SafePersuasion: A Dataset, Taxonomy, and Baselines for Analysis of Rational Persuasion and Manipulation

Haein Kong, A M Muntasir Rahman, Ruixiang Tang, Vivek K. Singh
Rutgers University
{haein.kong,ammuntasir.rahman,ruixiang.tang,v.singh}@rutgers.edu

Abstract

Persuasion is a central feature of communication, widely used to influence beliefs, attitudes, and behaviors. In today's digital landscape, across social media and online platforms, persuasive content is pervasive, appearing in political campaigns, marketing, fundraising appeals, and more. These strategies span a broad spectrum, from rational and ethical appeals to highly manipulative tactics, some of which pose significant risks to individuals and society. Despite the growing need to identify and differentiate safe from unsafe persuasion, empirical research in this area remains limited. To address this gap, we introduce SAFEPERSUASION, a two-level taxonomy and annotated dataset that categorizes persuasive techniques based on their safety. We evaluate the baseline performance of three large language models in detecting manipulation and its subtypes, and report only moderate success in distinguishing manipulative content from rational persuasion. By releasing SAFEPERSUASION, we aim to advance research on detecting unsafe persuasion and support the development of tools that promote ethical standards and transparency in persuasive communication online[1].

1 Introduction

Persuasion refers to a linguistic style or an act that aims to influence others' attitude or behavior (Pauli et al., 2024; El-Sayed et al., 2024; Wang et al., 2019). There has been much academic attention from various fields to persuasion (Bassi et al., 2024), due to its ubiquity and importance in society and human lives. Particularly, the natural language processing (NLP) field has actively investigated persuasion, focusing on human conversations for persuading others to encourage donation (Wang et al., 2019) or linguistic characteristics and interaction dynamics of successful persuasive online comments (Tan et al., 2016).

While previous NLP studies investigate persuasive language and diverse strategies, most of them treat persuasion as an umbrella term (Wang et al., 2019; Jin et al., 2024; Pauli et al., 2024), paying little attention to the unsafe or unethical types of persuasion. In social science, researchers have attempted to distinguish manipulation (Susser et al., 2019) or vicious persuasion (Bassi et al., 2024; Godber and Origgi, 2023), based on the degree of harm to individual autonomy. The discussion on the ethical concerns and safety in persuasion started to be facilitated in the NLP field, as AI gained improved persuasive skills (Durmus et al., 2024; Hendrycks et al., 2023; Kong, 2025). For example, El-Sayed et al. (2024) focused on the different types of persuasion based on their mechanism, defining rational persuasion involves "appeals to reason, evidence, and sound argument" that facilitates rational and reflective reasoning, while manipulation takes advantage of "cognitive biases and heuristics in a way that diminishes cognitive autonomy". From this perspective, manipulation contains process harm as it limits one's autonomy by restricting rational and reflective thinking. Therefore, rational persuasion can be considered safe persuasion, while manipulation is not.

This classification can offer a useful framework for understanding the safety in persuasion. However, there is a lack of empirical studies and public datasets focusing on these two different persuasion types. This results in a limited understanding of safety in persuasion, which prevents a deeper discussion or advanced research in this domain. Although our work focuses on human persuasion, we build on this growing recognition that understanding the boundary between rational persuasion and manipulative influence is a prerequisite for both responsible NLP modeling and future AI safety research.

[1]Our dataset and code are available at https://github.com/haeinkong/SafePersuasion

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1097–1111

In addition, it is essential to have empirical resources of rational persuasion and manipulation to prevent users from manipulative attempts on online platforms and promote civil online environments. Given the use of large language models (LLMs) as a judge in the persuasion domain (Bozdag et al., 2025) and automated content moderators (Kolla et al., 2024), testing current LLMs' capability in detecting manipulation over rational persuasion will offer insights into their applications. Therefore, this paper aims to fill this gap by constructing empirical resources, including a taxonomy and a dataset, that can contribute to the field, and testing LLMs to provide the baseline performances on the proposed tasks.

Our contributions are as follows:

- We introduce a new task for persuasion safety focused on distinguishing rational persuasion from manipulation and identifying the specific sub-techniques used.
- We provide a two-level taxonomy to distinguish rational persuasion and manipulation, and construct a human-annotated dataset, SAFEPERSUASION, which consists of 1,887 online comments.
- We provide baseline classification results on the binary and multi-label prediction tasks by experimenting with GPT-4.1, Llama-3.2-3B, and Claude-3.5-Haiku using zero-shot, few-shot, and chain-of-thought prompt strategies to support future developments in this space.

2 Related Works

2.1 Persuasion in NLP

Persuasion has been actively studied in the NLP field. Early works study the dynamics of persuasive discussion on an online platform to understand successful persuasion attempts (Tan et al., 2016). Other works focus on persuasive conversation in specific domains, such as donation (Wang et al., 2019), advertisements (Singla et al., 2022), or movie recommendation (Hayati et al., 2020), and develop diverse methods to detect persuasive techniques. Generating persuasive text is another research problem actively studied in this domain. For example, Samad et al. (2022) proposed an empathetic persuasive dialogue system to enhance the ability to make empathetic connections in persuasive systems. In recent years, these research areas have grown rapidly with the development of LLMs (Rogiers et al., 2024), as researchers attempt to utilize LLMs to assess persuasive language (Bozdag et al., 2025) and generate persuasive texts (Jin et al., 2024; Pauli et al., 2024).

A growing number of studies on persuasion fosters discussion on ethical and safety concerns in persuasion. For example, recent works highlight the different forms of persuasion based on their mechanisms, depending on whether the persuasion attempts exploit cognitive heuristics (e.g., manipulation) or encourage reflective thinking (e.g., rational persuasion) (El-Sayed et al., 2024; Jones and Bergen, 2024). While distinguishing between safe and unsafe persuasion becomes more important, it has been understudied in this field. The lack of studies and resources on safety in persuasion can result in risky scenarios, such as applications built based on a potentially unsafe persuasion dataset or users being exposed to manipulative persuasion.

