Pictures are Worth a Thousand Frames: The Impact of Images on the Discovery of Frames of Communication from Multimodal Social Media

Anonymous ACL submission

Abstract

Frames of Communication (FoCs) are evoked in multiple Social Media Postings (SMPs) that contain not only text, but also images. In this paper we introduce DA-FoC^{MM}, a new method capable of uncovering and articulating the FoCs evoked in SMPs while also pinpointing whether the FoC is evoked in the SMP text, image, or both. The DA-FoC^{MM} method successfully discovers FoCs from multimodal SMPs discussing two different controversial topics, namely COVID-19 vaccines and immigration, by using several constrained prompting approaches that determine the combination of counterfactual reasoning with Chain-of-Thought (CoT) reasoning performed by a Language Multimodal Model (LMM). In addition, we show that DA-FoC^{MM} enables the discovery of FoCs from multimodal SMPs across two platforms: Twitter / X and Instagram. Evaluations produced promising results, showing that 90%-91% of the FoCs identified by communication experts on the same collections of SMPs were also discovered by the method presented in this paper. We also found that 39% of FoCs would not have been discovered if the images from SMPs had been ignored.

Figure 1: A Frame of Communication (FoC) evoked in a Multimodal Social Media Posting (SMP). The SMP text and image address the same problem.

1 Introduction

In a polarized world like the one in which we live now, controversial communications are abundant on social media. Identical information can be presented, or "framed", in various manners (Keren, 2011), which can significantly impact how that information is interpreted. For instance: abortion can be framed as pro-life or pro-choice. An audience is differently primed based on the addressed problems, such as the sanctity of life or individual autonomy, as well as the causes ascribed to addressed problems (Rohlinger, 2002; Sonnett, 2019).

Frames evoked in social media communications refer to *problems*, or salient aspects of a controversial topic (e.g. abortion), addressing the causes of those problems, as a minimum, cf. (Entman, 2003; Reese et al., 2001; Scheufele, 2004; Chong and Druckman, 2012; Bolsen et al., 2014). But, how can Frames of Communication (FoCs) be discovered automatically, when communications involve not only texts, but also images, as it happens nowadays on social media? While recent work (Weinzierl and Harabagiu, 2024a) has shown significant promise for automatically discovering and articulating FoCs from textual Social Media Postings (SMPs), no research has yet addressed the problem of discovering FoCs when images are also present in SMPs. This was the focus of the research reported in this paper.

Figure 1 illustrates an SMP that contains both text and an image. The Figure also illustrates the problem addressed by the multimodal SMP, namely confidence in COVID-19 vaccines, one of the problems surrounding the controversial topic of vaccination. Moreover, Figure 1 shows the FoC that is evoked by the SMP. The text of the SMP in isolation appears to, at face value, communicate confidence in the COVID-19 vaccine because *"...you can trust the science and big pharma"*. However, the SMP's author is implying through the image that, just like how *"the science"* and *"big pharma"* sold smoking as safe, and even good for, pregnancies, the current recommendations for pregnant women to take the COVID-19 vaccine should not be trusted. Hence, based on the sarcasm used in the image, the illustrated FoC is evoked by the entirety of the SMP, spreading the misinformation that the COVID-19 vaccine is risky and unsafe for pregnant women and babies. Therefore, pictures included in SMPs enable the evocation of many FoCs, which the text alone would not evoke. Since pictures are said to be worth a thousand words, they can also be considered worth a thousand FoCs, hence the pun used in the title of our paper.

Evidently, automatically identifying the problem addressed by the text as well as the image of the SMP illustrated in Figure 1, namely confidence in vaccines, is not trivial. Furthermore, the automatic discovery and articulation of the evoked FoC is also extremely challenging, relying on (a) *reasoning* with cultural knowledge (e.g. smoking was once sold as beneficial for pregnant women); (b) analogical *reasoning* (e.g. as smoking was once considered safe, so is vaccination considered now); (c) commonsense *reasoning* (e.g. why would vaccination not be deemed harmful later for pregnant women, similarly to smoking?); and (d) *reasoning* with complex relationships between the text and the image of the SMP (e.g. a transcendent relation, where the image analogy to smoking entailing perils in vaccinating pregnant women picks up and contradicts the text). Therefore, to discover and articulate an FoC, similar to the one illustrated in Figure 1, we need to explore how we can elicit from Large Multimodal Models (LMMs) all these forms of knowledge and reasoning, and more importantly, to *link them together* for articulating the FoC. Moreover, given that each controversial problem is addressed by multiple FoCs, as shown in Figure 2, while each FoC is evoked by multiple multimodal SMPs, the discovery and the articulation of FoCs requires forms of reasoning that are distinct from those used by Visual Question Answering systems (Xenos et al., 2023; Subramanian et al., 2023; Dong et al., 2024), which focus on finding answers in images or combinations of texts and images as response to a question. To discover FoCs, we need to perform implicit reasoning on the multimodal content of multiple SMPs that evoke the same FoC, which is not known apriori. In addition, we need to articulate the FoC, while also reasoning to discover the implied problems it addresses.

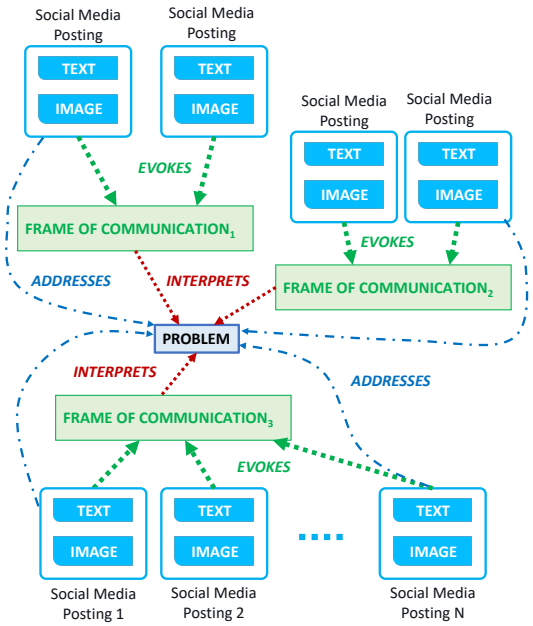

Figure 2: Several Frames of Communication (FoCs) provide different interpretations of the same problem, addressed by many multimodal Social Media Postings (SMPs). Each FoC is evoked by multiple SMPs.

Recently, Large Language Models (LLMs) have been used successfully to discover and articulate FoCs addressing controversial problems (Weinzierl and Harabagiu, 2024a) only from textual SMPs. FoCs and the problems they address were discovered by relying on a combination of curriculum learning, reasoning elicited by Chain-of-Thought (CoT) prompting (Wei et al., 2022), and active learning. However, curriculum learning operating on a multimodal dataset of SMPs and the FoCs they evoke is very challenging. Moreover, CoT prompting has been shown to struggle to reason with complex relations between text and images, cf. (Chen et al., 2024b,a). In addition, the CoT capabilities to reason with commonsense knowledge and perform analogical reasoning are subject to substantial human effort required to produce many demonstrations.

In this paper, we present a novel method that surmounts these limitations. Our first innovation provides an alternative to human-generated demonstrations. Instead, we consider automatically generated demonstrations. These demonstrations *explain* (a) why a controversial problem can be inferred from the text and/or image of an SMP; and (b) why an FoC is evoked from a multimodal SMP. Importantly, these explanations result from the combination of counterfactual prompting of LMMs, known to successfully produce explanations, cf. (Jacovi et al., 2021; He et al., 2022; Chen et al., 2023; Weinzierl and Harabagiu, 2024b) with (b) Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) which selects the demonstrations to be provided to CoT prompting.

Our second innovation consists of the usage of a novel constrained decoding approach (Zheng et al., 2023; Yang et al., 2023) for prompting LMMs, to enable the generation of detailed, **structured explanations** of the controversial problems implied in each multimodal SMP and of the evoked FoCs.

Our method uses these two innovations to Discover and Articulate FoCs from MultiModal SMPs, being named **DA-FoC**^{MM}. Our paper makes the following contributions:
◁**1**▷ We introduce the first method capable to discover and articulate FoCs from multimodal SMPs.
◁**2**▷ We introduce several constrained prompting approaches for LMMs to answer questions about *what*, *why* and *where* (a) controversial problems are addressed and (b) FoCs are evoked in SMPs.
◁**3**▷ We show that the combination of CoT reasoning with counterfactual reasoning helps the discovery of FoCs from multimodal SMPs.
◁**4**▷ We show that the DA-FoC^{MM} method operates successfully on SMPs discussing two different topics, allowing us to introduce a new dataset of multimodal SMPs annotated with the problems they address and the FoCs they evoke.
◁**5**▷ We explore how the DA-FoC^{MM} method can be used successfully across social network platforms.
◁**6**▷ The evaluation results indicate that 39% of FoCs discovered by the DA-FoC^{MM} method would not have been identified if the images from SMPs would have been ignored. Therefore, we provide a quantitative evaluation of the impact of images in discovering FoCs from social media.
◁**7**▷ We make available all code, prompts, annotations, and discovered FoCs on GitHub[1].

2 Datasets

In our experiments, we considered four datasets of multimodal SMPs:
Dataset 1: To our knowledge, the only existing dataset containing multimodal SMPs annotated with the (1) controversial problems they address; as well as (2) the FoCs they evoke is MMVAX-STANCE, reported and released in Weinzierl and Harabagiu (2023). This dataset contains 11,300 SMPs from Twitter / X addressing 7 possible problems and interpreted by 113 evoked FoCs. Details of the problem definitions, examples of annotated FoCs and discussion of the annotations are available in Appendix A. We note that from this dataset, we consider as a *Reference Dataset RF1* only the training split containing 5,464 SMPs, evoking 113 FoCs, which interpret all 7 problems. All the other multimodal SMPs from MMVAX-STANCE were considered as the *Test Dataset TS1*.
Dataset 2: Considering the same topic as in Dataset 1, namely COVID-19 vaccine hesitancy, we created a second dataset of 1,289 Instagram SMPs that are likely to evoke the same 113 FoCs annotated in RF1. We note that there are no annotations produced on this dataset, therefore it may be considered in its entirety as *Test Dataset TS2*. The manner in which TS2 was built is detailed in Appendix A.
Dataset 3: A new dataset of 1,878 multimodal SMPs from Twitter / X addressing the topic of immigration was annotated with the 27 problems introduced by Mendelsohn et al. (2021) and 57 newly-discovered FoCs. Details of the problem definitions, examples of annotated FoCs, and discussion of the annotations are available in Appendix A. A *Reference Dataset RF2* of 939 SMPs that evoke 57 FoCs was built from Dataset 3. All the other multimodal SMPs from this dataset were considered as the *Test Dataset TS3*.
Dataset 4: Considering the same topic as in Dataset 3, namely immigration, we created a fourth dataset of 956 Instagram SMPs that are likely to evoke the same 57 FoCs annotated in RF2. As with dataset 2, there are no annotations on this dataset, therefore it may be considered in its entirety as *Test Dataset TS4*. The manner in which TS4 was built is also detailed in Appendix A, along with examples.

[1] https://anonymous.4open.science/r/da-foc-mm-8817

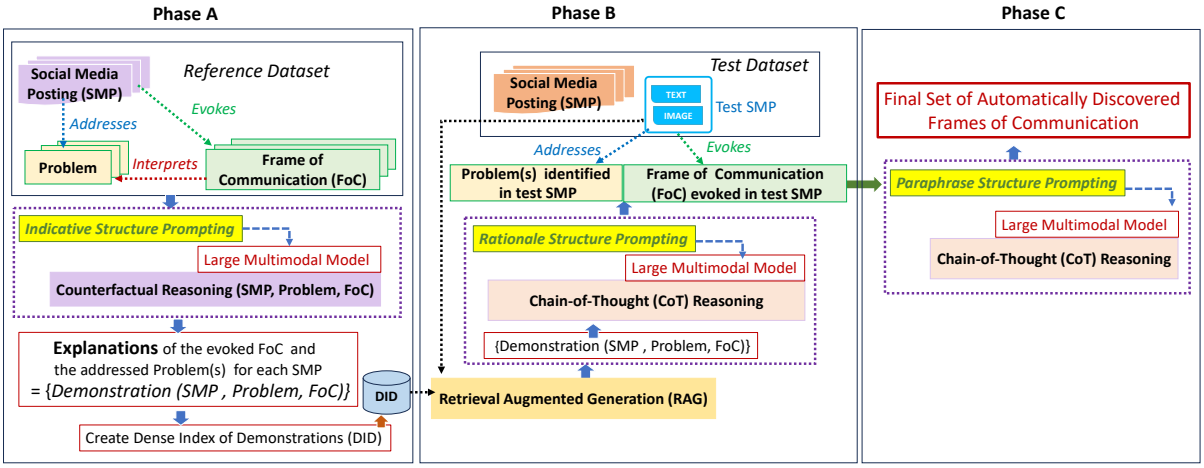

Figure 3: The architecture for Discovering and Articulating Multimodal Frames of Communication (DA-MFoC).

