Discovering and Articulating Frames of Communication from Social Media Using Chain-of-Thought Reasoning

Maxwell A. Weinzierl and **Sanda M. Harabagiu**
Human Language Technology Research Institute, The University of Texas at Dallas
{maxwell.weinzierl, sanda}@utdallas.edu

Abstract

Frames of Communication (FoCs) are ubiquitous in social media discourse. They define what counts as a problem, diagnose what is causing the problem, elicit moral judgments and imply remedies for resolving the problem (Entman, 1993). Most research on automatic frame detection involved the recognition of the problems addressed by frames, but did not consider the *articulation* of frames. Articulating an FoC involves reasoning with salient problems, their cause and eventual solution. In this paper we present a method for Discovering and Articulating FoCs (DA-FoC) that relies on a combination of Chain-of-Thought prompting (Wei et al., 2022a) of large language models (LLMs) with In-Context Active Curriculum Learning. Very promising evaluation results indicate that 86.72% of the FoCs encoded by communication experts on the same reference dataset were also uncovered by DA-FoC. Moreover, DA-FoC uncovered many new FoCs, which escaped the experts. Interestingly, 55.1% of the known FoCs were judged as being better articulated than the human-written ones, while 93.8% of the new FoCs were judged as having sound rationale and being clearly articulated.

1 Introduction

The way in which we interpret information depends on how the information is framed (Entman, 2003; Reese et al., 2001; Scheufele, 2004; Chong and Druckman, 2012; Bolsen et al., 2014). For instance, if information about vaccines is framed to build our confidence in them, we can become vaccine enthusiasts. The notion of Frame of Communication (FoC) has emerged from the Theory of Communication, studied in social sciences. Discovering FoCs is challenging because the FoCs are not directly expressed in texts, but rather texts *evoke* them, as shown in Figure 1. Framing entails emphasizing specific aspects of a topic within a text, guiding the audience towards a particular understanding. For the text illustrated in Figure 1, which is part of the discourse about COVID-19 vaccines on social media, the selected aspects are (1) the calculation people make about the personal costs and benefits of getting vaccinated; and (2) the complacency of getting vaccinated due to low perceived risk of infections. These aspects can be interpreted as problems related to vaccination. The two problems become *salient* to the FoC evoked by the text illustrated in Figure 1.

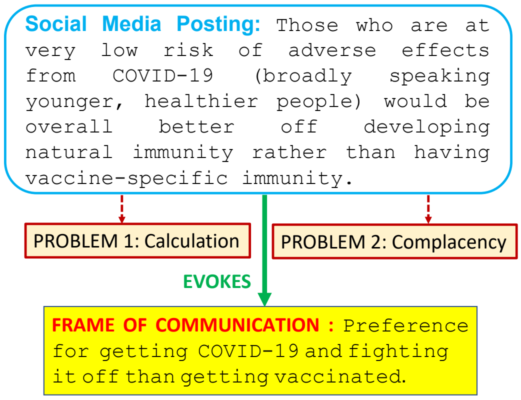

Figure 1: Frames of Communication (FoCs) evoked in Social Media Postings (SMPs).

In a widely cited definition, Entman (1993) notes that *"to frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described."* This means that, as a minimum, in addition to discovering the salient aspects of an FoC, we need to promote a causal interpretation of these aspects by *articulating* the FoC. In the FoC evoked by the text illustrated in Figure 1, the problem of calculation is caused by the preference for getting COVID-19 and fighting it off. The problem of complacency is caused by the assumption that getting COVID-19 is preferable to getting vaccinated. The final articulation of the FoC com-

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics Volume 1: Long Papers, pages 1617–1631
March 17-22, 2024

bines coherently both these causal interpretations of the problems. We note that the articulation of an FoC is expressing the reasons (or causes) of salient problems, but it is not explicitly mentioning the problems, instead it is implying them. Therefore the articulation of an FoC is a much harder NLP task than the discovery of FoCs and their salient problems.

Previous research addressing the problem of FoC discovery (Card et al., 2016; Naderi and Hirst, 2017; Field et al., 2018; Khanehzar et al., 2019; Kwak et al., 2020a; Mendelsohn et al., 2021) focused only on the discovery the salient problems implied by FoCs. This was due to the release of the Media Frames Corpus (MFC) (Card et al., 2015), which annotates fifteen dimensions of policy frames, addressing such problems as Constitutionality and Jurisprudence or Security and Defense. It is important to (1) discover when an FoC is evoked by a text; and (2) to be aware of which salient problems[1] are highlighted. However, without articulating the FoC, we cannot infer how the text should be interpreted. Moreover, without articulating FoCs, we ignore the many ways in which the same problem is framed in all texts that address it. But, as reported in (Van Gorp, 2010; Walter and Ophir, 2019; Vreese, 2005), the communication literature addresses mostly the inductive vs. deductive frame analysis, from which human inference of the articulation of the FoCs emerges. We believe that the reasoning capabilities of Large Language Models (LLMs) enable the automatic articulation of FoCs. This motivated us to design a method for Discovering and Articulating FoCs (DA-FoC).

Evidently, articulating FoCs involves reasoning with the problem(s) addressed in texts. Moreover, each articulated FoCs must be *relevant*, i.e. multiple texts should evoke it (Gamson, 1989). Therefore, discovering and articulating FoCs must consider that (1) FoCs may address one or more salient problems; (2) the FoC articulation needs to provide a rationale for each salient problem; and (3) the articulated FoC should be relevant. These requirements are very burdensome even for communication experts, who typically rely on codebooks emerging from their reasoning and painful inspection of large quantities of texts (Kwak et al., 2020b; Russell Neuman et al., 2014; Reese, 2007; Matthes and Kohring, 2008).

The recent ability of LLMs to perform complex reasoning provides an unprecedented opportunity for using them to simultaneously discover and articulate FoCs. In this paper we explore how Chain-of-Thought (CoT) prompting (Wei et al., 2022b) of LLMs can be used to reveal not only the problems addressed in texts but also the articulation of the FoCs. In addition, the CoT framework we used for DA-FoC benefits from *in-context active curriculum learning*, allowing the LLM to learn from its own mistakes. Because many FoCs discovered and articulated in this way may be paraphrasing each other, or they may be specializations of other FoCs, we also used CoT prompting to discover relations between FoCs. The relations between FoCs enabled us to select only FoCs that are relevant.