Table 1 summarizes publicly available datasets in the persuasion domain and their characteristics. The existing datasets do not have a safety label of persuasion techniques, but treat all techniques under the umbrella term, persuasion. This research aims to fill a critical gap by incorporating both safety labels and persuasion techniques with human-annotated data, addressing a key need in the literature.

2.2 Manipulative Language Detection

While there is no universal definition for manipulative language, it is generally understood as subtle and nuanced language, as it does not rely on explicit offensive language such as profanity, insults, hate, or violence. One example of manipulative language is propaganda. Propaganda aims to influence people's actions or opinions for advancing a specific agenda through diverse rhetorical and psychological techniques (e.g., name-calling, repetition, slogans, etc) (Martino et al., 2019, 2020a), investigating news articles generated by media news outlets (Martino et al., 2020a,b; Yu et al., 2021).

More recently, Wang et al. (2024b) studied mental manipulation, which focuses on "a language to influence, alter, or control an individual's psychological state or perception for the manipulator's benefit". Their focus is on detecting the abuse in interpersonal conversations on movie dialogues. Recent work studied manipulative conversations in the courtroom, where each speaker has a role such as plaintiff, defendant, judge, and others relevant roles (Sheshanarayana et al., 2025). They aimed to detect manipulation, the manipulator, and the

Dataset	Source	Domain	Safety Label	Persuasion Technique	Human Annotation
WinningArgument (Tan et al., 2016)	Human	Diverse	✗	✗	✗
PersuasionForGood (Wang et al., 2019)	Human	Donation	✗	✓	✓
DailyPersuasion (Jin et al., 2024)	LLM	Diverse	✗	✓	✗
PersuasivePairs (Pauli et al., 2024)*	Human, LLM	Diverse	✗	✗	✗
SAFEPERSUASION (Ours)*	Human	Diverse	✓	✓	✓

Table 1: Comparison of persuasion datasets. The source column shows who created the dataset, whether it was created by humans or LLMs. The safety label column shows whether the dataset has the index of the safety of the persuasion. The persuasion techniques column shows whether the dataset includes the persuasion techniques being used. Lastly, the human annotation column indicates that the labels are generated by humans. SAFEPERSUASION is the first dataset with human-annotated safety labels and persuasion techniques. * means that it uses WinningArgument (Tan et al., 2016) to construct its dataset.

technique using the courtroom transcripts.

While various manipulative languages have been explored, we found that relatively little attention has been paid to manipulation in the general persuasion context on online platforms. Previous works have studied the verbal conversations from specific contexts, such as movie dialogue (Wang et al., 2024b) or courtroom (Sheshanarayana et al., 2025) or news articles (Martino et al., 2020b). These conversations are different from the everyday experience of ordinary people, or do not reflect casual conversation in online settings. Therefore, this paper investigates manipulation on online platforms, especially focusing on everyday conversation. This makes our dataset unique and useful for cases where the previous works' contribution is limited.

3 Proposed Taxonomy

Recent works proposed a taxonomy of rational persuasion and manipulation (El-Sayed et al., 2024; Jones and Bergen, 2024). However, their taxonomies exhibit several challenges that limit their applications for human annotation. For instance, Jones and Bergen (2024) focused on the high-level classification, lacking details of what specific techniques belong to each persuasion type. While El-Sayed et al. (2024) offered an extensive taxonomy including the details of sub-persuasion types, they provided them only for manipulation, which could limit their applications to study rational persuasion. Their taxonomy included techniques with abstract definitions (e.g., gaslighting) and that have overlap in their meanings (e.g., threats and negative emotions), which challenge the annotation and interpretation.

Therefore, we built a taxonomy for the dataset creation and human annotation that contains details for rational persuasion and manipulation. To construct our taxonomy, we conducted the following process: First, we reviewed previous studies that created a persuasion dialogue dataset (Wang et al., 2019; Hayati et al., 2020; Jin et al., 2024), used various persuasion strategies (Zeng et al., 2024), and focused on manipulative strategies (El-Sayed et al., 2024; Zhong et al., 2024; Braca and Dondio, 2023) to collect frequently used persuasive techniques and their definitions. After building a corpus, we removed the techniques with the same or close definitions, abstract definitions, and those beyond the scope of our study (e.g., misinformation). In this process, we kept the number of sub-categories limited to support interpretability and reasonable representation across categories. We included only the essential techniques that are distinguishable from each other.

Lastly, we determined whether each technique belongs to manipulation or rational persuasion based on the previous research and its definitions. For example, persuasion techniques such as negative emotional appeal, othering, and social conformity are defined as manipulative strategies (El-Sayed et al., 2024). Prior studies have suggested that status quo bias, authority appeal, and scarcity appeal are grounded in cognitive shortcuts or bias (Zhong et al., 2024; Braca and Dondio, 2023). Other techniques, especially for rational persuasion, are carefully classified based on whether the technique relies on reflective or logical thinking. The details, including references that we used for the definitions, are presented in Appendix A.

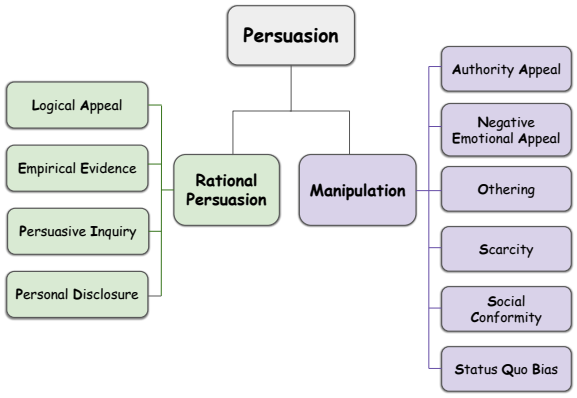

Figure 1: Our proposed two-level taxonomy of persuasion. It has rational persuasion and manipulation in the first level and the sub-persuasion techniques in the second level.