3 The Method

The DA-FoC^{MM} method operates in three distinct phases, each of them using a different prompting of the LMM, as illustrated in Figure 3. Phase A is informed by the Reference Dataset, e.g. dataset RF1 or RF2, consisting of a collection of multimodal SMPs, annotated with the problems they addressed and the FoCs they evoke. The Reference Dataset is used for generating explanations for the evoked FoCs and the addressed problems. The explanations are produced with counterfactual reasoning of an LMM, along with a special form of prompting, namely *indicative structure prompting*, detailed in Section 3.1. All obtained explanations are indexed in a Dense Index of Demonstrations (DID).

Phase B of DA-FoC^{MM} uses the Test Dataset, e.g. datasets TS1, TS2, TS3 or TS4, which contains only multimodal SMPs. The goal of this phase is to discover for each SMP_{Test} the problems it addresses articulate the FoCs it evokes. Therefore, given an SMP_{Test}, Retrieval Augmented Generation (RAG) operates on the DID to provide the demonstrations required by the Chain-of-Thought (CoT) reasoning, along with another special form of prompting, namely *rationale structure prompting*, detailed in Section 3.2.

Because sometimes FoCs discovered in Phase B are evoked by multiple SMPs from the Test Dataset, or paraphrase each other, Phase C of DA-FoC^{MM} has the goal to identify and filter out such paraphrases. Therefore *paraphrase structure prompting* of the LMM, detailed in Section 3.3, is used to discover the final set of FoCs.

3.1 Phase A

To automatically generate explanations for the problems addressed in the SMPs available in the Reference Dataset, as well as explanations for the FoCs the SMPs evoke, we have considered a special flavor of counterfactual reasoning. Counterfactual reasoning generally involves examining alternatives to facts, events, or states, drawing inferences about what could have occurred or been possible. For each alternative, explanations can be generated by an LMM, with Chain-of-Explanation (CoE) prompting, cf. (Weinzierl and Harabagiu, 2024b). However, this entails access to all counterfactual possibilities, which is not feasible, as we cannot be aware of all possible FoCs that can be evoked by an SMP. Alternatively, the Reference Dataset gives us *indications* of which FoCs are evoked by a particular SMP, as well as which problems are addressed both by the SMP and the FoC. Therefore, instead of using counterfactuals for eliciting explanations from an LMM, we use indications available from the Reference Dataset to ask for explanations. We consider this a special flavor of counterfactual reasoning.

The Reference Dataset can be viewed as containing indications I_i that connect each SMP $s_i = [t_i, v_i]$, consisting of a text t_i and an image v_i, with all FoCs $\{f^i_j\}$ evoked by s_i as well as all problems $\{p^i_k\}$ addressed by the pair $< s_i, f^i_j >$. Consequently, the *Indicative Structure Prompting* of the LMM, used for generating explanations for all problems $\{p^i_k\}$ and of all FoCs $\{f^i_j\}$ from I_i asks:
□**1**□ *Why* is each problem p^i_k addressed by s_i?
□**2**□ *Where* is p^i_k addressed: in t_i, in v_i, or in both?
□**3**□ *Why* is each FoC f^i_j evoked by s_i?

To implement the Indicative Structured Prompting we relied on a constrained decoding approach utilizing a "JSON schema", detailed in Appendix B.

In response to the Indicative Structured Prompting, the LMM generates for each problem p_k^i of I_i a rationale r_k^i which explains why p_k^i is addressed by the SMP s_i. The LMM also generates for each FoC f_j^i a rationale $re_{i,j}$ explaining why s_i evokes f_j and it also pinpoints where each FoC is evoked: in t_i, in v_i, or in both. Since each indication I_i has its own structure $I_i = [s_i, \{p_k^i\}, \{f_j^i\}]$, the rationales generated by the LMM for each I_i are created as structured explanations:

$SE_i = [s_i =< t_i, v_i >, \{p_k^i$ *and its rationale* $r_k^i\}$, $\{f_j^i$ *and its rationale* re_j^i *along with pinpointing whether* f_i^i *is evoked in* t_i, *in* v_i *or in both*$\}]$.

An example of the operation of indicative structure prompting is provided in Appendix C.

To select which demonstrations should be considered when performing CoT prompting using an SMP from the Test Dataset in Phase B of the DA-FoC^{MM} method, we also generate in Phase A a Dense Index of Demonstrations (DID). To build the DID, for each SMP s_i we produced an embedding of its text t_i with a CLIP (Radford et al., 2021) text encoder, generating the embedding e_i^t, while for its image v_i we used a CLIP image encoder, generating an embedding e_i^v. These embeddings are concatenated as $e_i = [e_i^t; e_i^v]$ and added to a dense FAISS index (Johnson et al., 2019). A link is generated from each e_i to its corresponding SE_i. Additional details are provided in Appendix C.

3.2 Phase B

Phase B operates on the Test Dataset, containing SMPs with no annotations. For each SMP s_i^T, consisting of a text t_i^T and an image v_i^T, the goal is to discover and articulate $\{f_j^T\}$, all its evoked FoCs, where each f_j^T is interpreting some problem p_k^T addressed in s_i^T, which also needs to be identified. By using CoT reasoning, the LMM not only identifies the problems addressed by each f_j^T, namely $\{p_k^T\}$, as well as all the evoked f_j^T, but it also generates detailed rationales for them. For this we used *Rationale Structure Prompting:*

⊙**1**⊙ *What* problems $\{p_k^T\}$ are addressed by s_i^T?
⊙**2**⊙ *Why* is each problem p_k^T addressed by s_i^T?
⊙**3**⊙ *Where* is problem p_k^T addressed, is it in t_i^T, in v_i^T, or in both?
⊙**4**⊙ *What* FoCs $\{f_j^T\}$ are evoked by s_i^T;
⊙**5**⊙ *Why* is each FoC f_j^T evoked by s_i^T?

Details of the implementation of the Rationale Structure Prompting are provided in Appendix D.

However, as reported in Weinzierl and Harabagiu (2024a) CoT reasoning used for the articulation of FoCs and the discovery of the problems they address functions best when it operates in a few-shot learning mode. Consequently, we need access to some demonstrations showing how some SMP s_x is addressing a problem p_y^x which is interpreted by an FoC f_z^x that evoked in s_x. Moreover, the demonstrations also need to contain rationales for p_y^x and f_z^x.

Instead of providing expert-created demonstrations, we make use of a special form of RAG which considers the demonstrations encoded in DID. RAG retrieves from the DID a ranked list of demonstrations $D(s_i^T) = \{D_i^1, D_i^2, ...\}$ for each SMP s_i^T. To do so, it uses a query Q_i^T created by concatenating the CLIP-generated embedding of t_i^T, the text contained in s_i^T, with the CLIP-generated embedding of v_i^T, the image used in s_i^T. $D(s_i^T)$ contains demonstrations listed in descending order of their relevance to s_i^T, where the relevance $r(D_i^j) = Q_i^T \cdot e_j$, with e_j as the embedding of a SMP s_j encoded in the DID. Each D_i^j represents the structured explanation SE_j of s_j. A small number K_D of the top demonstrations from $D(s_i^T)$ are used in CoT reasoning, to enhance the Rationale Structure Prompting. A detailed example of retrieval from the DID is provided in Appendix D. For each SMP s_i^T from the Test Dataset, $\{f_j^T\}$, the set of FoCs evoked by s_i^T are discovered and the problem addressed by each f_j^T is identified. The rationales of the FoCs and of the problems are also generated.

3.3 Phase C

The third phase of DA-FoC^{MM} concerns the identification of FoCs which paraphrase each other. Such paraphrases are explained by the fact that in Phase B, each multimodal SMP was processed independently of the other SMPs from the Test Dataset. Therefore, different articulations of the same FoC may be generated as paraphrases.

Paraphrase detection between pairs of FoCs from the set of FoCs discovered in Phase B of the DA-FoC^{MM} method, S_{FoC}^B, is cast as a sequential decision process that constructs a final, unique set of FoCs that contain no paraphrases, S_{FoC}^C. Initially $S_{FoC}^C = \{f_1\}$, where f_1 is an FoC selected from S_{FoC}^B. To decide which additional f_i from S_{FoC}^B

should be added to S^C_{FoC}, the Paraphrase Structure Prompting of the LMM is performed to determine if f_i paraphrases any of the FoCs already existing in S^C_{FoC}. CoT reasoning, which operates in zero-shot mode, also provides a rationale of the possible paraphrase. The prompt, further detailed in Appendix E with examples, is:

△**1**△ *What* FoC $f_j \in F^C_{FoC}$ does f_i paraphrase;

△**2**△ *What* problems $\{p_k\}$ do both FoCs f_j and f_i address;

△**3**△ *Why* do both f_i and f_j address p_k in the same way?

△**4**△ *Why* does f_i paraphrase f_j?

Only if f_i does not paraphrase any FoC from S^C_{FoC}, will it be inserted into S^C_{FoC}. After all FoCs from S^C_{FoC} are considered, we obtain the final set of FoCs evoked in the Test Dataset, which is available from S^C_{FoC}. Table 1 shows the reduction of the number of FoCs from S^B_{FoC} to those in S^C_{FoC} for the topic of hesitancy towards COVID-19 vaccination. Appendix F provides the same information for the second topic, namely immigration.

Method	System	K_D	S^B_{FoC}	S^C_{FoC}
PriorWork$_X$	LLaMa-2	50+	2,142	340
PriorWork$_X$	GPT-3.5	30+	2,238	386
PriorWork$_X$	GPT-4	15	2,374	292
DA-FoC$^{MM}_X$	GPT-4o-Mini	0	1,521	-
DA-FoC$^{MM}_X$	GPT-4o	0	1,390	-
DA-FoC$^{MM}_X$	GPT-4o	1	1,435	220
DA-FoC$^{MM}_X$	GPT-4o	5	1,404	181
DA-FoC$^{MM}_X$	GPT-4o-Mini	10	1,628	198
DA-FoC$^{MM}_X$	GPT-4o	10	1,407	153
DA-FoC$^{MM}_I$	GPT-4o	10	1,398	150

Table 1: Number of COVID-19 vaccine FoCs discovered in Phase B and the final number of FoCs resulting from Phase C when considering (1) the system reported in Weinzierl and Harabagiu (2024a), operating only on textual SMPs from Twitter / X of dataset TS1, denoted as PriorWork$_X$; (2) DA-FoC^{MM} operating on dataset TS1, denoted as DA-FoC$^{MM}_X$; and (3) DA-FoC^{MM} operating on dataset TS2, denoted as DA-FoC$^{MM}_I$. K_D represents the number of demonstrations used for CoT prompting.

4 Evaluation Results

Quantitative Results: DA-FoC^{MM} relies upon the constrained decoding capability of OpenAI's recent LMMs and their structured output functionality to produce detailed indicative structured explanations and structured CoT rationales. Therefore, the only two LMM systems we considered for our DA-FoC^{MM} framework were GPT-4o and GPT-4o-Mini. GPT-4o is the current flagship LMM by OpenAI, building upon the GPT-4 (OpenAI, 2023) and GPT-4V (OpenAI, 2024) architectures. GPT-4o has recently demonstrated high-quality multimodal content analysis capabilities (Wu et al., 2024; Shahriar et al., 2024), as well as benefitting from complex prompting paradigms (Yue et al., 2024). Additionally, GPT-4o-Mini was recently released to replace GPT-3.5 (Ouyang et al., 2022), bringing multimodal capabilities to a much smaller, cheaper LMM, while also performing well on visual understanding tasks (Yue et al., 2024). We also compare against the prior text-only system introduced by Weinzierl and Harabagiu (2024a) on CoVaxFrames for reference, which employs LLaMa-2 (Touvron et al., 2023), GPT-3.5, and GPT-4. Additionally, we utilize the ViT-bigG/14 CLIP model, trained with the LAION-2B English subset of LAION-5B (Schuhmann et al., 2022), for Phase B of DA-FoC^{MM}, as initial experiments demonstrated that this CLIP model worked best for retrieving demonstrations from the DID.

We also experimented with zero-shot DA-FoC^{MM}, where zero demonstrations are shown during Phase B, as well as 1, 5, and 10 demonstrations retrieved. Table 1 lists the number of discovered FoCs discussing COVID-19 vaccine hesitancy and final FoCs resulting from each approach across the prior text-only system on Twitter / X, multimodal DA-FoC$^{MM}_X$ on Twitter / X, as well as multimodal DA-FoC$^{MM}_I$ with Twitter / X demonstrations on the Instagram test dataset, introduced in Section 2. Similar results are presented in Appendix F for the topic of immigration. As Table 1 illustrates, zero-shot learning when prompting GPT-4o-Mini and GPT-4o failed to produce any meaningful FoCs, and therefore these approaches were not evaluated in the qualitative results.