In designing our DA-FoC method, we focused on social media platforms where millions of users express their opinions and participate in conversations about issues of their interest. In their Social Media Postings (SMPs), often users select particular aspects, or problems, of an issue, revealing the reasons for their interest in the problem. In doing so, they evoke FoCs, as shown in Figure 1. In addition to using only SMPs, which present the advantage of text brevity, we considered only the discovery and articulation of FoCs regarding COVID-19 vaccines. This allowed us to rely on knowledge about salient problems characterizing vaccine hesitancy, reported in Geiger et al. (2021). It also allowed us to make use of the only reference dataset having expert-annotated FoCs which are articulated. In Weinzierl and Harabagiu (2022) 14,180 SMPs have been expert-annotated with 113 FoCs. We have enriched this dataset by asking communication experts to also judge which of the problems reported in Geiger et al. (2021) were implied in each FoC. Using this enriched dataset allowed us to train and test DA-FoC and to make the following contributions:

◁**1**▷ We introduce the first method that does not only discover FoCs from texts available in SMPs, but also articulates the FoCs by using CoT prompting of Large Language Models (LLMs) with In-Context Active Curriculum Learning (ICACL), a promising new method for prompting LLMs.

◁**2**▷ We describe the first method of discovering relations between FoCs, identifying paraphrases, specializations, and contradictions between them. We make available all prompts, annotations, articulated frames, and relations discovered between

[1]The dimensions of the Media Frames Corpus correspond to the problems highlighted by an FoC. The notion of Frame of Communication and Media Frame are used interchangeably in Communication Theory (Chong and Druckman, 2007).

Problem	*Definition of Vaccination Problem*
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 1: Problems associated with vaccine hesitancy.

frames on GitHub[2].

◁3▷ A by-product of our method is the identification of all social media postings evoking the same FoC, which informs its relevance.

◁4▷ We present the first DA-FoC method which uncovers not only many of the frames identified by experts on the same dataset, but it is also capable of uncovering many *new* frames, which are both clearly articulated and sound.

Because FoCs are known to be influential in shaping public opinions, the discovery of frames and their articulation can inform the messaging used in various communication interventions. For example, knowing which FoCs contain misinformation about vaccines is crucial to interventions meant to inoculate the public against misinformation. The discovery of FoCs will also impact argumentation mining, an NLP area that has recently received plenty of interest (Palomino et al., 2022; Sun et al., 2022; Ziegenbein et al., 2023).

2 Reference Dataset

To our knowledge, the only existing dataset of SMPs annotated with FoCs is CoVaxFrames, reported in Weinzierl and Harabagiu (2022). This dataset includes FoCs related to COVID-19 vaccination hesitancy. Vaccine hesitancy, as reported in Geiger et al. (2021), is characterized by seven factors, or problems, that increase or decrease an individual's likelihood of getting vaccinated. For each of the FoCs annotated in CoVaxFrames, four researchers have annotated the problems that they address. The problems are listed in Table 1 along with their definitions and the number of FoCs addressing each problem. The researchers obtained a very high inter-annotator agreement of 81%, with the remaining disagreements adjudicated through discussions. The newly annotated dataset became the reference dataset used by the method described in Section 3 and Section 4. The same training and testing splits were utilized as in Weinzierl and Harabagiu (2022).

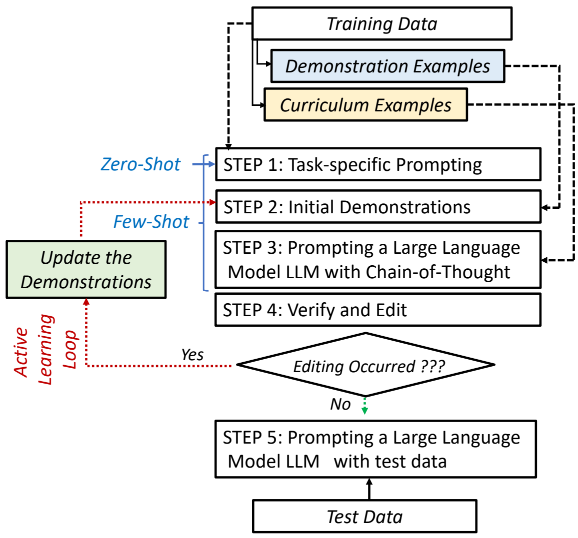

Figure 2: Chain-of-Thought Prompting with In-Context Active Curriculum Learning (CoT-ICACL).

3 The DA-FoC Method

The DA-FoC method has three distinct phases. In Phase A, FoCs are discovered and articulated using the CoT prompting with the In-Context Active Curriculum Learning (CoT-ICACL) framework illustrated in Figure 2. Since we noticed that some of the FoCs articulated in Phase A are paraphrases, while some FoCs were generalizations/ specializations of other FoCs, and also some FoCs contradicted each other, we used the same CoT-ICACL framework in Phase B to discover possible relations between FoCs. Because in Phases A and B we do not account for FoC relevance, in Phase C we tackle this necessary property, selecting the final set of FoCs.

3.1 Chain-of-Thought Prompting with In-Context Active Curriculum Learning

We considered the option of using CoT prompting of an LLM in three scenarios:

[2] https://github.com/Supermaxman/co-vax-frames-articulations

1. In a zero-shot learning scenario, the LLM prompt describes the task: in Phase A of the DA-FoC method, as detailed in Section 3.3, this involves the description of the task of FoC discovery and articulation, while in Phase B, as detailed in Section 3.4, this involves the definition of possible relations between the FoCs discovered in Phase A as well as the task of discovering them. This scenario is represented by Step 1 illustrated in Figure 2. However, the task of discovering and articulating FoCs is difficult because it requires not only knowledge, but also expert reasoning, as evidenced in the frame coding literature (Kwak et al., 2020b; Russell Neuman et al., 2014; Reese, 2007; Matthes and Kohring, 2008). Capturing the causal reasoning required by the articulation of FoCs or by the recognition of relations spanning FoCs is not possible in this scenario.

2. In a few-shot learning scenario, which corresponds to Steps 1-3 from Figure 2, following the task-specific prompting, we provide initial demonstrations of how the task is performed. Clearly, these demonstrations present how Phase-specific tasks are resolved and involve examples from the training data, as detailed in Section 3.3 and Section 3.4 respectively. Step 3 ends the few-shot learning, prompting the LLM to discover and articulate FoCs or to identify relations between FoCs , providing also their rationales. But, LLMs typically have a very restricted context length, which means only a few demonstrations may be provided to an LLM for in-context learning. Additionally, we need to decide the order in which the demonstrations are presented to the LLM, since this order can have a significant impact on performance (Dong et al., 2023; Zhao et al., 2021; Brown et al., 2020). This entails, as shown in Liu et al. (2022); Rubin et al. (2022) that for all the examples from the training data, we would need to have expert-quality rationales. This would generate a significant burden on communication experts, which we believe is not necessary. We could use instead Active Learning, which requires a smaller, manageable number of rationale examples to solve these issues.