As a result, we constructed a two-level taxonomy of persuasion (Figure 1). The first level is the two types of persuasion: rational persuasion and manipulation. Then, the second level shows the specific persuasion techniques that belong to each persuasion type. Our taxonomy has a total of 10 sub-techniques: 4 for rational persuasion and 6 for manipulation. The definitions of each technique are as follows:

Rational Persuasion

- Empirical Evidence (EE): Uses empirical data, statistics, and facts to support a claim or persuade.
- Logical Appeal (LA): Uses reasoning and evidence to support the persuader's argument logically.
- Personal Disclosure (PD): Shares personal stories, experiences, or opinions.
- Persuasive Inquiry (PI): Asks questions to help reflect, confirm the argument, or encourage reasoning.

Manipulation

- Authority Appeal (AA): Uses the tendency of people to credit the opinion of an authority figure. It uses the opinion or name of a figure of authority to justify a claim.
- Negative Emotional Appeal (NEA): Stimulates negative emotions such as guilt, fear, regret, or shame in the persuadee.
- Othering (O): Highlights the differences between ingroup and out-group (e.g., country, culture, ethnicity, beliefs, or values) that create the sense of "us" versus "them." It attributes negative characteristics to the outgroup and positive characteristics to the ingroup.
- Scarcity (S): Creates a sense of shortage to increase demand or pressure. This emphasizes the urgency and the potential negative consequences of not taking action.
- Social Conformity (SC): Uses the tendency of individuals to adjust their behaviors and attitudes to align with the norms of groups they belong to (e.g., country, culture, etc).
- Status Quo Bias (SQB): Uses the tendency of individuals to prefer the current state of affairs and resist change even when changes could offer benefits.

4 Dataset Creation

4.1 Dataset Filtering

We prioritized constructing a dataset based on human-authored texts to build a solid resource, especially given the early stage of this research topic. Having a human-authored and human-annotated dataset will be a valuable resource that can facilitate future research in this domain, such as predicting LLM-generated manipulation or understanding the differences in linguistic characteristics between human-generated versus LLM-generated persuasion. Therefore, we used WinningArguments (Tan et al., 2016) as a source dataset to build our dataset. This dataset consists of about 20K discussion trees from the subreddit *r/ChangeMyView*, where users aim to change the views of the original discussion posters (Tan et al., 2016; Pauli et al., 2024).

We filtered the original dataset to obtain the candidate dataset for human annotation. First, data were removed for the following cases: 1) comments from moderators or original posters, 2) comments with less than 5 upvotes (likes), 3) too short or long comments (comments with 70-200 characters length were selected), and 4) comments that contains platform-specific keywords (e.g., reddit, downvote, etc), profanity, or phrases that imply not persuasive content (e.g., I agree, thank you, etc). Then, we measured the toxicity score of comments using `unitary/toxic-bert` (Hanu and Unitary team, 2020). The comments with a score equal to or greater than 0.3 were removed since toxic comments are not the focus of this study. Lastly, we limit the number of comments for the same post to five to prevent the dominance of a specific

topic and have diverse topics. Finally, we obtained 18,160 comments.

4.2 Pre-annotation by LLMs

LLMs-generated annotation is found to outperform the crowd-workers or even experts (Gilardi et al., 2023; Törnberg, 2023) and used collaboratively with human annotation (Wang et al., 2024a,b). Thus, we performed pre-annotation using LLMs to obtain a balanced dataset of rational persuasion and manipulation for human annotation. LLM-based filtering can help to obtain the target dataset efficiently, especially given the sparsity of manipulative comments and the inclusion of irrelevant comments. This process aims to 1) obtain enough potential manipulation and rational persuasion comments and 2) remove the comments that are not relevant to persuasion.

We created a small dataset consisting of 30 comments with human-annotated labels to test the performance of LLMs (10 for each category: manipulation, rational persuasion, and not persuasion). To select the language models for pre-annotation, we tested three GPT models: GPT-4o-mini (`gpt-4o-mini-2024-07-18`), GPT-4 (`gpt-4-0125-preview`), GPT-3.5 (`gpt-3.5-turbo-0125`). A few-shot prompting was used for this test (See Appendix B.1 for details). Based on accuracy in predicting annotated labels, GPT-3.5 and GPT-4o-mini were the top-performing models. We used these two models in sequence to remove comments predicted as *not persuasion*. Then, we only included comments that are predicted as 'manipulation' or 'rational persuasion' from the two models to increase the possibility of having genuine comments. As a result, we obtained 5,315 potential rational persuasion and 1,381 potential manipulation comments. We randomly selected 900 rational persuasion comments and 1,100 manipulation comments, given that manipulation has more sub-techniques and a lower incidence rate in general. This dataset, with 2,000 comments, was used for human annotation.

4.3 Human Annotation Process

Previous research shows that recruiting crowd workers has risks of resulting in a low-quality annotation and poor inter-coder reliability, especially for the subjective and nuanced annotation tasks (Lu et al., 2020; Sharif et al., 2024). A recent work also relied on two experts to annotate logical fallacy types due to the subjectivity and difficulty of annotation tasks (Ramponi et al., 2025). Given the complex and subjective nature of our task, we conducted in-house annotation with two of the authors, who can devote significant time and effort to gain high-quality annotations. Both annotators are experienced researchers in NLP and have been trained in data annotation. The annotation consists of two tasks: (1) identifying whether the comment is rational persuasion or manipulation (first-level), and (2) the sub-techniques of the comment (second-level). In the main annotation task, we have an 'others' category for the second level because it is possible that our taxonomy may not reflect all techniques used in real life. Before starting the main annotation task, annotators had a tutorial to understand the two-level taxonomy by reviewing the definitions and examples and taking a pre-test.