Qualitative Results: Weinzierl and Harabagiu (2024a) introduced common measures for the discovery and articulation of FoCs: (a) the *soundness* of the rationales generated by LMMs when articulating an FoC; (b) the *clarity* of the final FoC articulation generated by the LMM; and (c) the *novelty* of the final set of FoCs when compared to the known FoCs in the reference dataset, introduced in Section 2. Two expert linguists were asked to judge the soundness, clarity, and novelty of the final FoCs produced by each approach, and agreement was measured on a sample of 1000 judgments. The agreement was measured with a Co-

Method & Dataset	System	Num. Demos	Z	A	R	R_K	F_1	P_A
PriorWork$_X$	LLaMa-2	50+	35.29	68.86	42.06	47.32	52.22	42.11
PriorWork$_X$	GPT-3.5	30+	39.38	53.37	89.57	78.76	66.88	39.39
PriorWork$_X$	GPT-4	15	97.60	95.89	94.92	86.73	95.40	93.81
DA-FoC$_X^{MM}$	GPT-4o	1	58.18	66.36	92.41	89.38	77.25	37.82
DA-FoC$_X^{MM}$	GPT-4o	5	79.01	86.19	**95.71**	**93.81**	90.70	66.67
DA-FoC$_X^{MM}$	GPT-4o-Mini	10	94.95	96.97	92.31	85.84	94.58	94.06
DA-FoC$_X^{MM}$	GPT-4o	10	**99.35**	**99.35**	93.83	91.15	**96.51**	**98.00**
DA-FoC$_I^{MM}$	GPT-4o	10	97.33	98.00	91.30	87.61	94.53	94.12

Table 2: Evaluation results of the final set of COVID-19 vaccine FoCs with (1) the system introduced by Weinzierl and Harabagiu (2024a) that operates only on textual SMPs from Twitter / X of dataset TS1, denoted as PriorWork$_X$; (2) DA-FoC^{MM} operating on the dataset TS1, denoted as DA-FoC$_X^{MM}$; and (3) DA-FoC^{MM} operating on dataset TS2, denoted as DA-FoC$_I^{MM}$.

hen's Kappa score of 0.74, indicating strong agreement (McHugh, 2012).

Soundness, clarity, and novelty judgments were then transformed into 6 established DA-FoC evaluation metrics, introduced by Weinzierl and Harabagiu (2024a): (1) the quality of *reasoning* (Z); (2) the quality of *articulation* (A); (3) the *recall* of clearly articulated FoCs (R); (4) the recall of *known* FoCs (R_K); (5) a combined measure of *articulation quality* and *recall* ($F_1 = 2AR/(A+R)$); and (6) the articulation quality of *novel* FoCs (P_A). Table 2 lists the evaluation metrics for each approach for DA-FoC^{MM} on Twitter / X and Instagram on the topic of COVID-19 vaccines, as well as a comparison against the text-only DA-FoC system on Twitter / X from Weinzierl and Harabagiu (2024a). Similar evaluation results obtained on the datasets TS3 and TS4, covering the topic of immigration, are presented in Appendix F.

5 Discussion

The DA-FoC$_X^{MM}$ method, when prompting GPT-4o, achieves remarkable performance on Twitter / X, generating the best results across both the topics of COVID-19 vaccine hesitancy and immigration. It also performs very well across all performance metrics when it uses $M = 10$ demonstrations retrieved from the DID, as presented in Table 2 and Appendix F. Our approach compares extremely favorably to the text-only approach on the topic of COVID-19 vaccines, scoring higher in almost every metric. Moreover, DA-FoC$_X^{MM}$ when prompting GPT-4o still achieves a known recall R_K of 89.38 on COVID-19 vaccines (87.27 on immigration) even considering only a single demonstration from the DID. However, as the number of demonstrations grows, DA-FoC$_X^{MM}$ when prompting GPT-4o produces increasingly sound rationales, as revealed by the results of the Z metric.

Articulation clarity, as measured by the A metric, also rises sharply along with the number of demonstrations, illustrating the value of indicative structured prompting. The clarity of newly discovered FoCs, as measured by the P_A metric, also achieved 98.00 on COVID-19 vaccines (74.95 on immigration) when 10 demonstrations were utilized, which marks a significant increase over text-only systems on COVID-19 vaccine hesitancy FoCs. Additionally, even GPT-4o-Mini compares favorably on DA-FoC$_X^{MM}$, while GPT-4o-Mini is a significantly smaller and cheaper LMM. These results inform us that the discovery and articulation of FoCs from multimodal SMPs can be accomplished by LMMs with DA-FoC^{MM}, with high measures of soundness, clarity, and novelty. Furthermore, DA-FoC$_I^{MM}$ when prompting GPT-4o achieved extremely high soundness, clarity, recall, and novelty, as shown in Table 2 and Appendix F, on SMPs from TS2, with only 10 demonstrations, retrieved from the DID. Further discussions of the results obtained when considering the datasets TS3 and TS4, covering the topic of immigration, are provided in Appendix F. Additionally, a comprehensive error analysis is presented in Appendix G, where we show that DA-FoC$_X^{MM}$ with GPT-4o and 10 demonstrations performs excellently - by producing sound rationales and clearly articulated FoCs - when compared to other system configurations.

Cross-Platform Insights: Insights into framing choices across social media platforms were revealed when we analyzed *where* vaccine hesitancy problems were addressed (text, image, or both) and utilized to evoke FoCs when SMPs discussed controversial problems surrounding COVID-19 vac-

cines. Only 24.1% of SMPs utilized *only* their text to evoke FoCs, with 23.6% on Twitter / X vs. 24.5% on Instagram. Furthermore, only 1.4% of SMPs employed *only* their image to evoke FoCs, with 2.3% on Twitter / X vs. 0.6% on Instagram.

These results indicate that approximately 39% of COVID-19 vaccine hesitancy FoCs would not be recognized if the DA-FoC^{MM} method had not considered both the texts and the images of SMPs found on either Twitter / X or Instagram. On Twitter / X we found that 37% (56 out of 153) of the final FoCs would not have been discovered without considering the images from SMPs. Similarly, on Instagram, 41% (62 out of 150) of FoCs would not have been discovered without considering the images of SMPs. Moreover, each FoC is evoked by many SMPs. Therefore there are *evocation relations* between each SMP and the FoC it evokes. When FoCs are missed, because the images of SMPs are ignored, we found that 76% of evocation relations are also missed. This means that 76% of the time, we would not discover that an SMP evokes an FoC. This clearly demonstrates the importance of considering not only text, but also images when analyzing framing on social media, as previously shown to work for framing analysis on television (Entman, 2003). A breakdown of the multimodal necessity of framing concerning COVID-19 vaccines is presented in Appendix H.

6 Related Work

Significant Social Science research has manually investigated the role of framing in news (Gamson, 1989; Entman, 1989, 1991; Pan and Kosicki, 1993; Entman and Rojecki, 1993; Miller, 1997; D'Angelo, 2002; Entman, 2004; Scheufele, 2006; Entman, 2007; Reese, 2007; Matthes and Kohring, 2008). The multimodal nature of framing was detailed in Entman (2003), which investigated the repeated use of culturally resonant terms in concert with searing images of the burning and collapsing World Trade Center during the aftermath of 9/11.

Early automatic frame identification methods on social media focused on detecting addressed problems (Meraz and Papacharissi, 2013; Neuman et al., 2014; de Saint Laurent et al., 2020; Baumer et al., 2015; Tsur et al., 2015; Field et al., 2018a), such as supervised NLP methods (Card et al., 2016; Naderi and Hirst, 2017; Field et al., 2018b; Khanehzar et al., 2019; Kwak et al., 2020; Roy and Goldwasser, 2020) that utilized the Media Frames Corpus (MFC) (Card et al., 2015). The MFC includes news articles annotated with fifteen policy frame *problems*, such as Constitutionality and Jurisprudence or Security and Defense. Mendelsohn et al. (2021) identified immigration policy problems in SMPs with multi-label classification methods, relying on RoBERTa (Liu et al., 2019), and then manually articulated FoCs. However, without automatic methods capable of reasoning about the *cause* and articulating the FoC, we cannot capture the causal nature of framing.

Recently, research interest has shifted towards automatic methods for frame discovery and articulation. Weinzierl and Harabagiu (2024a) built an In-Context Active Curriculum Learning (ICACL) methodology that relied upon CoT prompting with LLMs, such as GPT-4. However, their approach could only discover and articulate FoCs on text-only SMPs from Twitter / X, and therefore would not function on platforms such as Instagram, where images predominate. Additionally, their methodology required a human-in-the-loop to create and edit CoT rationales and demonstrations, which required significant human effort.

7 Conclusion

This paper introduces the first method capable of discovering and articulating Frames of Communication (FoCs) from multimodal social media, namely DA-FoC^{MM}. This method uses several new structured prompting methods operating on Large Multimodal Models (LMMs). DA-FoC^{MM} is the first method to also show capabilities for discovery and articulation of FoCs from multimodal SMPs across social media platforms. Thorough evaluations demonstrate that DA-FoC^{MM}, when prompting GPT-4o, re-discovered 91% of FoCs found by communication experts on the same Twitter / X dataset discussing COVID-19 vaccines (90% for immigration), while also uncovering new FoCs that were clearly articulated and had sound rationales. The detailed rationales produced by DA-FoC^{MM} when prompting GPT-4o also enabled us to make sense of the different framing choices made across social media platforms, providing insights into *where* and *why* controversial problems are discussed on social media. Importantly, the evaluation results revealed that 39% of FoCs would not have been recognized if DA-FoC^{MM} would have ignored the images of social media postings.

8 Ethical Statement

We protected the privacy and honored the confidentiality of the authors who posted the SMPs in all the datasets considered. We received approval from the Institutional Review Board at ANONYMIZED for working with these Twitter / X and Instagram social media datasets. IRB-XX-YYY stipulated that our research met the criteria for exemption #8(iii) of the Chapter 45 of Federal Regulations Part 46.101.(b). The experiments were conducted with rigorous professional standards, ensuring that evaluation on the test dataset was deferred until a final method was chosen. All experimental settings, configurations, and procedures are thoroughly documented in this paper, the supplementary materials in the appendix, and the associated GitHub repository. We do not anticipate any significant risks associated with our research, as it is aimed at enhancing the understanding of how COVID-19 vaccine hesitancy and immigration is framed on social media. The overarching priority throughout this research was the public good, with the dual aim of advancing natural language processing and public health research.

9 Limitations

The DA-FoC^{MM} method introduced to discover and articulate Frames of Communication (FoCs) from multimodal Social Media Postings (SMPs) focuses on SMPs from Twitter / X and Instagram. Our method will likely require modification to work as well on SMPs from longer-form social media platforms, such as Reddit. Furthermore, our method operates on only text and images in SMPs. Social media platforms, such as TikTok, employ video and audio, which will require additional approaches. Future work will address these additional social media platforms and modalities by extending our DA-FoC^{MM} method by considering additional modalities and content lengths.

Our approach is also limited by the need to have available a reference dataset of multimodal SMPs with evoked FoCs and addressed problems. First, these reference FoCs must be discovered with inductive frame analysis (Van Gorp, 2010) on thousands of SMPs, with additional efforts required to identify all the SMPs that evoke these reference FoCs. This involves non-trivial, significant effort from communication experts. We plan to extend our method to require significantly fewer demonstrations to mitigate these limitations.

References

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482, Denver, Colorado. Association for Computational Linguistics.

Rodney Benson. 2013. *Shaping Immigration News: A French-American Comparison*. Communication, Society and Politics. Cambridge University Press.

Toby Bolsen, James N. Druckman, and Fay Lomax Cook. 2014. How Frames Can Undermine Support for Scientific Adaptations: Politicization and the Status-Quo Bias. *Public Opinion Quarterly*, 78(1):1–26.

Amber E. Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. 2018. Tracking the development of media frames within and across policy issues.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Annual Meeting of the Association for Computational Linguistics*.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas. Association for Computational Linguistics.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024a. Measuring and improving chain-of-thought reasoning in vision-language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 192–210, Mexico City, Mexico. Association for Computational Linguistics.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. 2024b. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *AAAI Conference on Artificial Intelligence*.

Dennis Chong and James N. Druckman. 2012. Counterframing effects. *The Journal of Politics*, 75(1):1–16.

Paul D'Angelo. 2002. News Framing as a Multiparadigmatic Research Program: a Response to Entman. *Journal of Communication*, 52(4):870–888.

Constance de Saint Laurent, Vlad Petre Glăveanu, and Claude Chaudet. 2020. Malevolent creativity and social media: Creating anti-immigration communities on twitter. *Creativity Research Journal*, 32:66–80.

Junnan Dong, Qinggang Zhang, Huachi Zhou, Daochen Zha, Pai Zheng, and Xiao Huang. 2024. Modality-aware integration with large language models for knowledge-based visual question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2417–2429, Bangkok, Thailand. Association for Computational Linguistics.

Robert M. Entman. 1989. *Democracy without citizens: media and the decay of American politics*. Oxford University Press, New York, New York ;.

Robert M Entman. 1991. Framing u.s. coverage of international news: Contrasts in narratives of the kal and iran air incidents. *Journal of communication*, 41(4):6–27.

Robert M. Entman. 2003. Cascading activation: Contesting the white house's frame after 9/11. *Political Communication*, 20(4):415–432.