3. A scenario that (a) takes advantage of human intervention in the CoT prompting, by creating the active learning loop illustrated in Figure 2; as well as (b) curriculum learning, such that the examples presented in Step 3 have a growing level of difficulty. Because we still use (repeatedly) CoT prompting of the LLM, but also rely on In-Context Curriculum Learning and Active learning, we call this scenario Chain-of-Thought Prompting with In-Context Active Curriculum Learning (CoT-ICACL). We note that in this scenario, we present initially a small number of demonstrations in Step 2, while this number grows in the following usages of the active learning loop, because if in Step 4, edits are performed on the results of Step 3, all those edits become new demonstrations available to the LLM when Steps 2-4 are performed again. Finally, when reaching Step 5, the LLM is prompted in the same way as in Step 3, however, this time, all examples from the test data are used.

3.2 Curriculum Learning in DA-FoC

We were inspired by recent reports (Maharana and Bansal, 2022) on the impact of curriculum learning on common sense reasoning. Thus, when learning a curriculum of examples used in Step 3 of CoT-ICACL, we have considered the two functions a curriculum should have: (1) ranking of examples in terms of difficulty; and (2) transitioning of easy to difficult examples during training. As in Elman (1993); Bengio et al. (2009), this entails learning a list of examples ordered by values of difficulty. For this purpose, we relied on two hypothesis:

Hypothesis 1: In Phase A of DA-FoC, when modeling the difficulty of discovering FoCs evoked by SMPs, our hypothesis was that the more similar the language of an FoC is to the language of the SMP that evokes it, the easier it is to discover, articulate and explain the rationale for the FoC. We have experimented with measuring the similarity between an SMP_i and an FoC_j by considering (a) Sentence-BERT (SBERT) (Reimers and Gurevych, 2019); (b) BertScore (Zhang* et al., 2020); (c) the Cross-Encoder introduced by Nogueira and Cho (2020) and (d) Misinfo-GLP (Weinzierl and Harabagiu, 2021). Appendix A details our experiments, which led us to conclude that the best distance should use SBERT. The function quantifying the difficulty of discovering and articulating from an SMP_i an FoC_j was defined as: $f_D(SMP_i, FoC_j) = ||p_i - f_j||_2$, where $p_i = SBERT(SMP_i)$ and $f_j = SBERT(FoC_j)$. The Euclidean distance is used because the same distance was employed in the objective function of SBERT (Reimers and Gurevych, 2019).

Hypothesis 2: In Phase B of DA-FoC, the difficulty of discovering possible relations among the FoCs resulting from Phase A used the hypothesis that FoCs articulated with similar language are more likely to be related. Therefore, the function $f_{RD}(FoC_A, FoC_B)$ quantifying the difficulty of predicting a relation between a pair of FoCs is defined as: $f_{RD}(FoC_A, FoC_B) = ||f_A - f_B||_2$, where $f_A = SBERT(FoC_A)$ and $f_B = SBERT(FoC_B)$.

3.3 Phase A of DA-FoC: Discovering and Articulating Frames of Communication

For Phase A of the DA-FoC approach, Steps 1, 2, 3 and 4 need to be tailored for the task of discovering and articulating FoCs.

Step 1 represents the task-specific prompting, which (a) instructs the LLM to use the definition of FoCs from Entman (1993) and (b) details of the task. The prompt is illustrated in Appendix B. The LLM is instructed to first produce a rationale for each FoC it may discover in each exemplified SMP, and then it is asked to articulate the FoC. Moreover, since more than one FoC may be evoked by the same SMP, the LLM is instructed to discover *all* FoCs evoked in an SMP.

Step 2 provides the demonstrations to the LLM. *Demonstration Examples:* A demonstration contains (a) an example SMP; (b) the rationale explaining why it evokes a FoC, highlighting the salient problems; and (c) the articulation of the FoC. A demonstration example is:

Social Media Posting Example:
One shot of COVID-19 vaccine is sufficient to make #pregnancy more risky and unsafe for unborn babies.
Rationale:
This social media posting contains a framing, as the problem of confidence in vaccine is challenged due to the perceived risk for pregnancies, affecting the unborn babies.
Frame of Communication:
The COVID vaccine renders pregnancies risky, and it is unsafe for unborn babies.

The few demonstrations provided to the LLM are selected when satisfying the requirements: (C1) all the problems addressed by the SMPs from the training data should be represented across the demonstration examples; (C2) some SMP examples should not evoke any FoC; (C3) some SMP examples should evoke more than one FoC; and (C4) overall, a small number of demonstration examples should be used, such that they can fit in the context allowed by the LLM.

Step 3 continues to use examples from the curriculum to generate prompts for the LLM. In each prompt only the SMP example is presented, the LLM automatically generating the rationale and articulating the evoked FoC.

Step 4 follows the Verify-and-Edit paradigm (Zhao et al., 2023), where the LLM's rationale and articulated FoCs are verified and edited if necessary. Whenever necessary, the human expert edits the rationales and the FoC articulations.

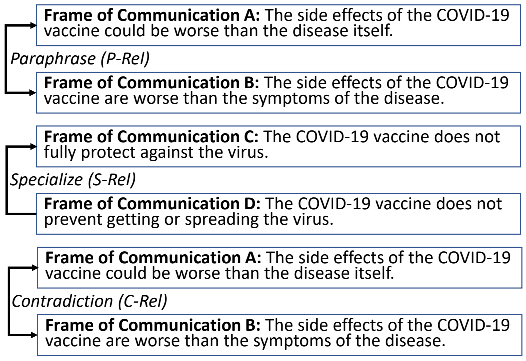

Figure 3: Examples of FoC relations.

3.4 Phase B: Discovering Relations between Frames of Communication

Three possible relations between the FoCs articulated by the LLM were observed, which are exemplified in Figure 3. Whenever a pair (FoC_A, FoC_B) used different words to address the same problems that had the same causes, we argue that they share a *Paraphrase* Relation (P-Rel). When a pair (FoC_D, FoC_E) address the same problem, but the cause articulated in FoC_D provides additional information than the cause articulated in FoC_E, we argue that they share a *Specialize* Relation (S-Rel). Unlike the P-Rel relations, which are symmetrical, the S-Rel relations are asymmetrical. Also, when a pair (FoC_E, FoC_F) address the same problems, but the causes are contradictory, we argue that they share a symmetrical *Contradiction* Relation (C-Rel).

