 2024a. IsamasRed: A Public Dataset Tracking Reddit Discussions on Israel-Hamas Conflict. *arXiv preprint arXiv:2401.08202*.

Chen, K.; He, Z.; Yan, J.; Shi, T.; and Lerman, K. 2024b. How Susceptible are Large Language Models to Ideological Manipulation? *arXiv preprint arXiv:2402.11725*.

Chong, D.; and Druckman, J. N. 2007. Framing theory. *Annu. Rev. Polit. Sci.*, 10: 103–126.

Cinelli, M.; Cresci, S.; Quattrociocchi, W.; Tesconi, M.; and Zola, P. 2022. Coordinated Inauthentic Behavior and Information Spreading on Twitter. *Decis. Support Syst.*, 160(C).

de Winter, J. C. F. 2023. Can ChatGPT Pass High School Exams on English Language Comprehension? *IJAIED*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *ACL*, 4171–4186.

Ehrett, C.; Linvill, D. L.; Smith, H.; Warren, P. L.; Bellamy, L.; Moawad, M.; Moran, O.; and Moody, M. 2021. Inauthentic Newsfeeds and Agenda Setting in a Coordinated Inauthentic Information Operation. *Soc. Sci. Comput. Rev.*, 0(0).

Entman, R. M. 1993. Framing: Toward clarification of a fractured paradigm. *J. Commun.*, 43(4): 51–58.

Esmaeilpour, S.; Liu, B.; Robertson, E.; and Shu, L. 2022. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *AAAI*, volume 36, 6568–6576.

Ferrara, E. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday*, 22(8).

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Giglietto, F.; Righetti, N.; Rossi, L.; and Marino, G. 2020a. Coordinated Link Sharing Behavior as a Signal to Surface Sources of Problematic Information on Facebook. In *SMSociety*, 85–91.

Giglietto, F.; Righetti, N.; Rossi, L.; and Marino, G. 2020b. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. *Information, Communication & Society*, 23(6): 867–891.

Graham, T.; Bruns, A.; Zhu, G.; and Campbell, R. 2020. *Like a virus: The coordinated spread of Coronavirus disinformation*. Canberra, A.C.T: The Australia Institute.

Gudibande, A.; Wallace, E.; Snell, C.; Geng, X.; Liu, H.; Abbeel, P.; Levine, S.; and Song, D. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Heinkelmann-Wild, T.; Kriegmair, L.; Rittberger, B.; and Zangl, B. 2020. Divided they fail: The politics of wedge issues and Brexit. *Journal of European Public Policy*, 27(5): 723–741.

Hickey, D.; Schmitz, M.; Fessler, D.; Smaldino, P. E.; Muric, G.; and Burghardt, K. 2023. Auditing Elon Musk's Impact on Hate Speech and Bots. *ICWSM*, 17(1): 1133–1137.

Hon, L. 2016. Social media framing within the Million Hoodies movement for justice. *Public Relations Review*, 42(1): 9–19.

Howard, P. N.; and Kollanyi, B. 2016. Bots,# StrongerIn, and# Brexit: computational propaganda during the UK-EU referendum. *arXiv preprint arXiv:1606.06356*.

Huang, W.; Ma, X.; Qin, H.; Zheng, X.; Lv, C.; Chen, H.; Luo, J.; Qi, X.; Liu, X.; and Magno, M. 2024. How Good Are Low-bit Quantized LLaMA3 Models? An Empirical Study. *arXiv preprint arXiv:2404.14047*.

Katz, D. M.; Bommarito, M. J.; Gao, S.; and Arredondo, P. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270): 20230254.

Kim, Y. M. 2018. Uncover: strategies and tactics of Russian interference in US elections. *Young*, 9(04).

Kirdemir, B.; Adeliyi, O.; and Agarwal, N. 2022. Towards Characterizing Coordinated Inauthentic Behaviors on YouTube. In *ROMCIR*. Stavanger, Norway.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Liyanage, C.; Gokani, R.; and Mago, V. 2023. GPT-4 as a Twitter Data Annotator: Unraveling Its Performance on a Stance Classification Task. *Authorea Preprints*.

Luceri, L.; Boniardi, E.; and Ferrara, E. 2023. Leveraging Large Language Models to Detect Influence Campaigns in Social Media. *arXiv preprint arXiv:2311.07816*.

Luceri, L.; Pantè, V.; Burghardt, K.; and Ferrara, E. 2023. Unmasking the web of deceit: Uncovering coordinated activity to expose information operations on twitter. *arXiv preprint arXiv:2310.09884*.

Martin, D. A.; and Shapiro, J. N. 2019. Trends in online foreign influence efforts.

Martínez, C. R. 2023. Examining the Role of Wedge Issues in Shaping Voter Behavior: Insights From the 2020 US Presidential Election. *Comillas Journal of International Relations*, (27): 101–121.

Mazza, M.; Cola, G.; and Tesconi, M. 2022. Ready-to-(ab)use: From fake account trafficking to coordinated inauthentic behavior on Twitter. *OSNM*, 31: 100224.

Mishra, S.; Khashabi, D.; Baral, C.; and Hajishirzi, H. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Nizzoli, L.; Tardelli, S.; Avvenuti, M.; Cresci, S.; and Tesconi, M. 2021. Coordinated Behavior on Social Media in 2019 UK General Election. *ICWSM*, 15(1): 443–454.

Pacheco, D.; Hui, P.; Torres-Lugo, C.; Truong, B. T.; Flammini, A.; and Menczer, F. 2021. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. *ICWSM*, 21: 455–466.

Paper, O. W. 2022. Suspicious Twitter Activity around the Russian Invasion of Ukraine.