Robert M. Entman. 2004. *Projections of Power: Framing News, Public Opinion, and U.S. Foreign Policy*. University of Chicago Press.

Robert M. Entman. 2007. Framing Bias: Media in the Distribution of Power. *Journal of Communication*, 57(1):163–173.

Robert M. Entman and Andrew Rojecki. 1993. Freezing out the public: Elite and media framing of the u.s. anti-nuclear movement. *Political Communication*, 10(2):155–173.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018a. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018b. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

William A. Gamson. 1989. News as framing: Comments on graber. *The American Behavioral Scientist*, 33(2):157–161.

Mattis Geiger, Franziska Rees, Lau Lilleholt, Ana P. Santana, Ingo Zettler, Oliver Wilhelm, Cornelia Betsch, and Robert Böhm. 2022. Measuring the 7cs of vaccination readiness. *European Journal of Psychological Assessment*, 38(4):261–269.

Xuehai He, Diji Yang, Weixi Feng, Tsu-Jui Fu, Arjun Akula, Varun Jampani, Pradyumna Narayana, Sugato Basu, William Yang Wang, and Xin Wang. 2022. CPL: Counterfactual prompt learning for vision and language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3407–3418, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jan Fredrik Hovden and Hilmar Mjelde. 2019. Increasingly controversial, cultural, and political: The immigration debate in scandinavian newspapers 1970–2016. *Javnost - The Public*, 26(2):138–157.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Gideon Keren. 2011. *Perspectives on framing*. Psychology Press.

Shima Khanehzar, Andrew Turpin, and Gosia Mikolajczak. 2019. Modeling political framing across policy issues and contexts. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 61–66, Sydney, Australia. Australasian Language Technology Association.

Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. In *Proceedings of the 12th ACM Conference on Web Science*, WebSci '20, page 305–314, New York, NY, USA. Association for Computing Machinery.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yinhan Liu, 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.

Jörg Matthes and Matthias Kohring. 2008. The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, 58(2):258–279.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Sharon Meraz and Zizi Papacharissi. 2013. Networked gatekeeping and networked framing on #egypt. *The International Journal of Press/Politics*, 18:138–166.

M. Mark Miller. 1997. Frame mapping and analysis of news coverage of contentious issues. *Social Science Computer Review*, 15(4):367–378.

Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 536–542, Varna, Bulgaria. INCOMA Ltd.

W. Russell Neuman, Lauren Guggenheim, S. Mo Jang, and So Young Bae. 2014. The dynamics of public attention: Agenda-setting theory meets big data. *Journal of Communication*, 64:193–214.

OpenAI. 2023. Gpt-4 technical report.

OpenAI. 2024. GPT-4V(ision) system card.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Zhongdang Pan and Gerald M. Kosicki. 1993. Framing analysis: An approach to news discourse. *Political communication*, 10(1):55–75.

Thomas E. Patterson. 1992. Is anyone responsible? how television frames political issues. *American Political Science Review*, 86(4):1060–1061.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Stephen D. Reese. 2007. The framing project: A bridging model for media research revisited. *Journal of Communication*, 57(1):148–154.

Stephen D. Reese, Oscar H. Gandy, and August E. (Eds.) Grant. 2001. *Framing Public Life: Perspectives on Media and Our Understanding of the Social World*. Routledge.

Deana A. Rohlinger. 2002. Framing the abortion debate: Organizational resources, media strategies, and movement-countermovement dynamics. *The Sociological Quarterly*, 43(4):479–507.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7698–7716, Online. Association for Computational Linguistics.

Bertram Scheufele. 2004. Framing-effects approach: A theoretical and methodological critique. *Communications*, 29(4):401–428.

Dietram A. Scheufele. 2006. Framing as a Theory of Media Effects. *Journal of Communication*, 49(1):103–122.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sakib Shahriar, Brady Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. 2024. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency.

John Sonnett. 2019. 226Priming and Framing: dimensions of communication and cognition. In *The Oxford Handbook of Cognitive Sociology*. Oxford University Press.

Sanjay Subramanian, Medhini Narasimhan, Kushal Khangaonkar, Kevin Yang, Arsha Nagrani, Cordelia Schmid, Andy Zeng, Trevor Darrell, and Dan Klein. 2023. Modular visual question answering via code generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 747–761, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China. Association for Computational Linguistics.

Baldwin Van Gorp. 2010. *Strategies to take subjectivity out of framing analysis*, pages 84–109. Routledge.

Jason Wei, 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. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Maxwell Weinzierl and Sanda Harabagiu. 2023. Identification of multimodal stance towards frames of communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12597–12609, Singapore. Association for Computational Linguistics.

Maxwell Weinzierl and Sanda Harabagiu. 2024a. Discovering and articulating frames of communication from social media using chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1617–1631, St. Julian's, Malta. Association for Computational Linguistics.

Maxwell Weinzierl and Sanda Harabagiu. 2024b. Tree-of-counterfactual prompting for zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–880, Bangkok, Thailand. Association for Computational Linguistics.

Maxwell A. Weinzierl and Sanda M. Harabagiu. 2022. From hesitancy framings to vaccine hesitancy profiles: A journey of stance, ontological commitments and moral foundations. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1087–1097.

Yiqi Wu, Xiaodan Hu, Ziming Fu, Siling Zhou, and Jiangong Li. 2024. Gpt-4o: Visual perception performance of multimodal large language models in piglet activity understanding.

Alexandros Xenos, Themos Stafylakis, Ioannis Patras, and Georgios Tzimiropoulos. 2023. A simple baseline for knowledge-based visual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14871–14877, Singapore. Association for Computational Linguistics.

Songlin Yang, Roger Levy, and Yoon Kim. 2023. Unsupervised discontinuous constituency parsing with mildly context-sensitive grammars. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5747–5766, Toronto, Canada. Association for Computational Linguistics.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark.

Jing Zheng, Jyh-Herng Chow, Zhongnan Shen, and Peng Xu. 2023. Grammar-based decoding for improved compositional generalization in semantic parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1399–1418, Toronto, Canada. Association for Computational Linguistics.

A Dataset Details

Our primary research question involved discovering how images impacted framing across social media platforms. However, we also wanted to ensure these findings held across multiple topics. Therefore, we utilized four distinct datasets. These datasets spanned across two topics - COVID-19 vaccines and immigration - and included SMPs from two social media platforms - Twitter / X and Instagram, as introduced in Section 2. Each dataset is further detailed below.

Figure 4: Examples of multimodal SMPs, evoked FoCs, and interpreted problems from Twitter / X in dataset TS1.

A.1 Datasets covering the Topic: COVID-19 Vaccines

☐ The dataset RF1, opiginating from **Twitter / X:** In addition to annotations of the evoked FoCs, produced by communication experts on the MMVAX-STANCE dataset, the problems addressed by each FoC are available. These problems are informed by the 7C model of vaccine hesitancy (Geiger et al., 2022). The 7C model consists of seven factors, considered as hesitancy problems, that impact an individual's likelihood of getting vaccinated. Table 3 lists the problems and their definitions. The Table also indicates the number and percentage of annotated FoCs that address each problem in the MMVAX-STANCE dataset.

☐ The dataset TS1 contains SMPs from **Twitter / X**, available also from the the MMVAX-STANCE dataset, Figure 4 (A) illustrates an SMP from dataset TS1 that employs multimodal sarcasm to evoke an FoC. This SMP appears to thank Minnesota for enabling the author to receive the first dosage of the "new" COVID-19 vaccine, and that the author "looks and feels wonderful". However, the included image stands in stark contrast to the text of this SMP, with the image illustrating a disfigured character named "Sloth" from "The Goonies." The superimposed text transforms this image into a "meme", with the top text reading "Got my COVID-19 vaccine" and the bottom text reading "Feeling great!!!". The SMP in Figure 4 (A) therefore evokes the FoC "The COVID-19 vaccine alters human DNA", and this FoC interprets the vaccine hesitancy problems of Confidence and Conspiracy. Additional examples of the SMPs from dataset TS1 are provided in Figure 4 along with evoked FoCs and interpreted problems.

☐ The dataset TS2 contains SMPs from **Instagram:** To search for Instagram SMPs discussing the COVID-19 vaccines we used the same query as in Weinzierl and Harabagiu (2023), namely: "(covid OR coronavirus) AND vaccine AND lang:en". The retrieved Instagram SMPs were created between January 1st, 2020, and January 1st, 2022. Each SMP was comprised of text and an image. This search produced 516,581 Instagram SMPs, retrieved from the CrowdTangle platform, from which we considered a subset for our cross-platform experiments.

We selected a representative subset of the 516,581 Instagram SMPs by utilizing the text-based FoC evocation detection system described in Weinzierl and Harabagiu (2023). Our goal was to find a smaller set of SMPs that had a higher likelihood than random sampling of evoking any of the 113 FoCs from MMVAX-STANCE. This selection process improved our ability to measure the impact of images on the evocation of FoCs across both platforms, providing a similar set of SMPs evoking similar FoCs. Our filtering process produced a list of 1,289 SMPs, referred to as dataset TS2, likely to evoke at least one FoC from the 113 reference FoCs from MMVAX-STANCE.

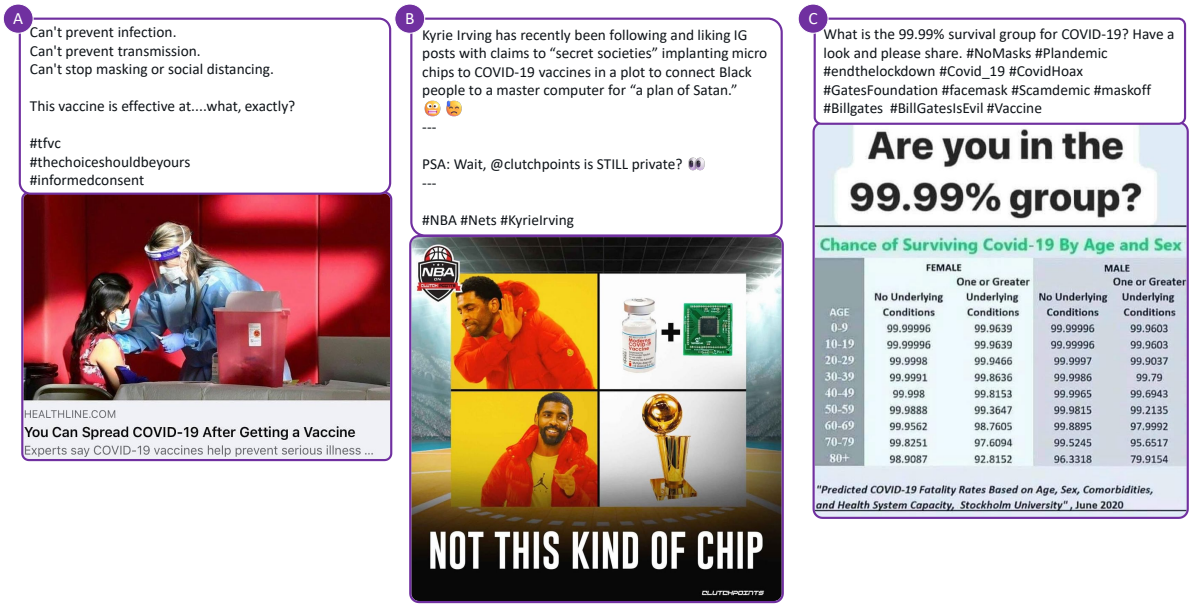

Figure 5: Examples of multimodal SMPs from our collection of Instagram SMPs discussing the COVID-19 vaccines in dataset TS2.

Figure 5 (B) illustrates an SMP in dataset TS2 from Instagram that discusses COVID-19 vaccines. The text of the SMP describes how Kyrie Irving, a professional basketball player in the NBA, has promoted Instagram posts that propagate COVID-19 vaccine conspiracy theories, such as those that state that the COVID-19 vaccine includes microchips in a satanic plan. The image included in this SMP further reinforces this message, using a common meme format, popularized with Drake (a popular Canadian rapper and singer) shrugging off something and then pointing with approval at something else. In this instance, Drake's face has been replaced with Kyrie, and Kyrie is shrugging off the Moderna COVID-19 vaccine and a microchip. This meme is therefore implying that Kyrie believes in the conspiracy theory that the COVID-19 vaccine includes microchips. In the next part of the meme, Kyrie shows approval and preference towards an NBA championship trophy, which is commonly referred to as a "chip" among players and fans. Together, this multimodal SMP employs significant cultural knowledge and certainly evokes the COVID-19 FoC that "the COVID-19 Vaccine is a satanic plan to microchip people" which interprets the problems of Confidence and Conspiracy. Additional examples of Instagram SMPs from dataset TS2 are provided in Figure 5.