In Phase B of the DA-FoC approach, we tailor Steps 1-3 from CoT-ICACL, illustrated in Figure 2, for the task of identifying relations between the FoCs discovered in Phase A.

Step 1: We instruct the LLM about the task of discovering relations between FoCs, showcasing each type of relation. The prompt is illustrated in Appendix B.

Step 2 provides a small number of demonstrations involving pairs of FoCs uncovered in Phase A and the relations between them. For each example, a rationale is provided along with the decision of the type of relation.

Demonstration examples: The demonstration examples of relations between FoCs had to also sat-

isfy the requirements: (T1) the arguments of the example relations had to address all the distinct problems addressed in the training set; (T2) some demonstration examples should use pairs of FoCs that do not participate in any relation and (T3) to account for the context size of the LLM, only a small number of demonstrations should be provided.
Building the rationale: For each demonstration example, a rationale of the relation is provided, explaining why a relation between the pair of FoCs exists as well as the type of relation.
Step 3 uses examples of pairs of CoTs from the curriculum to prompt the LLM to generate a rationale for a relation if one exists and to decide the type of relation.
Step 4 also follows the Verify-and-Edit paradigm, where whenever necessary, the human expert edits the rationales and the assigned FoC relations.

3.5 Phase C: Relevance of Frames of Communication

In addition to addressing salient problems, FoCs need to be relevant. In social media discourse, we measure the relevance of FoCs by the number of SMPs evoking each FoC, similarly to how relevance is measured for FoCs in news (Gamson, 1989). This number is available to us first from Phase A of the DA-FoC method, which allows us to collect all the examples of SMPs evoking each of the discovered FoC^*. However, due to the discovery of relations between FoCs made possible by Phase B, these relevance numbers need to be updated. First, we select only one FoC from each set of paraphrased FoCs PF_i, namely M-FoC, which is the most connected (through P-Rels) FoC in PF_i. The relevance of M-FoC is updated from the original number of SMPs evoking it to the sum of all SMPs evoking any FoC in PF_i. In this way, the discovery of P-Rels enables us to filter out FoCs that articulate the same causes of the same salient problems.

The S-Rels discovered in Phase B of the DA-FoC method enable us to organize FoCs in taxonomies, enabling us to implement the notion of *inherited* relevance. This entails that the relevance of an FoC_A having an S-Rel with FoC_B can be updated, to sum up its original relevance value to the relevance of FoC_B. Selecting a relevance threshold T_r results in the final set of FoCs, spanned by the final set of S-Rel and C-Rel relations. We note that because C-Rels reveal contrasting viewpoints of the problem causes, we retain all FoCs participating in such relations, to allow opposing interpretations due to these FoCs.

4 Evaluation Results

Quantitative Results: To compare the results of our method with a simple baseline, we considered a methodology that clustered all SMPs from the test data. Clustering was facilitated by creating SMP embeddings $p_i^* = SBERT(SMP_i^*)$ from the test set. Hierarchical Agglomerative Clustering (HAC) was employed from Ward (1963) with a variance gain threshold of 1.1, selected from initial experiments on the training data. For each cluster CL_j, the first sentence of the SMP_i closest to the centroid of CL_j was selected and placed in the set of final FoCs. Obviously, this baseline does not discover any relations between FoCs. Table 2 lists the number of FoCs uncovered by the HAC baseline method.

Four LLMs were considered in our evaluations of the DA-FoC framework: Vicuna-13B (Chiang et al., 2023; Zheng et al., 2023), LLaMa-2-70B (Touvron et al., 2023), GPT-3.5 (Ouyang et al., 2022), and GPT-4 (OpenAI, 2023). In Phase C we chose $T_r = 2$, corresponding to each FoC needing to be evoked by at least two SMPs. Further discussion surrounding this decision along with ablation results are provided in Appendix D. Furthermore, active learning loops with a minimum of 50 curriculum examples produced the best results from initial LLM experiments. Table 2 lists the number of discovered FoCs resulting from Phase A when using each LLM, the number of P-Rels, S-Rels, and C-Rels discovered in Phase B, and the number of final FoCs selected in Phase C. As Table 2 illustrates, zero-shot learning with GPT-3.5 and Few-Shot learning with Vicuna-13B failed to produce any meaningful FoCs, and therefore these configurations were not included in the qualitative results. A further discussion of the context limitations of the considered LLMs is provided in Appendix C.

Qualitative results: The quality of the final set of FoCs was evaluated in terms of three properties: (a) the *soundness* of the rationale provided by the LLM when articulating a FoC; (b) the *clarity* of the FoC articulation generated by the LLM; and (c) the *novelty* of the final set of FoCs when compared to the known FoCs in the reference dataset. Two linguists were tasked to judge the soundness, clarity, and novelty of final FoCs, with N_S FoCs deemed sound, and N_C FoCs deemed clear. With N_T final

CoT Prompting Method	System	Discovered FoCs	P-Rels	S-Rels	C-Rels	Final FoCs
-	HAC	-	-	-	-	321
Zero-Shot	GPT-3.5	-	-	-	-	-
Few-Shot	Vicuna-13B	27	-	-	-	-
Few-Shot	LLaMa-2-70B	2,006	49	615	567	48
Few-Shot	GPT-3.5	1,795	831	159	431	318
Few-Shot	GPT-4	2,021	875	499	177	331
CoT-ICACL	LLaMa-2-70B	2,142	293	132	384	340
CoT-ICACL	GPT-3.5	2,238	1,073	147	445	386
CoT-ICACL	GPT-4	2,374	586	636	146	292

Table 2: Number of FoCs discovered in Phase A; number and type of relations between FoCs discovered in Phase B, and final number of FoCs selected in Phase C.

CoT Prompting Method	System	Z	A	R	R_K	F_1	P_A
-	HAC	-	36.14	76.32	68.14	49.05	15.98
Few-Shot	LLaMa-2-70B	25.00	64.58	25.41	19.47	36.47	34.62
CoT-ICACL	LLaMa-2-70B	35.29	68.86	42.06	47.32	52.22	42.11
Few-Shot	GPT-3.5	5.03	41.19	70.43	51.33	51.98	28.08
CoT-ICACL	GPT-3.5	39.38	53.37	89.57	78.76	66.88	39.39
Few-Shot	GPT-4	79.46	78.25	89.62	73.45	83.55	70.97
CoT-ICACL	GPT-4	**97.60**	**95.89**	**94.92**	**86.73**	**95.40**	**93.81**

Table 3: Evaluation results of the final set of FoCs.

FoC proposed by each method, then *the quality of reasoning (Z)* involved in uncovering FoCs is $Z = N_S/N_T$ while *the quality of the articulation (A)* of FoCs is $A = N_C/N_T$.