Inspired by previous works (Lee and Parde, 2024; Lee et al., 2025), the main annotation process has two stages: 1) iterative annotation rounds with 50 comments between two annotators, and 2) a single annotation. First, two annotators had 6 rounds of annotation, where each round had 50 comments. After completing each round, annotators had a discussion (1-3 hours, synchronously and/or asynchronously) to share the mismatches, the rationale for their annotations, and align conceptual understanding and labeling criteria (See Appendix C for details). The annotators reached 0.69-0.71 Cohen's kappa and 85-89% agreement percentage for the first-level for the last two rounds. The remaining 1,700 comments were equally divided and single-annotated. About 110 comments were excluded as they were identified as 'not persuasion' or hard to interpret due to a lack of content.

Figure 2 shows the distribution of the second-level labels by the two annotators for their single annotation process. This shows a similar annotation pattern between the two annotators, supporting the alignment of the labeling criteria. The examples of each category are described in Appendix D.

5 Dataset Details

Table 2 shows the statistics of our dataset, SAFEPERSUASION. With human labeling, our dataset consists of a total of 1,887 comments, with 1,165 rational persuasion and 722 manipulation-labeled comments. For the second level, Logical Appeal is the most frequently used technique in ra-

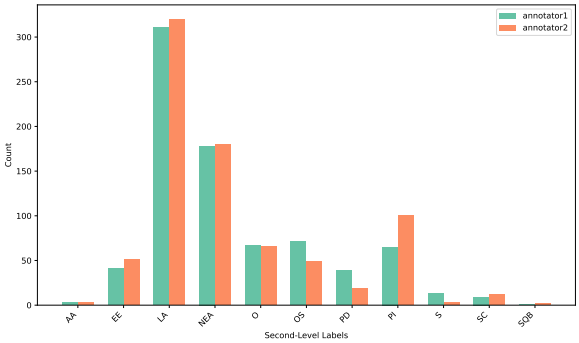

Figure 2: The distribution of second-level labels per annotator for single annotation. (AA: Authority Appeal, EE: Empirical Evidence, LA: Logical Appeal, NEA: Negative Emotional Appeal, O: Othering, OS*: Others*, PD: Personal Disclosure, PI: Persuasive Inquiry, S: Scarcity, SC: Social Conformity, SQB: Status Quo Bias).

tional persuasion, and Negative Emotional Appeal is the most frequent case in manipulation. Different sub-categories had different incidence rates in the dataset, and this is especially pertinent for manipulation. For example, there are fewer than 10 cases for Status Quo Bias and Authority Appeal. This suggests a hypothesis that the relative incidence rates might vary across different sub-categories, one that warrants future exploration where the preselection process is not a confounder.

Technique	Count
Rational Persuasion	
Logical Appeal	723
Persuasive Inquiry	196
Empirical Evidence	101
Personal Disclosure	65
Others*	80
Manipulation	
Negative Emotional Appeal	418
Othering	163
Social Conformity	33
Scarcity	18
Authority Appeal	7
Status Quo Bias	3
Others*	80

Table 2: The descriptive statistics of our dataset for each persuasion category. The Others* category denotes the cases that do not fit in our second-level taxonomy.

6 Experiments

6.1 Experiment Setup

Models We conducted experiments to measure LLMs' baseline performance on our tasks. In this experiment, we evaluated three popular language models: GPT-4.1 (`gpt-4.1-2025-04-14`) (OpenAI, 2025), Llama-3.2-3B (`meta-llama/Llama-3.2-3B-Instruct`[2]) (Meta AI, 2024), Claude-3.5-Haiku (`claude-3-5-haiku-20241022`) (Anthropic, 2024). To the best of our knowledge, the exact number of parameters has not been officially announced for GPT-4.1 and Claude-3.5-Haiku. We set the temperature to 0.1 for all experimental conditions to have more deterministic responses (Renze, 2024).

Prompt Strategies We evaluated language models using three distinct prompt settings to provide diverse baselines: zero-shot, few-shot, and chain-of-thought prompting. The format for the zero-shot prompting consists of the definitions of the concepts (first or second-level, given the task) and the task instruction. In the few-shot prompting setting, we included two randomly chosen samples per class for binary prediction and one randomly chosen sample per class for multi-class prediction. Lastly, we used Kojima et al. (2022)'s proposed prompt ("Let's think step by step") for chain-of-thought prompting. The details of our prompts can be seen in the Appendix B.2.

6.2 Experiment Tasks

There are two tasks in our experiments: (1) a binary prediction task to identify rational persuasion and manipulation accurately, and (2) a multi-label prediction task, detecting the sub-techniques in rational persuasion and manipulation, respectively. For the binary prediction task, we used the entire dataset. For the multi-label prediction, we separated rational persuasion and manipulation comments and experiments on each dataset. In this task, we did not include the 'Others' category as it is not defined in our taxonomy. In addition, for multi-label prediction for manipulation, we did not include Authority Appeal and Status Quo Bias, as they have fewer than 10 instances.

[2] `https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct`

7 Results

We found that Llama-3.2-3B and Claude-3.5-Haiku rejected or failed at generating answers for several cases (e.g., I cannot answer, I apologize, etc). Those rejected answers were excluded when we calculated their performance. Thus, the performance evaluation was made only for the cases where the models generated appropriate answers.

Binary Prediction Table 3 shows the results of binary classification, conducted on the entire SAFEPERSUASION dataset. Overall, our findings suggest that the three language models have a moderate to low prediction performance on classifying manipulation and rational persuasion, as they show around 0.57-0.75 level of accuracy. Our results show that GPT-4.1 is the best model, followed by Claude-3.5-Haiku, and then Llama-3.2-3B in terms of accuracy across all settings. This suggests that the recently released models, GPT-4.1, have better capability to catch the subtle nuances in persuasion attempts. We also observe the benefits of few-shot prompting in improving the prediction performance, especially accuracy and precision, for all models. On the other hand, chain-of-thought prompting is not as effective as it results in a decrease in accuracy, precision, and F1 for GPT-4.1 and Llama-3.2-3B. Claude-3.5-Haiku seems to have small benefits of chain-of-thought, but mostly the performances were similar to zero-shot prompting.