A.2 Datasets covering the Topic: Immigration

☐ The dataset RF2 originates from **Twitter / X**.The

Problem	*Definition*
Confidence - 43 FoCs (38%)	Trust in the security and effectiveness of vaccinations, the health authorities, and the health officials who recommend and develop vaccines.
Complacency - 7 FoCs (6%)	Complacency and laziness to get vaccinated due to low perceived risk of infections.
Constraints - 1 FoC (1%)	Structural or psychological hurdles that make vaccination difficult or costly.
Calculation - 19 FoCs (17%)	Degree to which personal costs and benefits of vaccination are weighted.
Collective Responsibility 10 FoCs (9%)	Willingness to protect others and to eliminate infectious diseases.
Compliance - 27 FoCs (24%)	Support for societal monitoring and sanctioning of people who are not vaccinated.
Conspiracy - 37 FoCs (33%)	Conspiracy thinking and belief in fake news related to vaccination.

Table 3: Problems associated with vaccine hesitancy.

salient problems surrounding the topic of immigration have been studied extensively (Patterson, 1992; Benson, 2013; Hovden and Mjelde, 2019; Mendelsohn et al., 2021). Table 4 lists the problems and their definitions. These problems are informed by the Policy Frames Codebook, which provides a general-purpose way to structure and describe frame problems in political communication content (Boydstun et al., 2018). However, little work has studied the ways in which these problems are interpreted and framed on social media. Therefore, we decided to construct a new dataset to explore how multimodality impacts immigration framing.

Problem	*Description*
Economic	Financial implications of an issue.
Capacity & Resources	The availability or lack of time, physical, human, or financial resources.
Morality & Ethics	Perspectives compelled by religion or secular sense of ethics or social responsibility.
Fairness & Equality	The (in)equality with which laws, punishments, rewards, and resources are distributed.
Legality, Constitutionality & Jurisdiction	Court cases and existing laws that regulate policies; constitutional interpretation; legal processes such as seeking asylum or obtaining citizenship; jurisdiction.
Crime & Punishment	The violation of policies in practice and the consequences of those violations.
Security & Defense	Any threat to a person, group, or nation and defenses taken to avoid that threat.
Health & Safety	Health and safety outcomes of a policy issue, discussions of health care.
Quality of Life	Effects on people's wealth, mobility, daily routines, community life, happiness, etc.
Cultural Identity	Social norms, trends, values, and customs; integration/assimilation efforts.
Public Sentiment	General social attitudes, protests, polling, interest groups, public passage of laws.
Political Factors & Implications	Focus on politicians, political parties, governing bodies, political campaigns and debates; discussions of elections and voting.
Policy Prescription & Evaluation	Discussions of existing or proposed policies and their effectiveness.
External Regulation & Reputation	Relations between nations or states/provinces; agreements between governments; perceptions of one nation/state by another.
Victim: Global Economy	Immigrants are victims of global poverty, underdevelopment, and inequality.
Victim: Humanitarian	Immigrants experience economic, social, and political suffering and hardships.
Victim: War	Focus on war and violent conflict as reasons for immigration.
Victim: Discrimination	Immigrants are victims of racism, xenophobia, and religion-based discrimination.
Hero: Cultural Diversity	Highlights positive aspects of differences that immigrants bring to society.
Hero: Integration	Immigrants successfully adapt and fit into their host society.
Hero: Worker	Immigrants contribute to economic prosperity and are an important source of labor.
Threat: Jobs	Immigrants take nonimmigrants' jobs or lower their wages.
Threat: Public Order	Immigrants threaten public safety by breaking the law or spreading disease.
Threat: Fiscal	Immigrants abuse social service programs and are a burden on resources.
Threat: National Cohesion	Immigrants' cultural differences are a threat to national unity and social harmony.
Episodic	Message provides concrete information about specific people, places, or events.
Thematic	Message is more abstract, placing stories in broader political and social contexts.

Table 4: Descriptions of salient problems interpreted by Frames of Communication in immigration discourse.

We used the same query as in Mendelsohn et al. (2021) to find multimodal Twitter / X SMPs discussing immigration: "(immigration OR immigrant(s) OR emigration OR emigrant(s) OR migration OR migrant(s) OR illegal alien(s) OR illegals OR undocumented) AND lang:en". The retrieved Twitter / X SMPs were posted between January 1st, 2020, and January 1st, 2022, and each SMP included an image and text. This search produced 264,237 multimodal Twitter / X SMPs, retrieved from the Twitter / X historical API.

We randomly selected 2,000 unique SMPs for annotation from the full set of 264,237 multimodal Twitter / X SMPs. Two linguistic experts from ANONYMOUS followed the same procedure as Weinzierl and Harabagiu (2022) to perform inductive frame analysis (Van Gorp, 2010) on these 2,000 SMPs. After removing irrelevant SMPs, a total of 57 newly discovered FoCs were identified as being evoked by 1,878 multimodal SMPs. Each FoC was also annotated as interpreting any of the 27 immigration-specific problems, outlined in Table 4.

□ The dataset TS3 contains multimodal SMPs originating on Twitter /X . Figure 6 (C) illustrates an SMP from Twitter / X that discusses immigration from dataset TS3. The text of the SMP discusses how successful vaccine policy has been by the Biden administration. However, the image attached demonstrates what the author is trying to communicate: that Republicans scapegoat immigrants when politically convenient to distract from successful Democrat policies. Together, this SMP evokes the FoC which states that "immigrants are often scapegoated in political disputes, distracting from core issues like economic policy or governance." This FoC interprets the problems of public sentiment, political factors & implications, and the thematic problem, as defined in Table 4. Additional examples of SMPs from dataset TS3 and evoked FoCs are illustrated in Figure 6.

□ The dataset TS4 contians multimodal SMPs originating from the **Instagram** platform. We searched CrowdTangle for Instagram SMPs discussing the topic of immigration. We found 259,281 Instagram

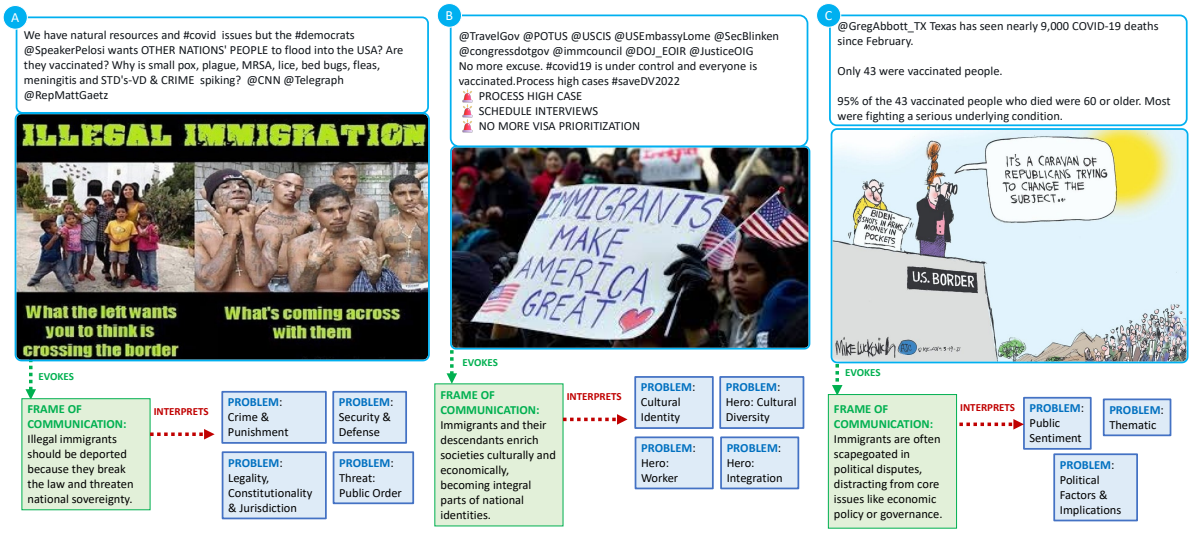

Figure 6: Examples of multimodal SMPs, evoked FoCs, and interpreted problems from Twitter / X discussing immigration in dataset TS3.

SMPs posted between January 1st, 2020, and January 1st, 2022, with each SMP containing an image and text. We also similarly selected a representative subset of 956 Instagram SMPs, utilizing the system from Weinzierl and Harabagiu (2023) to identify SMPs likely to evoke any of the same 57 immigration FoCs discovered on Twitter / X. These SMPs comprised the TS4 dataset.

Figure 7 (C) illustrates an SMP from the dataset TS4, originating from Instagram that discusses immigration. The text of the SMP outlines how Joe Biden wants to “rip our borders wide open and let thousands of illegals in." The text further raises the fear that these “illegals" will steal jobs - particularly the newly available $15 per hour minimum wage jobs - from Americans. Finally, the text touches on how Americans may end up paying for “illegals" to receive healthcare and COVID-19 vaccines. All of these fears are strengthened by an image from a riot in Venezuela involving anti-government protesters. Additional examples of SMPs from dataset TS4 discussing immigration on Instagram are provided in Figure 7.

B Constrained Decoding Prompts and Schema

Our constrained decoding approach is based on constrained decoding with Context-Free Grammars (CFGs) (Zheng et al., 2023; Yang et al., 2023), which drastically improves the reliability of generating structured outputs from generative models. Constrained decoding deterministically modifies the output probabilities of a next-token prediction model, such that all non-valid tokens are assigned probability zero, based on the defined CFG. This approach can be utilized to specify an exact output format, which can greatly assist in ensuring LLMs and LMMs follow a specific “thought" process when generating rationales and explanations.

For example, a CFG can be defined such that an LLM is required to first generate a step-by-step list of reasoning steps before a final answer, enforcing granular CoT generation. We employ three prompting templates and three constrained decoding schemes for Phases A, B, and C. These schemas ensure that the LMM adheres to precise syntactic and semantic constraints when producing outputs. By restricting the search space of possible next tokens, constrained decoding enhances both interpretability and consistency in generated outputs.

In our particular task, constrained decoding forces the LMM to reason separately about each modality explicitly, after which the LMM is presented an opportunity to reason jointly about both modalities. Additionally, structured outputs enable us to manipulate the generated indicative explanations from Phase A to make them appear as rationales for yet-to-be-articulated FoCs in Phase B.

C Indicative Explanations and Demonstration Creation

For Phase A, the prompting template is provided in Figure 8, while the constrained decoding schema is illustrated in Figure 16. This prompt template and schema ensure that the LMM generates the exact indicative explanation structure we outline in

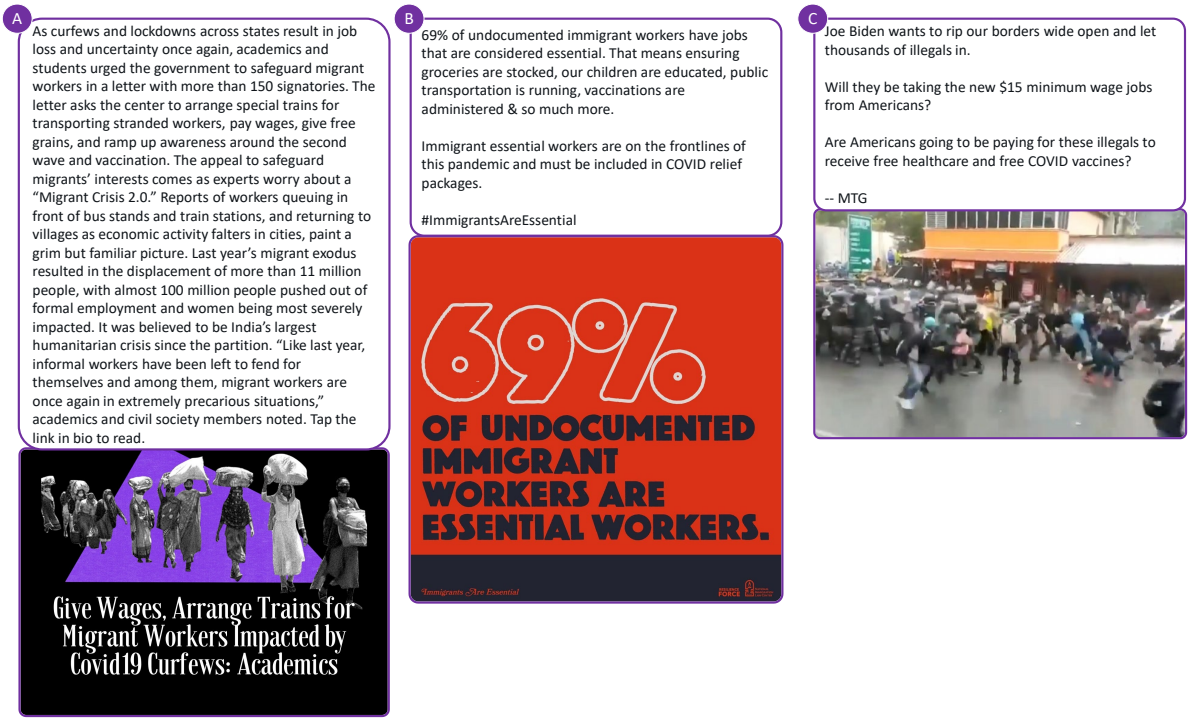

Figure 7: Examples of multimodal SMPs from our collection of Instagram SMPs discussing immigration in dataset TS4.

```yaml
system_prompt: >-
  You are an expert linguistic assistant.
  Frames of communication select particular aspects of an
  issue and make them salient in communicating a message.
  Salient aspects are referred to as problems, which are
  addressed through articulated causes when authors
  communicate via framing.
  Frames of communication are ubiquitous in social media
  discourse and can impact how people understand issues and,
  more importantly, how they form their opinions.
  Each frame of communication will be provided, along with
  problems addressed by the frame of communication.
  You will be tasked with explaining why a post evokes a
  particular frame of communication by articulating the
  causes provided for the addressed problems.
  Not necessarily every problem addressed by the frame of
  communication will be addressed in the post, so be sure to
  only consider problems supported by articulated causes.
  You should discuss your reasoning in detail, thinking
  step-by-step.
user_prompt: |-
  Frame of Communication:
  {frame_text}

  Problems Addressed by Frame of Communication:
  {frame_problems}

  Post:
  {post_text}

  Image:
  {post_image}
```

Figure 8: The prompt template utilized for Phase A, in YAML format.

```yaml
system_prompt: >-
  You are an expert linguistic assistant.
  Frames of communication select
  particular aspects of an issue and make
  them salient in communicating a message.
  Salient aspects are referred to as
  problems, which are addressed through
  articulated causes when authors
  communicate via framing.
  Frames of communication are ubiquitous
  in social media discourse and can impact
  how people understand issues and, more
  importantly, how they form their
  opinions.
  You will be tasked with identifying the
  problems a social media post addresses,
  as well as articulating evoked frames of
  communication by articulating the causes
  provided for the addressed problems.
  You should discuss your reasoning in
  detail, thinking step-by-step.
user_prompt: |-
  Post:
  {post_text}

  Image:
  {post_image}
```

Figure 9: The prompt template utilized for Phase B, in YAML format.