While metrics Z and A capture the soundness and clarity of the final set of FoCs, we also considered four additional evaluation metrics that account for the novelty of the FoCs. For each F, which is a clearly articulated FoC, an expert linguist was asked to find if F conveys the same information as any F_R, representing the FoCs available from the reference dataset. When F and some F_R state the same thing, we consider F to be *known*, and thus not novel. Let N_K represent the number of known FoCs judged in this way, and N_F the total number of reference FoCs. This allows us to define two additional evaluation metrics: (1) the R metric, defined as $R = N_C/(N_C + N_F - N_K)$, which models *the recall of clearly articulated FoCs*; and (2) $R_K = N_K/N_F$ which accounts for *the recall of known FoCs* from all those available in the reference dataset. Finally, as we desire the FoCs to be both clearly articulated and fully recalled, we combine the A measure with the R measure into $F_1 = 2AR/(A + R)$. We also are interested in measuring the clarity of the novel FoCs, and therefore we use the evaluation metric $P_A = (N_C - N_K)/(N_T - N_K)$. Table 3 lists the results of all these evaluation metrics across all methods for discovering FoCs. However, because the clustering baseline does not involve any reasoning, it has no results for Z. Agreement between linguists was measured on a sample of 1000 judgments, with a Cohen's Kappa of 0.62 indicating moderate agreement (McHugh, 2012).

We also performed an evaluation of the relations between FoCs discovered by GPT-4 employing CoT-ICACL, given that this method produced the best results for discovering FoCs. Expert inspection revealed that 96.56% of these relations were correct. More specifically, 99.15% of P-Rels were correct, 96.54% of S-Rels were correct and 86.30% of C-Rels were correct. Mistakes are further analyzed in Appendix F.

System	Better	Equivalent	Worse
HAC	2.60%	18.18%	79.22%
GPT-3.5	26.97%	29.21%	43.82%
GPT-4	**55.10%**	**35.71%**	**9.18%**

Table 4: Comparing the articulation clarity of uncovered FoCs against reference FoCs.

5 Discussion

The results obtained when using CoT-ICACL with GPT-4 as the LLM are not only the best, but they are also impressive across all evaluation metrics. Even when using CoT-ICACL with GPT-3.5 as the LLM, our method obtained a substantial improvement over the baseline for all evaluation metrics.

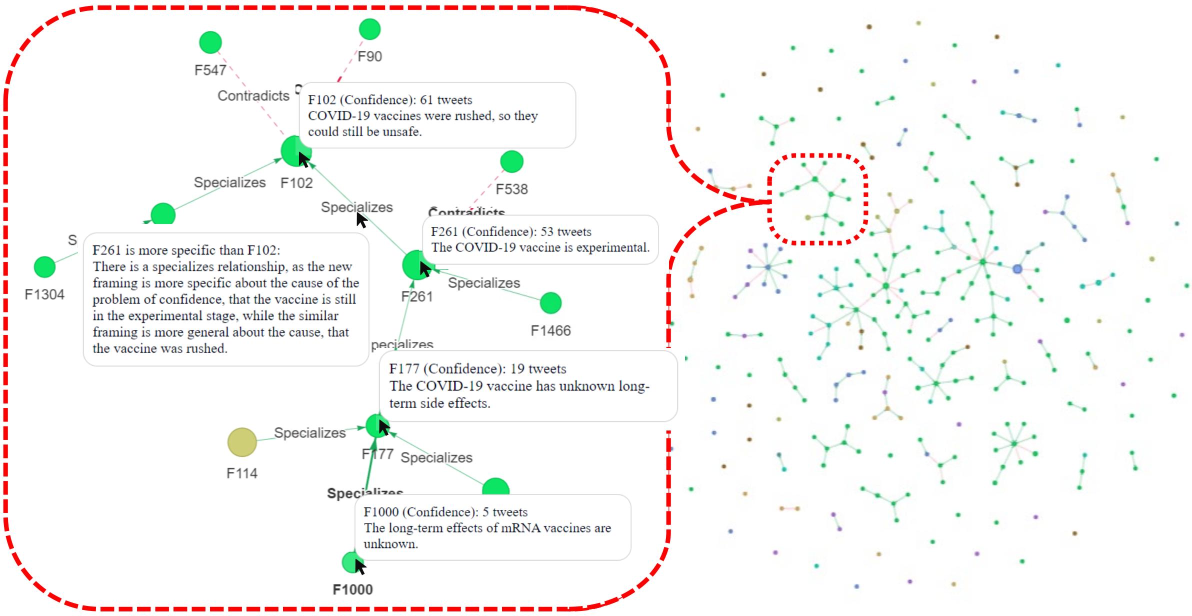

Figure 4: Interactive website enabling an exploration of the discovered FoCs, FoC relations, and FoC taxonomies discovered by GPT-4 employing CoT-ICACL for DA-FoC.

But unlike GPT-4, GPT-3.5 does not produce many sound rationales, as revealed by the results of the Z metric, showing that its reasoning capabilities are limited when compared to GPT-4 (Espejel et al., 2023). GPT-4 enabled the uncovering of many more clearly articulated FoCs, as captured by the A metric. Interestingly, many of the prompting methods were able to have good recall of the known FoCs, created by experts. But in terms of both clearly articulating FoCs and revealing all FoCs, only methods powered by GPT-4 were competitive, given the interpretation of the values of the F_1 metric. Furthermore, the values of the P_A evaluation results indicate that novel FoCs, which were not discovered by experts, were well articulated only when the used LLM was GPT-4. This makes us conclude that uncovering FoCs from SMPs can be performed with high values of soundness, clarity, and novelty when using GPT-4 and can be further improved with CoT-ICACL.

Articulation Quality: A different way of assessing the clarity of the FoC articulation is made possible when focusing only on the final FoCs (resulting from Phase C) which had the same content as some of the reference FoCs annotated in the reference dataset. For each pair of FoCs (F_K, F_R), where the uncovered F_K was judged by a computational linguist to convey the same information as a reference FoC F_R, the linguist was asked whether the articulation of F_K was (a) better, (b) worse, or (c) of the same clarity as F_R. The results of these judgments are listed in Table 4. As expected, the baseline method uncovers FoCs with vastly worse articulation clarity (79.22%) than the reference FoCs. The CoT-ICACL prompting of GPT-3.5 significantly improves the clarity of FoC articulation, uncovering 29.21% of known FoCs with the same clarity quality as the reference FoCs and even improving 26.97% of the clarity of uncovered known FoCs. The percentage of known FoCs articulated more clearly is an impressive 55.10% when CoT-ICACL used GPT-4, and only 9.18% of the known FoCs are articulated with poorer clarity. This indicates that CoT-ICACL with GPT-4 is capable of better articulating FoCs uncovered from social media than experts 55.10% of the time, while 37.71% of the time the FoCs are articulated with equivalent clarity. A 9.18% reduced clarity indicates that the need for expert intervention is greatly reduced. Examples are provided in Appendix E of discovered FoCs and their quality of articulation.