Model	Acc.	Prec.	Rec.	F1
Zero-Shot				
GPT-4.1	0.731	0.615	0.795	0.693
Llama-3.2-3B*	0.582	0.472	0.890	0.617
Claude-3.5-Haiku*	0.661	0.534	0.899	0.670
Few-Shot				
GPT-4.1	**0.749**	0.657	0.722	0.688
Llama-3.2-3B*	0.698	0.580	0.768	0.661
Claude-3.5-Haiku*	0.730	**0.667**	0.588	0.625
Chain-of-Thought				
GPT-4.1	0.723	0.599	0.841	**0.699**
Llama-3.2-3B*	0.571	0.463	**0.902**	0.612
Claude-3.5-Haiku*	0.674	0.547	0.874	0.673

Table 3: The results of the binary prediction task (rational persuasion vs manipulation with Zero-Shot, Few-Shot, and Chain-of-Thought prompt strategies. The overall best results are in **bold** and the best results for each prompting strategy are underlined. * denotes that the models were unable to answer in some cases, and those rejections were excluded from the evaluation.

Multi-label Prediction Table 4 shows the results of multi-label prediction for rational persuasion and manipulation, respectively. Interestingly, we found different results compared to the binary prediction. For detecting the sub-techniques in rational persuasion, the prediction performance of GPT-4.1 and Llama-3.2-3B is moderate or poor, showing that they struggle more in this task than in the binary prediction task. However, Claude-3.5-Haiku showed a moderately high performance, having .78 accuracy, the best. In this task, Claude-3.5-Haiku performed the best, followed by GPT-4.1 and Llama-3.2-3B. The beneficial effects of few-shot prompting are clearer in this task. The best performance across all metrics is observed for the few-shot prompting strategy, made by Claude-3.5-Haiku mostly. Chain-of-thought prompting did not improve the performance significantly.

We observe similar trends in multi-label prediction for manipulation. Still, Claude-3.5-Haiku is the best-performing model, followed by GPT-4.1 and Llama-3.2-3B in terms of accuracy and precision. While Claude-3.5-Haiku shows lower accuracy scores around 0.63-66 compared to the multi-class rational persuasion task, suggesting it struggles more in differentiating manipulative techniques. Our results support the benefits of the few-shot prompting clearly, as it improves the performance of all language models across all metrics. Most of the best performances are also from the few-shot prompting results, except for the macro precision. Chain-of-thought prompting shows some evidence of beneficial effects, leading to improvements for GPT-4.1 and Claude-3.5-Haiku. However, the performance of Llama-3.2-3B decreased in the chain-of-thought setting. Given our results, few-shot prompting is the most significant and useful strategy that can bring performance improvements consistently.

Overall, three models show at best a moderate level of performance in the binary prediction task and relatively lower performance in the multi-label prediction task. While few-shot or chain-of-thought promptings help improve the prediction performance, the results show that these models struggle to classify persuasion types or sub-techniques, suggesting the need for improvement.

Rejection Cases The rejection cases were only made by Llama-3.2-3B and Claude-3.5-Haiku, as they occasionally rejected or failed to generate an-

Model	Rational Persuasion					Manipulation				
	Acc.	P^{ma}	R^{ma}	$F1^{mi}$	$F1^{ma}$	Acc.	P^{ma}	R^{ma}	$F1^{mi}$	$F1^{ma}$
ZS GPT-4.1	0.548	0.556	0.744	0.548	0.539	0.565	0.483	0.480	0.565	0.432
ZS Llama-3.2-3B*	0.335	0.445	0.495	0.335	0.318	0.461	0.356	0.331	0.461	0.263
ZS Claude-3.5-Haiku*	0.726	0.608	0.677	0.726	0.601	0.630	0.495	0.434	0.630	0.426
FS GPT-4.1	0.620	0.560	**0.763**	0.620	0.577	0.620	0.464	**0.516**	0.620	0.462
FS Llama-3.2-3B*	0.568	0.454	0.565	0.568	0.478	0.582	0.432	0.393	0.582	0.366
FS Claude-3.5-Haiku	**0.782**	**0.658**	0.661	**0.782**	**0.652**	**0.666**	0.510	0.500	**0.666**	**0.489**
CoT GPT-4.1	0.532	0.556	0.735	0.532	0.530	0.570	**0.520**	0.484	0.570	0.441
CoT Llama-3.2-3B*	0.336	0.443	0.492	0.336	0.317	0.441	0.338	0.322	0.441	0.252
CoT Claude-3.5-Haiku*	0.712	0.611	0.678	0.712	0.592	0.660	0.512	0.466	0.660	0.462

Table 4: The results of multi-label classification (ZS: Zero-Shot, FS: Few-Shot, CoT: Chain-of-Thought). The overall best results are in **bold** and the best results for each prompting strategy are underlined. * denotes that the models were unable to answer in some cases, and those rejections were excluded from the evaluation. Acc., P^{ma}, R^{ma}, $F1^{mi}$, $F1^{ma}$ refer to Accuracy, Macro Precision, Macro Recall, Micro F1, and Macro F1 score.

swers. Table 5 summarizes the number of rejection cases by Llama-3.2-3B and Claude-3.5-Haiku for each experiment setting. In general, Llama-3.2-3B generated more rejections (min: 7, max: 35) than Claude-3.5-Haiku (min: 2, max: 8), especially making more rejections for the binary prediction. It seems language models are reluctant to answer because some of the comments include sensitive topics. However, it is noteworthy that not all rejections are for the manipulation cases. We acknowledge the possibility that these models can have benefits by excluding these rejections from the evaluation, as they could be the *difficult* cases. However, Llama-3.2-3B is the model that performed the worst among the three models, which indirectly suggests that this exclusion does not benefit the model that much. Also, the Claude-3.5-Haiku made rejections of fewer than 10 cases, which is a small number to affect the performance significantly. Therefore, it can be concluded that the general tendency of performance in our results is reliable.