Section 3.1.

The prompt template in Figure 8 specifies detailed instructions to the system for producing structured explanations. The system prompt provides a comprehensive context, emphasizing the importance of Frames of Communication (FoCs) and their associated problems, while explicitly guiding the model to think step-by-step. The structured format guarantees consistency in responses, enabling precise mapping of inputs to FoCs and their respective addressed problems. The prompt aligns with this goal by specifying the input elements—text, image, and frame-related annotations—to ensure clarity in the generation process.

The JSON schema illustrated in Figure 16 for-

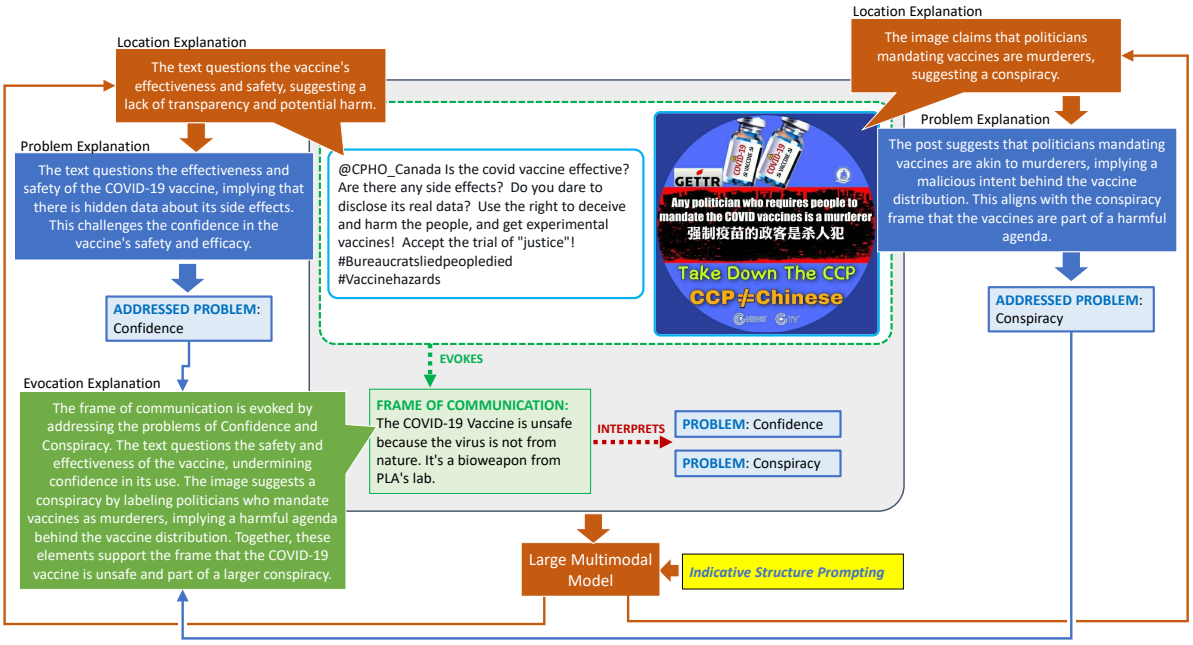

Figure 10: Example of the indicative explanations generated as part of Phase A.

```yaml
system_prompt: >-
  You are an expert linguistic assistant.
  Frames of communication select particular aspects of an
  issue and make them salient in communicating a message.
  Salient aspects are referred to as problems, which are
  addressed through articulated causes when authors
  communicate via framing.
  Frames of communication are ubiquitous in social media
  discourse and can impact how people understand issues
  and, more importantly, how they form their opinions.
  You will be tasked with judging if a new frame of
  communication is novel or a paraphrase of a known frame
  of communication.
  First, you should identify any overlap with addressed
  problems. Exact overlap is not necessary, but there
  likely should be some overlap in the addressed problems
  of paraphrasing frames of communication.
  The novel frames of communication will have the
  addressed problems listed, along with how often those
  problems were addressed by social media posts evoking
  the novel frame of communication.
  Next, you should ensure paraphrasing frames of
  communication share the same causes for the addressed
  problems they share.
  Finally, you should determine which known frame of
  communication, if any, is a paraphrase of the provided
  novel frame of communication. You will identify the
  paraphrasing frame of communication by providing the
  frame ID.
  You should discuss your reasoning in detail, thinking
  step-by-step.
user_prompt: |-
  Problem Definitions:
  {problem_definitions}

  Known Frames of Communication:
  {known_frames}

  Novel Frame of Communication:
  {novel_frame_text}

  Problems Addressed by Frame of Communication:
  {novel_frame_problems}
```

Figure 11: The prompt template utilized for Phase C, in YAML format.

malizes this process further by defining the permissible structure of the output. The schema ensures that each problem identified is linked to specific parts of the input (text or image) with clear explanations. It enforces strict adherence to the required components, including problem explanations and the overarching frame explanation, making the output highly interpretable and robust. By constraining the decoding process with this schema, we also minimize the risk of generating invalid or incomplete responses.

An example of indicative structure prompting is provided in Figure 10 on an SMP from RF1. Figure 10 demonstrates how an SMP from RF1 is processed to generate indicative explanations. The SMP's text raises questions about the vaccine's safety and effectiveness, suggesting hidden risks and a lack of transparency. The LMM identifies this as addressing the problem of Confidence, with a detailed explanation of how the text undermines trust in the vaccine's efficacy.

Simultaneously, the image in the SMP addresses a different problem: Conspiracy. The image portrays politicians mandating vaccines as murderers, implying malicious intent behind the vaccination campaign. This aligns with conspiracy theories suggesting that the COVID-19 vaccines are part of a harmful agenda. The LMM provides a location-specific explanation for how the image addresses the Conspiracy problem, ensuring that the visual and textual elements of the SMP are analyzed separately but cohesively.

The final explanation synthesizes these components to explain the FoC evoked by the SMP. In this case, the frame posits that "The COVID-19 Vaccine is unsafe because the virus is not from nature. It's a bioweapon from PLA's lab." The generated explanation highlights how the combination of text and image elements contributes to framing the vaccine

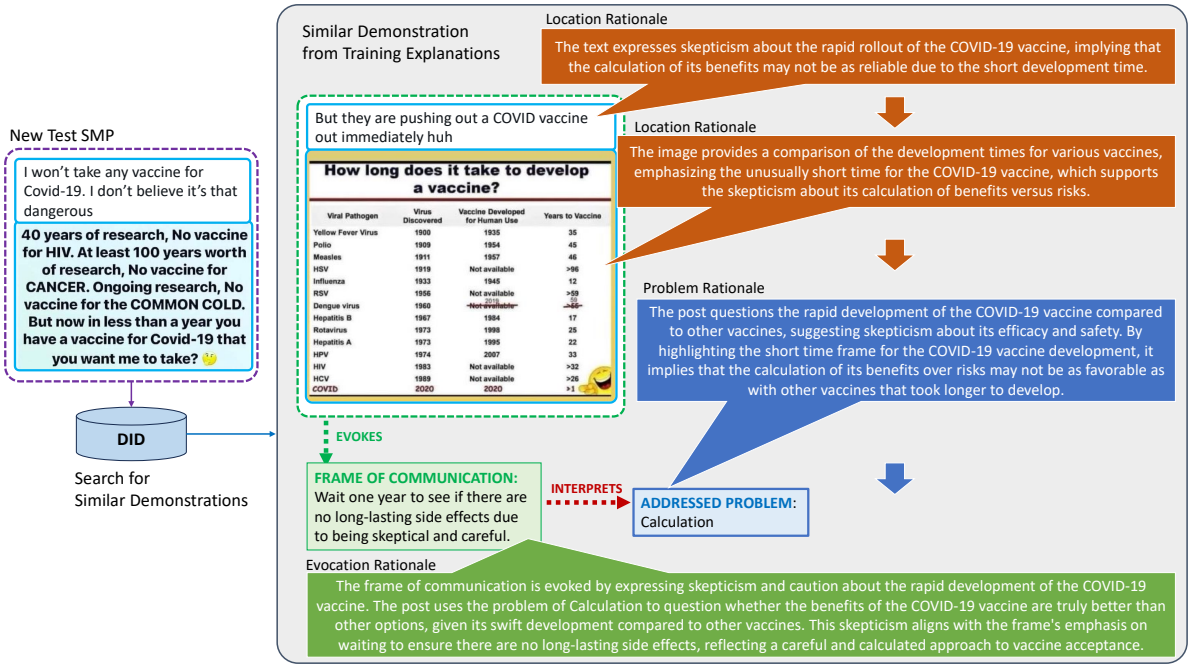

Figure 12: Example of demonstration retrieval in Phase B.

as part of a larger conspiracy.

As each explanation component from Figure 10 is generated in a structured format, we are able to easily re-arrange and manipulate these explanations to appear to the LMM in Phase B as demonstrations with rationales. This is the key insight into how our method is capable of producing CoT demonstrations entirely automatically - by exploiting post-hoc explanations of indicative examples we are able to transform these explanations into CoT demonstrations for Phase B to operate on dataset TS1 or dataset TS2.

D Demonstration Retrieval and Frame Discovery

For Phase B, the prompting template is provided in Figure 9, while the constrained decoding schema is illustrated in Figure 17. These together ensure that the LMM generates the structured rationales we seek in Phase B, introduced in Section 3.2. Figure 9 details the system and user prompts designed for Phase B. The system prompt guides the LMM to identify problems and articulate FoCs in an SMP. The JSON schema, illustrated in Figure 17, defines the expected structure of the model's output. Each addressed problem must be linked to specific locations in the post (either text or image), with clear rationales for how the problem is addressed. Furthermore, the schema enforces that the FoC evoked by the post is explicitly articulated, drawing upon the identified problems and their associated rationales. However, the key to Phase B is the retrieval of demonstrations of rationales produced by explanations generated in Phase A.

The retrieval process for demonstrations for an example SMP from dataset TS1 is illustrated in Figure 12. In this example, a new SMP questions the rapid rollout of the COVID-19 vaccine, expressing skepticism about its safety compared to vaccines developed over much longer periods. The retrieval mechanism identifies a similar demonstration from the training explanations, indexed in the DID, which also discusses vaccine development timelines. This retrieved explanation provides context and structure for the LMM's reasoning, and helps guide the LMM towards an accurate discovery and articulation from the new test SMP.

By integrating demonstration retrieval with rationale structure prompting, Phase B ensures that the LMM's outputs are forced to include rationales with all identified and articulated FoCs, while also being guided by the demonstrations retrieved from the DID. This approach not only improves the quality of generated rationales but, also facilitates deeper insights into how SMPs evoke FoCs and address salient problems.