Organizing the FoCs: The rationales generated by CoT prompting with GPT-4 indicate the problems addressed by the uncovered FoCs. This allowed us to inspect the distribution of problems in the final set of FoCs obtained when using CoT-ICACL prompting with GPT-4. Our inspection indicates that a total of 174 FoCs (59.6%) address Confidence in vaccines; 39 FoCs (13.4%) address Collective Responsibility; 28 FoCs (9.6%) address Complacency; 23 FoCs (7.9%) address Compliance; 19 FoCs (6.5%) address Constraints; 15 FoCs

(5.1%) address Conspiracy; and 14 FoCs (4.8%) address Calculation. Surprisingly, one FoC (0.3%) addressed a new problem, namely Morality.

When using the CoT-ICACL prompting with GPT-4, we found that the 586 P-Rels between FoCs discovered allowed us to filter out 1,216 of the uncovered FoCs, as they were paraphrasing other FoCs. In addition, the S-Rels allowed us to generate 130 FoC taxonomies, spanned by S-Rels. These taxonomies contained on average 6 FoCs. The largest taxonomy contained 49 FoCs, with a depth of 7. Sometimes, in a FoC taxonomy, there were FoCs specialized as many as 13 times. The taxonomies will enable further research on the ideal specialization of an FoC articulation. We also found that the final set of FoCs contained 43 pairs of contradicting FoCs, demonstrating that opposing viewpoints were common.

An interactive website enabling an exploration of the discovered FoCs, FoC relations, and FoC taxonomies has been made available[3]. Figure 4 illustrates how this interactive website operates. Each node represents one of the final FoCs discovered when using the CoT-ICACL promting of GPT4, with the colors corresponding to the problems identified by CoT reasoning. Edges in the graph represent specializing and contradicting relations, since all paraphrases have been eliminated. Zooming in on the full graph enables an exploration of the various automatically constructed FoC taxonomies, and hovering over each node provides the articulated FoC along with the identified problems and the number of SMPs evoking the FoC. Hovering over the edges also provides the rationale justifying the relation spanning the pair of FoCs.

6 Related Work

Initial large-scale research on frame identification from social media has generally relied on unsupervised approaches (Neuman et al., 2014; Meraz and Papacharissi, 2013; de Saint Laurent et al., 2020) which revealed interesting framing patterns, highlighted by lexical terms, but did neither articulate any FoC nor discover any problems that FoCs address. Classifiers aiming to identify frame-invoking language were reported in Baumer et al. (2015), but these classifiers did not identify the problems addressed by FoCs. The assumption that frames can be associated with certain stock phrases was challenged in Tsur et al. (2015), showing that frames can also be associated with certain topics.

A growing body of research using supervised NLP methods uses the Media Frames Corpus (MFC) (Card et al., 2015). These methods detect frame salient problems with techniques including logistic regression (Card et al., 2016), recurrent neural networks (Naderi and Hirst, 2017), lexicon induction (Field et al., 2018), and fine-tuning pre-trained language models (Khanehzar et al., 2019; Kwak et al., 2020a). Furthermore, subcategories of the policy frame dimensions annotated in MFC were extracted with a weakly-supervised approach (Roy and Goldwasser, 2020).

The only prior work that considered the analysis of frames in social media was reported in Mendelsohn et al. (2021), where immigration policy problems were identified in SMPs with multi-label classification methods, relying on RoBERTa (Liu et al., 2019). All these prior methods do not articulate FoCs, they only discover them. We believe that the release of the reference dataset used in our work, which annotates both FoCs and the problems they address, will facilitate new research in the difficult problem of discovering and articulating FoCs. Finally, none of the previous methods have considered the need to learn to automatically provide a rationale for the discovered FoCs or for their salient problem(s), which our DA-FoC method enables by using Chain-of-Thought prompting of LLMs with In-Context Active Curriculum Learning.

7 Conclusion

This paper presents a new method capable to discover and articulate Frames of Communication from social media. By combining Chain-of-Thought prompting of LLMs with In-Context Active Curriculum Learning, both previously known and especially new frames were revealed. Extensive evaluations show that when using GPT-4 with CoT-ICACL, 86.73% of the frames identified by experts were re-discovered on the same dataset while also uncovering many new frames, which are both clearly articulated and sound. The rationales generated by GPT-4 with CoT-ICACL help us to make sense of these uncovered FoCs, providing additional insights for understanding why certain problems are discussed on social media. The relations between frames help us discover when some frames specialize others and when some frames contradict others.

[3] https://personal.utdallas.edu/~maxwell.weinzierl/discovery

8 Ethical Statement

We respected the privacy and honored the confidentiality of the users that have produced the SMPs pertaining to the dataset from Weinzierl and Harabagiu (2022). We received approval from the Institutional Review Board at the University of Texas at Dallas for working with this Twitter social media dataset. IRB-21-515 stipulated that our research met the criteria for exemption #8(iii) of the Chapter 45 of Federal Regulations Part 46.101.(b). Experiments were performed with high professional standards, avoiding evaluation on the test collection until a final method was selected from training performance. All experimental settings, configurations, and procedures were clearly laid out in this work, the supplemental material, and the linked GitHub repository. We do not perceive any major risks related to our research, as our work is in service of improving understanding of how COVID-19 vaccine hesitancy is framed on social media. The public good was the central concern during all enclosed research, with a primary goal of benefiting both natural language processing and public health research.

9 Limitations

The method capable to discover and articulate Frames of Communication that is introduced in this work focuses on social media posts from Twitter / X. Therefore, our methodology may not work as well on posts originating from other social media platforms, particularly platforms such as Reddit, where longer textual content is typical. Furthermore, our method relies only on the textual content of posts. Many social media posts use also images, videos, and other multimedia content. In future work, we plan to extend our methods by enabling them to discover and articulate Frames of Communication by considering the entire multimodal content of social media posts. In addition, we plan to also extend the social media platforms on which our methods can operate.