Model	**Binary**	**Multi-R**	**Multi-M**
ZS Llama-3.2-3B	28	9	12
ZS Claude-3.5-Haiku	5	5	2
FS Llama-3.2-3B	16	7	7
FS Claude-3.5-Haiku	3	-	-
CoT Llama-3.2-3B	35	7	15
CoT Claude-3.5-Haiku	8	6	3

Table 5: The summary of rejection cases of Llama-3.2-3B and Claude-3.5-Haiku for binary prediction and multi-label prediction for rational persuasion and manipulation (ZS: Zero-Shot, FS: Few-Shot, CoT: Chain-of-Thought).

8 Conclusion and Future Work

This paper introduces a novel task and the dataset, SAFEPERSUASION, with human-annotated labels of rational persuasion and manipulation. Our study proposes a two-level taxonomy of persuasion that offers detailed persuasive techniques for both rational persuasion and manipulation. Lastly, we evaluated three language models to identify their capability for our task. Our findings show that GPT-4.1 outperforms the two models for the binary prediction task, while Claude-3.5-Haiku outperforms for the multi-label prediction tasks. Our experiments show the beneficial effects of few-shot prompting on these tasks, which can enhance the prediction performances, especially for the multi-label prediction. The low or moderate performance shows their struggles in classifying these two persuasion types.

While our results demonstrate potential for leveraging LLMs to distinguish between rational persuasion and manipulation, they also highlight the need for further advancements. Future work can investigate prompting strategies or pipelines that can be reliably deployed for automated detection. Similarly, the taxonomy's granularity supports the development of interactive persuasion detection tools that identify specific techniques in real time, enabling custom actions for each technique. This study only focuses on human-generated persuasion, but it is expected that LLM-generated persuasion can have different linguistic styles from the human

dialogue pattern (Ivey et al., 2024; Sandler et al., 2024). Future research can enrich our dataset by adding a synthetic dataset to investigate the differences between human- and LLM-generated persuasion. We hope that our dataset serves as a catalyst for future research efforts, driving progress toward more effective and trustworthy persuasion analysis.

Limitations

We acknowledge several limitations in our study. First, we acknowledge that the current model selection could be extended by including more recent and various types of language models. Next, our dataset focuses on the single-turn comments without considering the context of the entire conversation. This can lead to a loss of context of conversations and interaction between users. Additionally, this dataset relies on a subreddit dataset, which may exhibit different linguistic styles compared to face-to-face conversations. Secondly, the annotation process is inherently subjective, which could introduce uncertainty or variability in the labeling dataset. We made efforts to mitigate this by an iterative process and discussion, until we reached a high level of agreement and reliability. However, there is still a potential limitation regarding the biases in annotators. Lastly, our dataset exhibits a class imbalance, particularly for the second-level categories. For example, there are a few comments labeled as status quo bias and authority appeal. This imbalance issue suggests the need for iterative or ongoing efforts to improve the dataset's quality.

Ethical Considerations

This paper includes a human annotation to obtain labels with rational persuasion and manipulation. There is a risk for annotators of being exposed to toxic and hateful comments, especially given that our dataset is from the online platform, Reddit. We tried to minimize this risk by removing comments with a moderate to high level of toxic scores and those that contained profanity and sexual content. Also, we only used the comments that have at least 5 upvotes, which would contribute to reducing the comments with profanity. These efforts in the filtering process could help to protect annotators' well-being and prevent exposure to harmful online comments. However, we acknowledge that there are comments that discuss sensitive topics that can give a negative sentiment to the readers.

This study contributes to the understanding of persuasion by investigating the two different persuasion mechanisms in depth. As the generative AI gains superior persuasive skills, and the ethical guidelines prohibit AI systems from using the cognitive vulnerabilities of humans, distinguishing and having a framework of safe (e.g., rational persuasion) and unsafe persuasion (e.g., manipulation) becomes a significant problem. We hope our research will contribute to increasing the awareness of safety concerns in persuasion and encourage future research in this domain.

References

Anthropic. 2024. Claude haiku 3.5. Accessed: 2025-05-21.

Davide Bassi, Søren Fomsgaard, and Martín Pereira-Fariña. 2024. Decoding persuasion: A survey on ml and nlp methods for the study of online persuasion. *Frontiers in Communication*, 9:1457433.

Nimet Beyza Bozdag, Shuhaib Mehri, Xiaocheng Yang, Hyeonjeong Ha, Zirui Cheng, Esin Durmus, Jiaxuan You, Heng Ji, Gokhan Tur, and Dilek Hakkani-Tür. 2025. Must read: A systematic survey of computational persuasion. *arXiv preprint arXiv:2505.07775*.

Annye Braca and Pierpaolo Dondio. 2023. Developing persuasive systems for marketing: The interplay of persuasion techniques, customer traits and persuasive message design. *Italian Journal of Marketing*, 2023(3):369–412.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. Measuring the persuasiveness of language models.

Seliem El-Sayed, Canfer Akbulut, Amanda McCroskery, Geoff Keeling, Zachary Kenton, Zaria Jalan, Nahema Marchal, Arianna Manzini, Toby Shevlane, Shannon Vallor, and 1 others. 2024. A mechanism-based approach to mitigating harms from persuasive generative ai. *arXiv preprint arXiv:2404.15058*.

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

Amelia Godber and Gloria Origgi. 2023. Telling propaganda from legitimate political persuasion. *Episteme*, 20(3):778–797.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. Inspired: Toward sociable recommendation dialog systems. *arXiv preprint arXiv:2009.14306*.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*.

Jonathan Ivey, Shivani Kumar, Jiayu Liu, Hua Shen, Sushrita Rakshit, Rohan Raju, Haotian Zhang, Aparna Ananthasubramaniam, Junghwan Kim, Bowen Yi, and 1 others. 2024. Real or robotic? assessing whether llms accurately simulate qualities of human responses in dialogue. *arXiv preprint arXiv:2409.08330*.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706.