E Paraphrase Detection Details

For Phase C, the prompting template is provided in Figure 11, while the constrained decoding schema

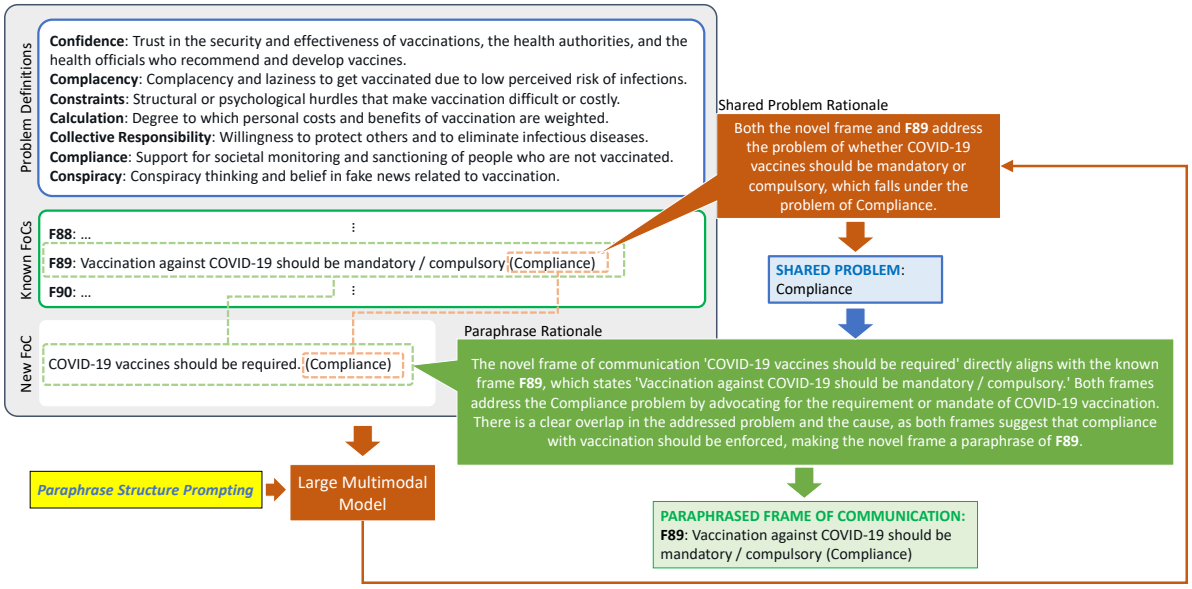

Figure 13: Example of the paraphrase identification as part of Phase C.

is illustrated in Figure 18. These together enable us to follow the sequential decision process, provided in Section 3.3, which identifies paraphrase relations and organizes a final set of FoCs.

Figure 13 illustrates an example of zero-shot paraphrase identification on FoCs discovered from dataset TS1. In this example, a novel FoC articulates that "COVID-19 vaccines should be required," while a known FoC, F89, states that "Vaccination against COVID-19 should be mandatory/compulsory." Both FoCs address the problem of Compliance, which is defined as support for societal monitoring and sanctioning of individuals who are not vaccinated.

The rationale for identifying this pair of FoCs as paraphrases is grounded in their shared problem, Compliance, and the overlapping causes articulated in both FoCs. The novel FoC emphasizes the necessity of COVID-19 vaccination, aligning closely with F89's advocacy for mandatory vaccination policies. This shared problem rationale highlights how both FoCs address Compliance in a similar manner, justifying their classification as paraphrases.

The paraphrase rationale further explains that the novel FoC does not introduce a new perspective or additional problems beyond those addressed by F89. Consequently, the LMM identifies the novel FoC as a paraphrase of F89, avoiding redundancy in the final set of FoCs. This decision process is guided by the paraphrase structure prompting schema, illustrated in Figure 18, which ensures consistency and transparency in paraphrase detection.

The JSON schema in Figure 18 formalizes the paraphrase identification process. It requires explicit identification of shared problems and their rationales, as well as a clear rationale for why one FoC paraphrases another. By enforcing these requirements, the schema supports rigorous analysis of paraphrase relations and ensures that the final set of FoCs is both concise and comprehensive.

This paraphrase detection approach addresses the challenges posed by independent processing of SMPs in Phase B, which may result in multiple articulations of the same FoC. By consolidating paraphrases, Phase C refines the set of discovered FoCs, reducing redundancy and improving the interpretability of the results.

We also evaluated the quality of the paraphrase relations between FoCs, discovered in Phase C of the DA-FoC^{MM} when prompting GPT-4o and using SMPs from Twitter / X. Two linguistic experts made judgments and found that 99.24% of paraphrase relations were correct. This result also improves upon the prior method of discovering and articulating FoCs only from textual SMPS, reported in Weinzierl and Harabagiu (2024a), which had achieved a 99.15% accuracy for paraphrase relations.

F Evaluation Results for the Topic of Immigration and Discussion

Table 5 shows the number of FoCs discovered in Phase B and the final FoCs produced in Phase C for the immigration test datasets TS3 and TS4. This includes results for DA-FoC$_X^{MM}$ operating on Twitter / X and DA-FoC$_I^{MM}$ operating on Instagram. Table 6 presents the evaluation metrics, including reasoning quality (Z), articulation quality (A), recall (R), recall of known FoCs (R_K), combined F_1 score, and clarity of novel FoCs (P_A).

The results indicate strong performance for DA-FoC^{MM} on the topic of immigration, with GPT-4o achieving the best outcomes across all evaluation metrics. For DA-FoC$_X^{MM}$, prompting GPT-4o with 10 demonstrations achieves the highest scores across all metrics. The reasoning quality (Z) reaches 90.48, while articulation quality (A) improves to 91.70. The recall of known FoCs (R_K) reaches 89.90, demonstrating GPT-4o's strong ability to rediscover manually identified and articulated FoCs. Additionally, the F_1 score improves to 92.31, illustrating the system's balance between articulation clarity and recall. Notably, DA-FoC$_X^{MM}$ produces novel FoCs with a high degree of clarity, achieving a P_A score of 78.07.

DA-FoC$_I^{MM}$, which operates on Instagram SMPs, also achieves impressive performance when GPT-4o is prompted with 10 demonstrations. The reasoning quality (Z) and articulation quality (A) are 89.55 and 90.16, respectively, showing only a minor reduction compared to Twitter / X results. However, the recall (R) and recall of known FoCs (R_K) are lower, at 75.19 and 67.11, respectively.

Method	System	K_D	S^B_{FoC}	S^C_{FoC}
DA-FoC$_X^{MM}$	GPT-4o-Mini	0	964	-
DA-FoC$_X^{MM}$	GPT-4o	0	803	-
DA-FoC$_X^{MM}$	GPT-4o	1	784	120
DA-FoC$_X^{MM}$	GPT-4o	5	724	93
DA-FoC$_X^{MM}$	GPT-4o-Mini	10	824	100
DA-FoC$_X^{MM}$	GPT-4o	10	758	82
DA-FoC$_I^{MM}$	GPT-4o	10	587	63

Table 5: Number of immigration FoCs discovered in Phase B and the final number of FoCs resulting from Phase C when considering (1) DA-FoC^{MM} operating on multimodal SMPs from Twitter / X in dataset TS3, denoted as DA-FoC$_X^{MM}$ and (2) DA-FoC^{MM} operating on multimodal SMPs from Instagram in dataset TS4, denoted as DA-FoC$_I^{MM}$. K_D represents the number of demonstrations used for CoT prompting.

This can be attributed to the distinct nature of Instagram content, which places greater emphasis on visual elements and often lacks the textual detail present in Twitter / X SMPs. Despite this, the combined F_1 score remains strong at 81.99, and the clarity of novel FoCs (P_A) achieves a competitive 74.95.

These results underscore the effectiveness of DA-FoC^{MM} in discovering and articulating FoCs across both Twitter / X and Instagram datasets. The higher performance on Twitter / X reflects the platform's text-centric nature, which aligns well with CoT prompting techniques. In contrast, Instagram's multimodal emphasis presents additional challenges but still yields strong outcomes, demonstrating the robustness of DA-FoC^{MM} in handling diverse modalities.

The results for immigration confirm that DA-FoC^{MM} can successfully identify, articulate, and refine FoCs across different platforms and topics. While GPT-4o with 10 demonstrations consistently produces the best performance, GPT-4o-Mini also achieves competitive results, highlighting the efficiency of the framework. These findings reinforce the value of indicative explanations with constrained decoding and CoT prompting in enabling high-quality multimodal frame discovery across varied datasets.

G Error Analysis

In this section, we compare the performance of three systems on an example multimodal SMP from dataset TS1 discussing COVID-19 vaccination, as illustrated in Figure 14. These systems include DA-FoC$_X^{MM}$ with GPT-4o-Mini and 10 demonstrations, DA-FoC$_X^{MM}$ with GPT-4o and 1 demonstration, and DA-FoC$_X^{MM}$ with GPT-4o and 10 demonstrations. This analysis highlights the strengths of the best-performing system, as discussed in Section 4, while exposing limitations and errors in the first two systems.

The multimodal SMP, illustrated in Figure 14, consists of a post that rejects COVID-19 vaccination mandates, using both textual and visual elements to frame the vaccine as coercive and untrustworthy. The image of a chaotic enforcement scenario, paired with sarcastic commentary in the text, evokes skepticism toward vaccines and distrust in government health initiatives. An undertone of conspiracy thinking is also present.

The first system, DA-FoC$_X^{MM}$ with GPT-4o-

Method & Dataset	System	Num. Demos	Z	A	R	R_K	F_1	P_A
DA-FoC$_X^{MM}$	GPT-4o	1	52.48	59.86	90.83	87.27	72.16	31.49
DA-FoC$_X^{MM}$	GPT-4o	5	70.87	77.31	90.10	86.07	83.21	52.20
DA-FoC$_X^{MM}$	GPT-4o-Mini	10	88.30	90.18	88.84	80.08	89.50	**81.96**
DA-FoC$_X^{MM}$	GPT-4o	10	**90.48**	**91.70**	**92.92**	**89.90**	**92.31**	78.07
DA-FoC$_I^{MM}$	GPT-4o	10	89.55	90.16	75.19	67.11	81.99	74.95

Table 6: Evaluation results of the final set of immigration FoCs with (1) DA-FoC^{MM} operating on multimodal SMPs from Twitter / X in dataset TS3, denoted as DA-FoC$_X^{MM}$ and (2) DA-FoC^{MM} operating on multimodal SMPs from Instagram in dataset TS4, denoted as DA-FoC$_I^{MM}$.

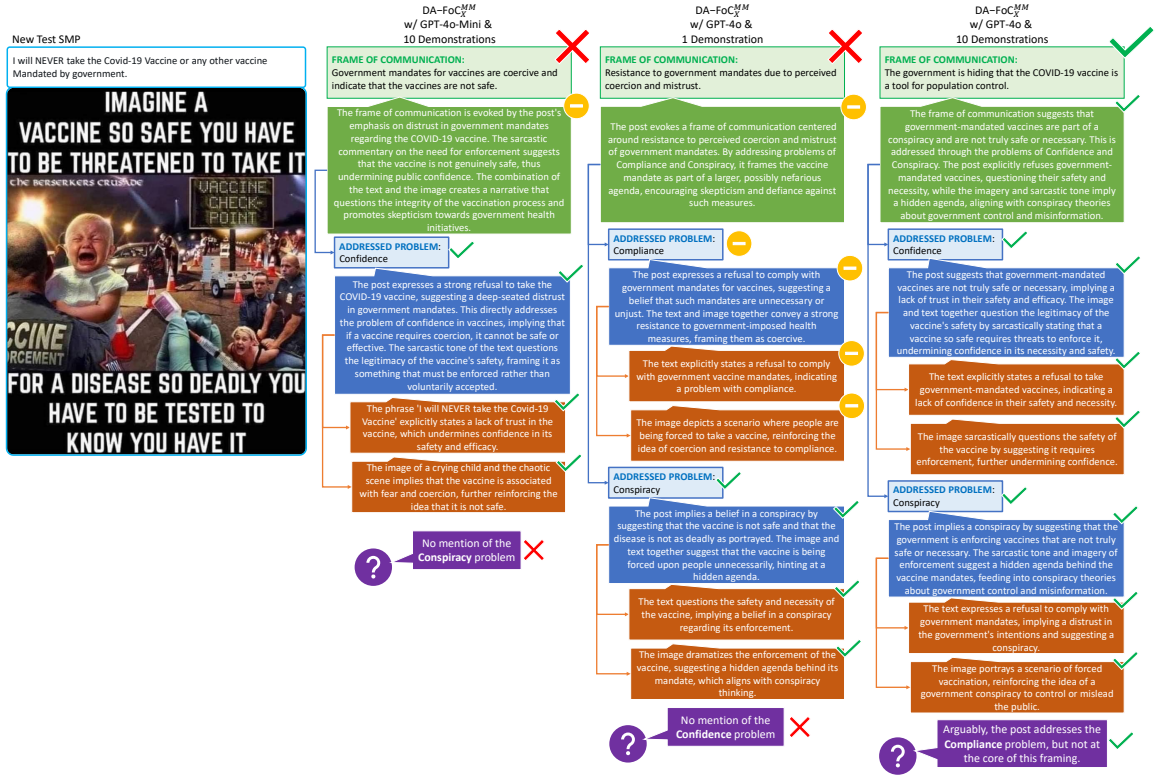

Figure 14: An error analysis of a multimodal SMP from Twitter / X across three of the evaluated systems.