Piña-García, C. A.; and Espinoza, A. 2022. Coordinated campaigns on Twitter during the coronavirus health crisis in Mexico. *Tapuya: Latin American Science, Technology and Society*, 0(0): 2035935.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Goncalves, B.; Flammini, A.; and Menczer, F. 2021. Detecting and Tracking Political Abuse in Social Media. *ICWSM*, 5(1): 297–304.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T. L.; Raja, A.; et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Sayyadiharikandeh, M.; Varol, O.; Yang, K.; Flammini, A.; and Menczer, F. 2020. Detection of Novel Social Bots by Ensembles of Specialized Classifiers. In *CIKM*, 2725–2732.

Schliebs, M.; Bailey, H.; Bright, J.; and Howard, P. 2021. China's inauthentic UK Twitter diplomacy: a coordinated network amplifying PRC diplomats.

Shahid, U.; Di Eugenio, B.; Rojecki, A.; and Zheleva, E. 2020. Detecting and understanding moral biases in news. In *WNU 2020*, 120–125.

Sharma, K.; Zhang, Y.; Ferrara, E.; and Liu, Y. 2021. Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours. In *KDD*, 1441–1451.

Shu, L.; Xu, H.; and Liu, B. 2018. Unseen class discovery in open-world classification. *arXiv preprint arXiv:1801.05609*.

Starbird, K. 2019. Disinformation's spread: bots, trolls and all of us. *Nature*, 571(7766): 449–450.

Stella, M.; Ferrara, E.; and Domenico, M. D. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *PNAS*, 115(49): 12435–12440.

Tao, Z.; Jin, Z.; Bai, X.; Zhao, H.; Feng, Y.; Li, J.; and Hu, W. 2023. Eveval: A comprehensive evaluation of event semantics for large language models. *arXiv preprint arXiv:2305.15268*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tucker, J. A.; Guess, A.; Barberá, P.; Vaccari, C.; Siegel, A.; Sanovich, S.; Stukal, D.; and Nyhan, B. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization*.

Van de Wardt, M.; De Vries, C. E.; and Hobolt, S. B. 2014. Exploiting the cracks: Wedge issues in multiparty competition. *The Journal of Politics*, 76(4): 986–999.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*, 359(6380): 1146–1151.

Weber, D.; and Falzon, L. 2021. Temporal Nuances of Coordination Network Semantics.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Yao, J.-Y.; Ning, K.-P.; Liu, Z.-H.; Ning, M.-N.; and Yuan, L. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Zhang, S.; Dong, L.; Li, X.; Zhang, S.; Sun, X.; Wang, S.; Li, J.; Hu, R.; Zhang, T.; Wu, F.; et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures?
   Yes, see Ethical considerations.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope?
   Yes, for example, we made sure to state both the promise and limitations of our results throughout the paper.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made?
   Yes, we were careful to state the purpose of our methods and its limitations.

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions?
   Yes, see Ethical Considerations.

(e) Did you describe the limitations of your work?
Yes, see Limitations within the Discussion and Conclusions.

(f) Did you discuss any potential negative societal impacts of your work?
Yes, see Ethical Considerations.

(g) Did you discuss any potential misuse of your work?
Yes, see Ethical Considerations.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings?
Yes, see Ethical Considerations.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them?
Yes.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results?
NA.

(b) Have you provided justifications for all theoretical results?
NA.

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results?
NA.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study?
NA.

(e) Did you address potential biases or limitations in your theoretical framework?
NA.

(f) Have you related your theoretical results to the existing literature in social science?
NA.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain?
NA.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results?
NA.

(b) Did you include complete proofs of all theoretical results?
NA.

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
Yes, see code link in Introduction.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
Yes, see Methods and code.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?
Yes, see Tables 1 and 2.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?
Yes, see Methods.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made?
Yes, see Results and Discussion.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance?
Yes, see Ethical Considerations.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators?
NA.

(b) Did you mention the license of the assets?
NA.

(c) Did you include any new assets in the supplemental material or as a URL?
Yes, see

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating?
Yes, see Data subsection with in Methods.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
Yes, see Data subsection with in Methods.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))?
NA.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))?
NA.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots?
NA.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals?
Yes, see Methods for IRB approval and Ethical considerations for potential participant risks.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant

compensation?
NA.

(d) Did you discuss how data is stored, shared, and deidentified?
Yes.

Appendix

Prompting Template

```
You are using an AI model specialized
in classifying tweets into relevant
[event topic] concern (wedge issues)
categories. The available concern
categories are [concern categories].
Your task is to analyze the following
tweet and categorize the following
tweet into one or more of these concern
categories. If the tweet does not align
with any of these categories, label it
as #NoneOther. Please respond with only
the relevant category labels.

Tweet: [tweet]
```

Instruction Template

```
Instruction: You are using an AI model
specialized in classifying tweets into
relevant [event topic] concern (wedge
issues) categories. The available
concern categories are [concern
categories]. Your task is to analyze
the following tweet and categorize
the following tweet into one or more
of these concern categories. If the
tweet does not align with any of these
categories, label it as #NoneOther.
Please respond with only the relevant
category labels.

Input: [tweet]

output: [labels]
```

LLM X post generation examples

Political goal	LLM-generated post
Polarize Australian politics	Why is #Australia letting in so many Muslim refugees when they bring nothing but trouble? #StopIslam
Discredit Sebastian Kurz after the 2017 Presidential elections in Austria	BREAKING: Shocking revelations have emerged about Sebastian Kurz's connections to Islamic immigration! #DiscreditKurz #Austria
Undermine Belarusian government	Just heard that Russia offered Belarus a sweet deal to join forces. Maybe it's time for change #BelarusStrong

Table 3: Political goals in various information operations (Martin and Shapiro 2019) and LLM-generated posts for these campaigns.