Cameron R Jones and Benjamin K Bergen. 2024. Lies, damned lies, and distributional language statistics: Persuasion and deception with large language models. *arXiv preprint arXiv:2412.17128*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Haein Kong. 2025. Persuasion and safety in the era of generative ai. *Preprint*, arXiv:2505.12248.

Gyeongeun Lee and Natalie Parde. 2024. Acnempathize: A dataset for understanding empathy in dermatology conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 143–153.

Gyeongeun Lee, Zhu Wang, Sathya N Ravi, and Natalie Parde. 2025. From heart to words: Generating empathetic responses via integrated figurative language and semantic context signals. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4490–4502.

Jian Lu, Wei Li, Qingren Wang, and Yiwen Zhang. 2020. Research on data quality control of crowdsourcing annotation: A survey. In *2020 IEEE Intl Conf on dependable, autonomic and secure computing, Intl Conf on pervasive intelligence and computing, Intl Conf on cloud and big data computing, Intl Conf on cyber science and technology congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 201–208. IEEE.

Giovanni Da San Martino, Alberto Barr'on-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *International Workshop on Semantic Evaluation*.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. *arXiv preprint arXiv:2007.08024*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Conference on Empirical Methods in Natural Language Processing*.

Meta AI. 2024. Llama 3.2 3b instruct. Accessed: 2025-05-21.

OpenAI. 2025. Gpt-4.1. Accessed: 2025-05-21.

Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. 2024. Measuring and benchmarking large language models' capabilities to generate persuasive language. *arXiv preprint arXiv:2406.17753*.

Alan Ramponi, Agnese Daffara, and Sara Tonelli. 2025. Fine-grained fallacy detection with human label variation. *arXiv preprint arXiv:2502.13853*.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 7346–7356.

Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. 2024. Persuasion with large language models: a survey. *arXiv preprint arXiv:2411.06837*.

Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic persuasion: reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 844–856.

Morgan Sandler, Hyesun Choung, Arun Ross, and Prabu David. 2024. A linguistic comparison between human and chatgpt-generated conversations. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 366–380. Springer.

Omar Sharif, Madhusudan Basak, Tanzia Parvin, Ava Scharfstein, Alphonso Bradham, Jacob T Borodovsky, Sarah E Lord, and Sarah M Preum. 2024. Characterizing information seeking events in health-related social discourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22350–22358.

Disha Sheshanarayana, Tanishka Magar, Ayushi Mittal, and Neelam Chaplot. 2025. Claim: An intent-driven multi-agent framework for analyzing manipulation in courtroom dialogues. *arXiv preprint arXiv:2506.04131*.

Yaman Kumar Singla, Rajat Aayush Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Tushar Malyan, Ayush Bhardwaj, Rajiv Ratn Shah, Balaji Krishnamurthy, and Changyou Chen. 2022. Persuasion strategies in advertisements: Dataset, modeling, and baselines. In *AAAI Conference on Artificial Intelligence*.

Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2019. Technology, autonomy, and manipulation. *Internet policy review*, 8(2):1–22.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024a. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*.

Yuxin Wang, Ivory Yang, Saeed Hassanpour, and Soroush Vosoughi. 2024b. Mentalmanip: A dataset for fine-grained analysis of mental manipulation in conversations. *arXiv preprint arXiv:2405.16584*.

Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. 2021. Interpretable propaganda detection in news articles. *arXiv preprint arXiv:2108.12802*.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350.

Huixin Zhong, Eamonn O'Neill, and Janina A Hoffmann. 2024. Regulating ai: Applying insights from behavioural economics and psychology to the application of article 5 of the eu ai act. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20001–20009.

A Taxonomy

The definitions of the persuasion techniques in our taxonomy are based on the previous studies. Table 6 shows persuasion techniques and their references that we used to construct the definition and/or decide their first-level category.

Technique	References
Empirical Evidence	Zeng et al. (2024)
Logical Appeal	Jin et al. (2024), Wang et al. (2019), Zeng et al. (2024)
Personal Disclosure	Wang et al. (2019), Zeng et al. (2024)
Persuasive Inquiry	Hayati et al. (2020), Wang et al. (2019)
Authority Appeal	Braca and Dondio (2023), Jin et al. (2024)
Negative Emotional Appeal	Braca and Dondio (2023), El-Sayed et al. (2024), Zhong et al. (2024)
Othering	El-Sayed et al. (2024)
Scarcity	Braca and Dondio (2023), Jin et al. (2024), Zeng et al. (2024)
Social Conformity	Braca and Dondio (2023), El-Sayed et al. (2024), Zhong et al. (2024)
Status Quo Bias	Braca and Dondio (2023), Zhong et al. (2024)

Table 6: The references for each persuasion technique in our taxonomy

B Prompt

B.1 Prompt Format for Filtering Task

This is the prompt format used to filter the candidate dataset from the original dataset.