An important limitation of our approach stems from the need to have available a reference dataset of social media posts annotated with frames of communication that were discovered to be evoked in them. These frames of communication need to be discovered with inductive frame analysis (Van Gorp, 2010) on the set of social media posts. The postings evoking each frame from this repertoire of frames of communication also need to be known. This requires significant efforts from communication experts. In addition, the problems revealed by each frame need to be annotated such that our chain-of-thought prompting methodology may have demonstrations. Semi-automatic methods that propose the frames of communication evoked in social media posts and predict the problems that are addressed by the frames are considered in our future work, to alleviate these limitations.

Finally, our method only considered frames of communication for *"COVID-19 Vaccines"* due to the only existing dataset where frames of communication are annotated. Therefore, we could consider additional datasets that may cover a variety of topics, such as the policy problems addressing immigration, tobacco, or same-sex marriage, which are covered in the Media Frames Corpus (MFC) (Card et al., 2015). In future work, we shall contemplate the discovery of frames of communication for a variety of topics and domains.

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.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Dennis Chong and James N. Druckman. 2007. Framing public opinion in competitive democracies. *American Political Science Review*, 101:637 – 655.

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

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Jeffrey L. Elman. 1993. Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71–99.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communications*.

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

Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. 2023. Gpt-3.5 vs gpt-4: Evaluating chatgpt's reasoning performance in zero-shot learning.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. In *Conference on Empirical Methods in Natural Language Processing*.

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. 2021. Measuring the 7cs of vaccination readiness. *European Journal of Psychological Assessment*, pages 1–9.

Shima Khanehzar, Andrew Turpin, and Gosia Mikołajczak. 2019. Modeling political framing across policy issues and contexts. In *Australasian Language Technology Association Workshop*.

Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020a. A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. *Proceedings of the 12th ACM Conference on Web Science*, pages 305–314.

Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020b. 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.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

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. *ArXiv*, abs/1907.11692.

Adyasha Maharana and Mohit Bansal. 2022. On curriculum learning for commonsense reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 983–992, Seattle, United States. Association for Computational Linguistics.

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.

Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. In *Recent Advances in Natural Language Processing*.

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.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert.

OpenAI. 2023. Gpt-4 technical report.

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.

Alonso Palomino, Khalid Al Khatib, Martin Potthast, and Benno Stein. 2022. Differential bias: On the perceptibility of stance imbalance in argumentation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 411–421, Online only. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Conference on Empirical Methods in Natural Language Processing*.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

W. Russell Neuman, Lauren Guggenheim, S. Mo Jang, and Soo Young Bae. 2014. The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data. *Journal of Communication*, 64(2):193–214.

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

Yang Sun, Bin Liang, Jianzhu Bao, Min Yang, and Ruifeng Xu. 2022. Probing structural knowledge from pre-trained language model for argumentation relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3605–3615, Abu Dhabi, United Arab Emirates. 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 M. J. Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Annual Meeting of the Association for Computational Linguistics*.

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

Claes H. Vreese. 2005. News framing: Theory and typology. *Information Design Journal*, 13(1):51–62.

Dror Walter and Yotam Ophir. 2019. News frame analysis: An inductive mixed-method computational approach. *Communication Methods and Measures*, 13(4):248–266.

Joe H. Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022a. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. 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 A. Weinzierl and Sanda M. Harabagiu. 2021. Automatic detection of covid-19 vaccine misinformation with graph link prediction. *Journal of Biomedical Informatics*, 124:103955.

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.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, Toronto, Canada. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363, Toronto, Canada. Association for Computational Linguistics.

A Difficulty Modeling Experiments

Model	Accuracy
Cross-Encoder	59%
Misinfo-GLP	63%
BERTScore	67%
SBERT	71%

Table 5: Difficulty function results from initial experiments with different difficulty models.

Initial experiments were conducted on the CoVaxFrames dataset to determine which models of difficulty could serve to guide curriculum learning. 5 FoCs were manually selected from CovAxFrames to serve as a reference for difficulty models. For each of the selected FoCs, 20 pairs of SMPs were sampled for a total of 100 pairs of SMPs. An expert linguist judged which of the two SMPs in each pair was more difficult to recognize as evoking the respective FoC, which enabled measuring how accurately different difficulty models aligned with these human preferences, similar to Reinforcement Learning with Human Feedback (Christiano et al., 2017). Table 5 illustrates the accuracy of the various difficulty models considered in Section 3.

The Cross-Encoder approach, introduced by Nogueira and Cho (2020), employs a BERT-based model to measure relevance and was trained on MS-MARCO (Nguyen et al., 2016). The Misinfo-GLP method (Weinzierl and Harabagiu, 2021) employs graph-link prediction to identify whether an SMP evokes a misinformation FoC about COVID-19 vaccines. BERTScore (Zhang* et al., 2020) employs BERT to measure the F_1 score between the contextualized embeddings of a reference sequence and a candidate sequence. Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) produces sentence-level embeddings trained contrastively to be close together in Euclidean distance if the semantics of the sentences are similar. SBERT clearly resulted in the closest aligned measure of difficulty, with an accuracy of 71% in modeling human judgments of difficulty for recognizing frame evocation. Therefore, we utilized SBERT for all difficulty modeling in In-Context Active Curriculum Learning.

T_r	Final FoCs	Z	A	R	R_K	F_1	P_A
2	292	97.60	95.89	94.92	86.73	95.40	93.81
3	157	96.82	95.54	74.63	54.87	83.80	92.63
4	99	96.97	93.94	56.02	35.40	70.19	89.83
5	73	97.26	91.78	44.97	27.43	60.36	85.71

Table 6: Ablation evaluation results over the relevance threshold from Phase C, producing the final set of FoCs for CoT-ICACL with GPT-4.

B Chain-of-Thought Prompting Details

The task-specific prompt provided for Phase A of DA-FoC (a) instructs the LLM to use the definition of FoCs from Entman (1993) and (b) details of the task. The prompt is illustrated in Figure 5.

Frames of communication select particular aspects of an issue and make them salient in communicating a message. Social science stipulates that discourse almost inescapably involves framing – a strategy of highlighting certain issues to promote a certain interpretation or attitude. It has been argued that "to frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote problem definition, causal interpretation, moral evaluation, and/or treatment recommendation."

The Task:

You will be tasked with identifying and articulating vaccine hesitancy framings on the social media postings. You should discuss your reasoning first, and then provide a final decision. Each social media posting provided may or may not contain one or more frames of communication, so your first step is:

(a) Reason about whether the posting contains a frame (or more frames), or just states something factual or an experience.

If the posting contains a frame, the next step is

(b) Articulate that frame succinctly.

You will perform these steps until the answer to (a) is false, either because there are no frames in the posting, or because you have already articulated all the frames.

Figure 5: Task definition prompt for Phase A, the articulation of FoCs from SMPs for DA-FoC.