Mini and 10 demonstrations, generates an FoC stating that "Government mandates for vaccines are coercive and indicate that the vaccines are not safe," as presented in Figure 14. While this system correctly identifies distrust toward government mandates, it makes two notable errors. First, it fails to identify the Conspiracy problem despite clear indications in both the text and the image. The imagery, which depicts a chaotic checkpoint scene with vaccine enforcement personnel, strongly implies a hidden agenda and aligns with conspiracy theories about government control. The omission of this problem leads to an incomplete interpretation of the SMP's framing. Second, the system overemphasizes skepticism regarding vaccine safety, which it identifies as the Confidence problem. Although this problem is relevant, the system's focus on safety results in neglecting the Conspiracy problem, which is central to the post's portrayal of coercion and control.

The second system, DA-FoC$_X^{MM}$ with GPT-4o and 1 demonstration, generates an FoC that frames the SMP as "Resistance to government mandates due to perceived coercion and mistrust." This FoC primarily addresses the Compliance problem but exhibits two critical issues, as shown in Figure 14. First, it fails to identify the Confidence problem, which is explicitly conveyed in the post's textual statement, "I will NEVER take the Covid-19 Vaccine." This statement indicates a lack of trust in the vaccine's safety and efficacy, which is a central aspect of the framing. The system's inability to cap-

ture this problem weakens its interpretation of the SMP's message. Second, the system provides only a limited interpretation of the Conspiracy problem. While it hints at mistrust, it does not explicitly recognize the conspiracy implications of the imagery, which strongly suggests government overreach and hidden agendas. This limitation results in an incomplete analysis of the post's visual components and fails to capture the interplay between the text and image.

The third system, DA-FoC$_X^{MM}$ with GPT-4o and 10 demonstrations, produces the most accurate and comprehensive analysis when compared to the others in Figure 14. It generates an FoC stating that "The government is hiding that the COVID-19 vaccine is a tool for population control." This FoC demonstrates a nuanced understanding of the SMP and correctly addresses multiple problems. The Confidence problem is identified through the explicit textual statement, "I will NEVER take the Covid-19 Vaccine," which conveys distrust in the vaccine's safety and necessity. The system also captures the Conspiracy problem by interpreting the imagery as portraying a scenario of government coercion and control. The chaotic enforcement checkpoint evokes associations with hidden agendas and misinformation, aligning with common conspiracy narratives. Finally, the system acknowledges the Compliance problem indirectly, with "The text expresses a refusal to comply with government mandates..." and "The image portrays a scenario of forced vaccination...", recognizing the refusal to comply with government mandates as part of the broader skepticism toward enforced vaccination policies. However, the system correctly determines that Compliance is not at the core of this FoC, and that Conspiracy is actually the more salient problem addressed.

The error analysis highlights significant differences in the performance of the three systems. The first system, DA-FoC$_X^{MM}$ with GPT-4o-Mini and 10 demonstrations, and the second system, DA-FoC$_X^{MM}$ with GPT-4o and 1 demonstration, fail to fully interpret the SMP due to omissions of key problems, particularly Confidence and Conspiracy. In contrast, the third system, DA-FoC$_X^{MM}$ with GPT-4o and 10 demonstrations, captures the interplay of all three addressed problems by the SMP - Confidence, Conspiracy, and Compliance - producing a nuanced and accurate FoC. This comparison underscores the importance of our high-quality demonstrations and advanced structured reasoning capabilities in achieving robust FoC discovery in multimodal content.

H Detailed Problem Analysis

We performed a deep dive into where the problems are addressed in SMPs discussing the COVID-19 vaccines, in order to measure the impact of images on framing vaccine hesitancy. Figure 15 illustrates how many SMPs addressed each of the 7C problems of vaccine hesitancy in (1) only the text of the SMP, (2) only the image, or (3) both the text and the image, across both Twitter / X and Instagram in dataset TS1 and dataset TS2.

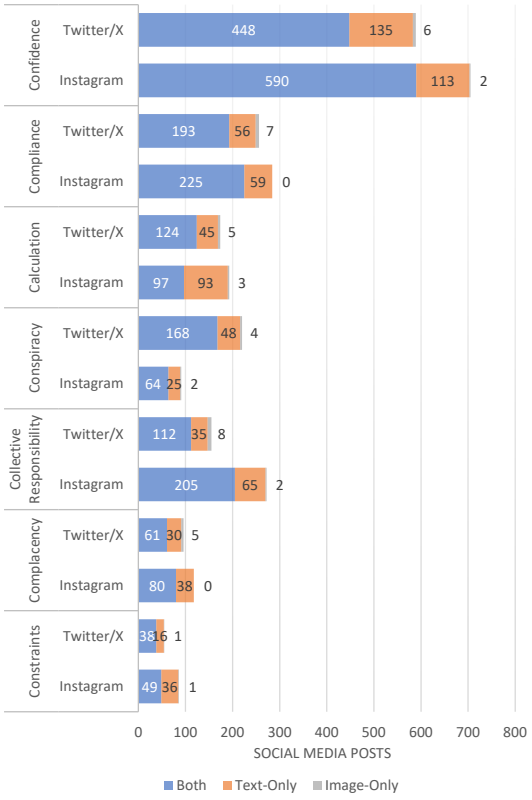

Figure 15: Number of Social Media Postings (SMPs) that address each of the COVID-19 vaccine hesitancy problems and evoke corresponding FoCs in different modalities, across Twitter / X and Instagram.

The rationales generated for Twitter / X and Instagram SMPs also revealed differences in *what* COVID-19 vaccine hesitancy problems were addressed on each platform. We found that SMPs on Instagram more often evoked FoCs that address Confidence, Collective Responsibility, and Constraints than Twitter / X, while SMPs on Twitter / X heavily focused on problems of Conspiracy. FoCs discovered from Twitter / X tended to address problems of Compliance (34%) and Complacency (12%) more often than FoCs from Instagram (33%

```yaml
response_format:
  type: "json_schema"
  json_schema:
    name: irag_demos
    schema:
      type: object
      properties:
        problem_rationales:
          type: array
          items:
            type: object
            properties:
              problem:
                description: The name of the frame problem addressed by the post. The problem must have a cause
                             articulated in the post, and must also be addressed by the provided frame of communication.
                type: string
                enum:
                  - "Confidence"
                  - "Complacency"
                  - "Constraints"
                  - "Calculation"
                  - "Collective Responsibility"
                  - "Compliance"
                  - "Conspiracy"
              explanation:
                description: Explain why this problem is addressed through a cause articulated in the post. Make sure to
                             explain how the image and text together contribute to addressing the problem.
                type: string
              locations:
                description: The locations in the post where this problem addressed.
                type: array
                items:
                  type: object
                  properties:
                    explanation:
                      description: Explain why this problem is addressed by this location in the post.
                      type: string
                    location:
                      description: The location in the post where this problem is addressed.
                      type: string
                      enum:
                        - "Text"
                        - "Image"
                  required:
                    - explanation
                    - location
                  additionalProperties: false
            required:
              - problem
              - explanation
              - locations
            additionalProperties: false
        frame_rationale:
          description: Explain why the provided frame of communication is evoked by this post, drawing upon each
                       addressed problem. Specifically articulate the frame of communication in your explanation.
          type: string
      required:
        - problem_rationales
        - frame_rationale
      additionalProperties: false
    strict: true
```

Figure 16: The JSON constrained decoding schema for Phase A, in YAML format.

and 10% respectively). Alternatively, FoCs from Instagram more often addressed problems of Constraints (21% vs. 11%), Calculation (27% vs. 25%), and Collective Responsibility (15% vs. 13%).

As Figure 15 demonstrates, significant context is lost when only considering the text of SMPs on either Twitter / X or Instagram, as a majority of the SMPs from both platforms employed both the text and the included image of their SMPs to evoke FoCs.

The figure provides important insights into the role of multimodality in framing COVID-19 vaccine hesitancy problems. The most striking observation is that the "Both" modality - where text and images work together - dominates across all seven problems of the 7C model, indicating that multimodal framing is a key strategy employed by social media users to evoke FoCs. For example, for the problem of Confidence, the highest number of SMPs utilize both text and images (448 for Twitter / X and 590 for Instagram). In contrast, SMPs that evoke Confidence using only the text or only the image are much less frequent.

A notable trend emerges when comparing platforms. Instagram consistently has more SMPs addressing Confidence and Collective Responsibility compared to Twitter / X, particularly in the multimodal category. This suggests that Instagram users may rely more heavily on visual components to evoke trust or solidarity-related FoCs. For example, the "Both" category for Confidence on Instagram (590) significantly outnumbers the corresponding figure on Twitter / X (448), highlighting the importance of visual persuasion on Instagram.

```yaml
response_format:
  type: "json_schema"
  json_schema:
    name: irag_frames
    schema:
      type: object
      properties:
        frames:
          type: array
          items:
            type: object
            properties:
              problems:
                type: array
                items:
                  type: object
                  properties:
                    explanation:
                      description: Explain why a problem is addressed through a cause articulated in the post. Make sure to
                                   explain how the image and text together contribute to addressing the problem.
                      type: string
                    locations:
                      description: The locations in the post where this problem addressed.
                      type: array
                      items:
                        type: object
                        properties:
                          explanation:
                            description: Explain why this problem is addressed by this location in the post.
                            type: string
                          location:
                            description: The location in the post where this problem is addressed.
                            type: string
                            enum:
                              - "Text"
                              - "Image"
                        required:
                          - explanation
                          - location
                        additionalProperties: false
                    problem:
                      description: The name of the frame problem addressed by the post. The problem must have a cause
                                   articulated in the post, and must be addressed by the discovered frame of communication.
                      type: string
                      enum:
                        - "Confidence"
                        - "Complacency"
                        - "Constraints"
                        - "Calculation"
                        - "Collective Responsibility"
                        - "Compliance"
                        - "Conspiracy"
                  required:
                    - explanation
                    - locations
                    - problem
                  additionalProperties: false
              frame_rationale:
                description: Explain why a frame of communication is evoked by this post, drawing upon each addressed problem.
                type: string
              frame:
                description: Articulate the evoked frame of communication.
                type: string
            required:
              - problems
              - frame_rationale
              - frame
            additionalProperties: false
      required:
        - frames
      additionalProperties: false
    strict: true
```

Figure 17: The JSON constrained decoding schema for Phase B, in YAML format.

For problems like Conspiracy and Calculation, multimodal SMPs are again the majority, but text-only posts play a more prominent role on Twitter / X. This reflects the platform's tendency for users to articulate conspiratorial or analytical reasoning through text, which may not require as strong of a visual component. For example, 48 Twitter / X SMPs addressed the Conspiracy problem using text only, compared to only 25 on Instagram. Similarly, Calculation sees higher numbers for text-only SMPs on Instagram (93) compared to Twitter / X (45).

Compliance is another notable category where Twitter / X demonstrates a greater prevalence of text-only posts (56) compared to Instagram (59 text-only posts, with zero relying on images alone). This reflects the platform-specific discourse styles, where Twitter / X often fosters debate over policies and mandates using textual arguments, while Instagram relies less on text alone to evoke Compliance-related FoCs.

On the other hand, FoCs interpreting Constraints and Complacency exhibit smaller numbers overall, but the trend remains consistent: multimodal SMPs dominate, followed by text-only posts, with image-only posts being the least frequent. For Constraints, the multimodal category accounts for 38 SMPs on Twitter / X and 49 on Instagram, while image-only

```yaml
response_format:
  type: "json_schema"
  json_schema:
    name: irag_judge_para
    schema:
      type: object
      properties:
        shared_problems:
          type: array
          items:
            type: object
            properties:
              explanation:
                description: Explain why this problem is addressed by both paraphrasing frames of communication (if any).
                type: string
              problem:
                description: The name of the frame problem addressed by both paraphrasing frames of communication (if any).
                type: string
                enum:
                  - "Confidence"
                  - "Complacency"
                  - "Constraints"
                  - "Calculation"
                  - "Collective Responsibility"
                  - "Compliance"
                  - "Conspiracy"
            required:
              - explanation
              - problem
            additionalProperties: false
        paraphrase_rationale:
          description: Explain why the novel frame of communication paraphrases any of the known frames of communication,
                       drawing upon each addressed problem, or why the novel frame of communication is new.
          type: string
        frame_id:
          description: The frame ID of the corresponding known frame of communication, which the novel frame of communication
                       paraphrases (if any).
          type:
            - "string"
            - "null"
      required:
        - shared_problems
        - paraphrase_rationale
        - frame_id
      additionalProperties: false
    strict: true
```

Figure 18: The JSON constrained decoding schema for Phase C, in YAML format.

contributions are negligible.

The importance of multimodal framing is further underscored by the observation that image-only SMPs contribute minimally across all problems. This suggests that images alone, while capable of evoking FoCs, are often insufficient to address complex vaccine hesitancy problems without accompanying textual support.

Figure 15 highlights the prevalence and significance of multimodal framing in vaccine hesitancy discourse. The combined use of text and images allows users to more effectively evoke and amplify FoCs, particularly for problems like Confidence, Conspiracy, and Compliance. Platform differences also emphasize the need to analyze multimodal content within its unique context, as Instagram places greater emphasis on visual persuasion, while Twitter / X exhibits a stronger reliance on text for FoC articulation.