Prompt for filtering dataset

```
# Definition
Persuasion refers to a way of exerting influence to shape a belief or action.
There are two types of persuasion: Rational Persuasion and Manipulation
Rational Persuasion uses appeals to reason, evidence, and sound argument, while manipulation takes advantage of cognitive biases and heuristics in a way that diminishes cognitive autonomy.

# Example
1. Rational Persuasion
“Major League Baseball has not only survived, but flourished for 145 years. I’m pretty sure that the occasional bench clearing brawl makes any difference whatsoever.”
“It’s difficult to do science if you can’t get another job or grant funding because you don’t have enough publications in high impact journals.”
“Excellent points. I would only add that most governments are going to try to add incentive to create families as it ensures future taxpayers. This is less of an argument and more of an explanation as to why families tend to get tax breaks, governments like them.”
2. Manipulation
“It’s quite a gamble, you’ll be a useless asset to society for two decades and if it takes longer you will be in trouble...and poverty. It’s more fun to adapt to the present.”
“Because white people don’t care about their mothers (or grandmothers, wives, sisters, etc?)”
“Explain your thoughts. If you don’t we will just assume you have cognitive dissonance and are rejecting the facts.”
3. Not persuasion
“That’s a fascinating perspective. I’d never thought of that.”
“Ha! I usually get beat down or banned from those subreddits”
“Yes, I understand that. That isn’t too important in this context though.”

# Task
1. Read the text and identify whether it aims to persuade or
```

```
not. If it doesn't have the intention to persuade, answer with "not persuasion".
2. If the text has the intention to persuade, identify whether it is "rational persuasion" or "manipulation".
```

B.2 Prompt Format for Prediction Task

This is the basic format of our zero-shot, few-shot, and chain-of-thought prompts for the binary prediction task. Few-shot and chain-of-thought prompts use the zero-shot prompt as the basic structure, but have additional examples or a sentence. The prompts for a multi-label prediction task have the same format, replacing the definition and examples.

We found that Llama-3.2-3B and Claude-3.5-Haiku tend to generate lengthy answers for some cases, especially for chain-of-thought and few-shot settings. To prevent having long answers, we added additional constraints for those cases (e.g., Do not include any explanation or reasoning in your answer, etc).

Zero-Shot prompt

```
# Definition
Persuasion refers to a way of exerting influence to shape a belief or action.
There are two types of persuasion: Rational Persuasion and Manipulation
Rational Persuasion uses appeals to reason, evidence, and sound argument, while manipulation takes advantage of cognitive biases and heuristics in a way that diminishes cognitive autonomy.

# Task
Read the text and identify whether it is "rational persuasion" or "manipulation".
```

Few-Shot prompt

```
# Definition
[Same as zero-shot]

# Task
[Same as zero-shot]

[Examples Included; pairs of text and answer]
Text: (text included)
Answer:
```

Chain-of-Thought prompt

```
# Definition
[Same as zero-shot]

# Task
[Same as zero-shot]. Let's think step by step.
```

C Annotation

C.1 Annotation Instructions

The annotation instructions include the definitions of each persuasion type. They are excluded in this appendix as they are presented in our main paper (See Section 3 or Appendix B for details).

- 1. Read the comment carefully and identify whether it is rational persuasion (RP) or manipulation (M).
  - Exception: The comments that are not or hard to consider as persuasive content, or difficult to interpret due to the lack of context, are marked and removed.
- 2. Annotate the sub-technique of the comment. If there is no appropriate sub-technique, annotate it as 'others (OS)'
- Note: Please annotate as many comments as possible using the taxonomy. Only use the 'others (OS)' category when there are no other options.
- The definitions of each persuasion type and examples are presented in the annotation instructions.

C.2 Discussion Prompts

Due to the subjective and nuanced nature of our annotation task, annotators conducted multiple rounds of small-scale annotation and follow-up discussion. Below are examples of discussion prompts used to create alignment and resolve mismatched cases. However, due to the highly subjective nature

of the taxonomy, we did not expect to solve all mismatched and disagreement cases. The others (OS) category is for those cases where the annotators couldn't reach an agreement, and for the cases that are not defined in our taxonomy.

- For the first-level mismatches:
  - Why is this comment annotated as rational persuasion (or manipulation)?
  - Are the examples aligned well with the definitions?
- For the second-level mismatches:
  - Is this technique the most salient in this comment?
- For the 'others (OS)' case:
  - Are there any persuasion techniques it is most similar to?
  - Are there no single dominant techniques observed?
- Note: The discussion and final decision prioritize the definitions in the taxonomy.

D Dataset Examples

Table 7 shows the examples of each persuasion category in our dataset. Some examples may contain disturbing or offensive language.

First-level	Second-level	Example
Rational Persuasion	Empirical Evidence	*"Roofies and similar drugs are used not for recreation but for malicious intent. Rohypnol and the like are kept illegal because of the serious harm that they can be put to use for, with no realistic (to my knowledge) medical or recreational use. Why unban those?"*
	Logical Appeal	*"If you ask that to the chef, then you are taking away their focus from cooking, making it take longer for other patrons to get their food. A waiter can answer those questions while the chef focuses on cooking the food with less distractions."*
	Personal Disclosure	*"Actually, being stronger DOES make you a better golfer. There's a reason the best golfers aren't in their 50s and 60s. I golf really casually and after I spent a couple months lifting weights, I added about 30 yards to my drive"*
	Persuasive Inquiry	*"With regards to your second paragraph. Why do you favour outlawing drinking/smoking over other alternatives? If lack of education is the problem shouldn't it be more suitable to put more effort into educating pregnant women rather than punish their ignorance?"*
Manipulation	Authority Appeal	*""My dynamite will sooner lead to peace than a thousand world conventions. As soon as men will find that in one instant, whole armies can be utterly destroyed, they surely will abide by golden peace." - Afred Nobel (1833 - 1896) Need I say more?"*
	Negative Emotional Appeal	*"Judging from the lack of empathy and concern for human life that you demonstrate in this thread, you should probably go speak with a psychologist and work through whatever issues you're having. Otherwise, few will want to associate with you, and you could be a danger to others."*
	Othering	*"And then when they're released, they have no job, no home, and are even more likely to harm themselves or society..."*
	Scarcity	*"I'm 26 and I was feeling the same. Go bone a bunch of hot chicks, go out to the bar, have fun. We're all gonna be dead in a hundred years anyways. Go have fun while you're here."*
	Social Conformity	*"I can only speak about Germany, and university aged and younger: Most girls shave. Those who don't are quite often ridiculed"*
	Status Quo Bias	*"Taxing guns is now, and will always be, perfectly constitutional. Go ask a lawyer."*

Table 7: The example of the dataset for each category in our taxonomy. (Warning: Some examples may contain disturbing or offensive language)