The LLM is asked to first produce a rationale for each FoC it may uncover in each exemplified SMP, and then it is asked to articulate the FoC. Moreover, since more than one FoC may be evoked by the same SMP, the LLM is instructed to uncover *all* FoCs evoked in an SMP. Similarly, the task-specific prompt provided for Phase B of DA-FoC is illustrated in Figure 6.

C Context Length Limitations

All LLMs considered in Section 4 have a limited context length, defined by the number of tokens the LLM can consider in a single prompt. Table 7 presents the maximum context lengths possible for each of the considered LLMs. We note that Vicuna-13B has such a small context that it can barely fit the task-specific prompt and necessary demonstrations for few-shot learning, and this limitation is likely why Vicuna-13B performed so poorly in our evaluations, discussed in Section 4. However, LLaMa-2-70B, GPT-3.5, and GPT-4 had no problem including demonstrations for few-shot learning and In-Context Active Curriculum Learning.

Model	Max Context Length
Vicuna-13B	2,048
LLaMa-2-70B	4,096
GPT-3.5	4,096
GPT-4	8,192

Table 7: Maximum context length comparisons between LLMs used for CoT-ICACL.

D Ablation Experiments over Relevance Threshold

The relevance threshold $T_r = 2$ corresponds to requiring two or more SMPs to evoke each FoC for that FoC to be considered relevant. Higher relevance thresholds can be considered, which produce a different final number of FoCs when employing CoT-ICACL with GPT-4, illustrated in Table 6. Further manual judgments were performed on $T_r > 2$, also provided in Table 6. As the threshold for relevance increased, fewer and fewer final FoCs were produced leading to a major decrease in recall metrics. Interestingly, we also see a noticeable decline in the quality of new FoCs, measured by P_A, which could indicate that the new high-quality FoCs discovered with $T_r = 2$ correspond more often to FoCs with lesser relevance. Human annotators likely missed these FoCs in constructing CoVaxFrames because much fewer SMPs evoke them. Furthermore, as the test collection is only a representative sample of 2,113 SMPs, it was difficult to justify $T_r > 2$, as $T_r = 2$ already corresponds to 0.1% of the population of SMPs. If we assume this sample is representative, then $T_r = 2$ would correspond to a minimum evocation of approximately 470 SMPs per month for each

Frames of communication select particular aspects of an issue and make them salient in communicating a message. Social science stipulates that discourse almost inescapably involves framing – a strategy of highlighting certain issues to promote a certain interpretation or attitude. It has been argued that "to frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote problem definition, causal interpretation, moral evaluation, and/or treatment recommendation."
The Task:
You will be tasked with identifying relationships between vaccine hesitancy framings. You should discuss your reasoning first, and then provide a final decision. Each framing provided may or may not be involved in a single relationship with one framing from a provided set of similar framings. We will consider three possible relationships:
1. Paraphrases(X,Y): X and Y say essentially the same exact thing, with different words or phrasing. If one person agreed with X, they would agree with Y, and vice versa. Frames should share the same cause and the same problem to be considered paraphrases.
2. Specializes(X,Y): X is a more specific or detailed framing of Y. Notice the order of X and Y is important for this relationship, as X is more specific and Y is more general. Frames should share the same problem, but have more specific or general causes to be considered specializes.
3. Contradicts(X,Y): X and Y contradict each other, such that they frame the same exact issue from opposing perspectives. If one person agreed with X, they would disagree with Y, and vice versa. Be extremely careful with the contradicts relationship, as we do not want two frames to contradict simply because they say the vaccine is safe vs unsafe, the frames need to have the same cause to contradict, such as safe due to being tested vs unsafe due to being rushed. The two frames X and Y should essentially paraphrase each other, sharing the same problem and cause but from opposing perspectives.
4. No relationship: There are no relationships between the new framing and any of the provided framings.
You should
(a) Reason about if the framing holds one of the above relationships with any of the provided framings.
Multiple relationships could be true, but prioritize in the order provided: If a paraphrase relationship holds, it must be provided.
If there is no paraphrase, then look for specialize. If there is a specialize relationship, provide it, otherwise look for contradicts.
Finally, if there is no contradicts relationship, answer no relationship.
If a relationship is identified, then
(b) State that relationship, using the IDs for each framing.

Figure 6: Task definition prompt for Phase B, the discovery of FoC relations for DA-FoC.

FoC, using the collection criteria from Weinzierl and Harabagiu (2022).

E Successful and Erroneous FoC Examples and Relations Spanning Them

An example of a known uncovered FoC which was judged to be more clear than an FoC discovered by experts on CoVaxFrames is FoC_2 :*"Preference for getting COVID-19 and fighting it off than getting vaccinated"*, the known FoC, and FoC_3 : *"Natural immunity is better than vaccine immunity"*, a FoC discovered by GPT-4 with CoT-ICACL. An example of an uncovered FoC that was not known and is clear as well as sound is FoC_4: *"Avoiding people is a better strategy than getting the COVID-19 vaccine"*. The rationale generated by CoT for FoC_4 is: *"The problem of calculation is due to the cause that a trade-off is being made, where taking the vaccine is not worth the calculated risk when compared to avoiding people."* Also, an example of a newly discovered FoC_5 which specializes some FoC_6 can be provided for FoC_5: *"People should make their own decisions about COVID-19 vaccination without being chastised"* and FoC_6 : *"People should make informed decisions about COVID-19 vaccination."* An example of contradictory FoCs is established between FoC_7: *"Getting the COVID-19 vaccine will protect those who cannot get the vaccine"* and FoC_8: *"The COVID-19 vaccine only benefits the recipient."* These examples show that in addition to uncovering and articulating FoCs from social media, the method that we have presented discovers interesting and informative relations between FoCs. Moreover, the rationales generated to make sense of these FoCs provide additional insights for understanding why certain problems are discussed on social media.

F Errors in Articulated FoCs and FoC Relations

A closer inspection of the edited demonstrations from Phase A of the curriculum built for GPT-4 demonstrates the kinds of early mistakes, which were corrected through editing with CoT-ICACL. GPT-4 mistakenly only articulated a single FoC, when the prompted SMP evoked multiple FoCs, for five out of the six edited demonstrations. The sixth demonstration had sound rationale, but an overly verbose articulation of the FoC. In Phase B, GPT-4 required 20 examples to be edited, where 7 edited examples involved incorrect P-Rels on FoCs which shared problems; 6 edited examples included missed P-Rels; 4 examples were edited where GPT-4 incorrectly directed the S-Rel, and 3 edited examples were added for C-Rels which were incorrectly identified once as a P-Rel, and twice as no relation.