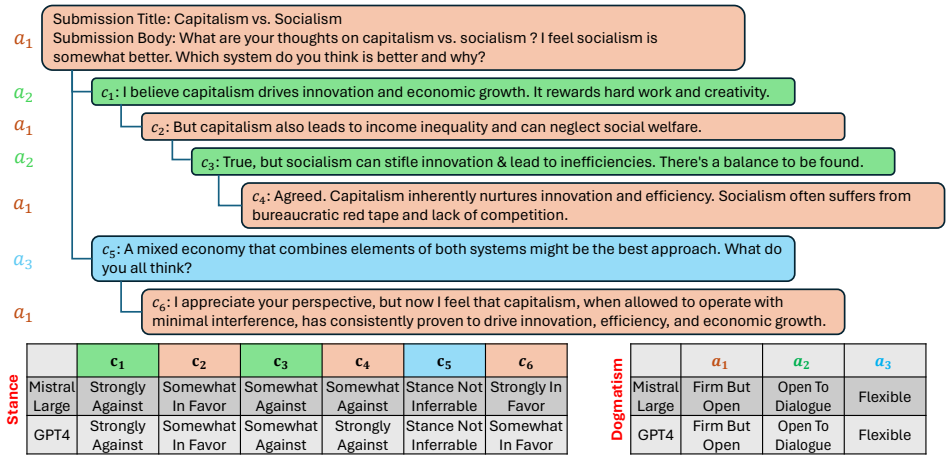
USDC: A Dataset of User Stance and Dogmatism in Long Conversations

Anonymous Author(s)
Affiliation
Address
email



Stance

	c_1	c_2	c_3	c_4	c_5	c_6
Mistral Large	Strongly Against	Somewhat In Favor	Somewhat Against	Somewhat Against	Stance Not Inferrable	Strongly In Favor
GPT4	Strongly Against	Somewhat In Favor	Somewhat Against	Strongly Against	Stance Not Inferrable	Somewhat In Favor

Dogmatism

	a_1	a_2	a_3
Mistral Large	Firm But Open	Open To Dialogue	Flexible
GPT4	Firm But Open	Open To Dialogue	Flexible

Figure 1: Sample Reddit conversation on "Capitalism vs. Socialism" with Stance (for every comment $\{c_i\}_{i=1}^{6}$) and Dogmatism (for every author $\{a_j\}_{j=1}^{3}$) labels from Mistral Large and GPT-4. The submission content favors to socialism and examines how the authors position their opinions regarding socialism vs. capitalism.

Abstract

Identifying user's opinions and stances in long conversation threads on various topics can be extremely critical for enhanced personalization, market research, political campaigns, customer service, conflict resolution, targeted advertising and content moderation. Hence, training language models to automate this task is critical. However, to train such models, gathering manual annotations has multiple challenges: 1) It is time-consuming and costly; 2) Conversation threads could be very long, increasing chances of noisy annotations; and 3) Interpreting instances where a user changes their opinion within a conversation is difficult because often such transitions are subtle and not expressed explicitly. Inspired by the recent success of large language models (LLMs) for complex natural language processing (NLP) tasks, we leverage Mistral Large and GPT-4 to automate the human annotation process on the following two tasks while also providing reasoning: i) User Stance classification, which involves labeling a user's Stance of a post in a conversation on a five-point scale; ii) User Dogmatism classification, which deals with labeling a user's overall opinion in the conversation on a four-point scale. The majority voting on zero-shot, one-shot, and few-shot annotations from these two LLMs on 764 multi-user Reddit conversations helps us curate the USDC dataset. USDC is then used to finetune and instruction-tune multiple deployable small

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

language models for the 5-class stance and 4-class dogmatism classification tasks. We make the code and dataset publicly available [1].

.

1 Introduction

Understanding the user's (or author's) opinion in a conversation is a fundamental aspect of successful interpersonal interactions, and it is essential for developing better interpersonal communication skills, empathy development, and informed decision-making. This user understanding is particularly relevant in the context of dogmatism, a phenomenon observed in various areas such as politics, religion, culture, intellect, and science, where rigid adherence to beliefs often hinders open-mindedness and empathy (Rokeach, 1954). Advertisers can target their campaigns more effectively by aligning with the opinions and stances of potential customers. Companies can use this information for market research to tailor products and services to meet consumer needs and preferences. Political groups can gauge public reaction to policies and campaigns and adjust their strategies accordingly. Identifying differing opinions can help conflict resolution by understanding the perspectives of all parties' perspectives. Society can promote tolerance and maintain social harmony by recognizing and respecting diverse opinions.

Fig. 1 shows a sample Reddit conversation on the topic of *Capitalism vs. Socialism*. We refer to an author's initial post (containing title and body) as a submission. Multiple authors can then share their opinions as comments on the submission. Specifically this example contains 6 comments $\{c_i\}_{i=1}^{6}$ from 3 authors $\{a_j\}_{j=1}^{3}$. We also show stance and dogmatism predictions from two large language models (LLMs): Mistral Large and GPT-4. Some authors like a_1 change their views during the discussion based on the beliefs or opinions of others. At the beginning of the dialogue, we note that author a_1 is somewhat favoring socialism (in submission and c_2). But the author shifts their stance to somewhat favors capitalism (in c_4) after considering the viewpoints of author a_2 in comments c_1 and c_3, illustrating author a_1's firm yet open-minded approach. On the other hand, author a_3 seems very flexible based on their comment c_5. Understanding conversations requires understanding the fine-grained topics being discussed and the dynamic viewpoints of the individual users.

Given the importance of understanding these user dynamics in conversations, training language models to perform this task automatically at scale is critical. While numerous datasets are available for analyzing individual user posts (Fast & Horvitz, 2016; Sakketou et al., 2022), typically through random subsampling of posts or selecting posts with a limited number of tokens, the exploration of a specific user's opinion across each post within an entire conversational thread remains under-explored.

Crowdsourcing is one possible approach to address the need for a suitable dataset. However, a significant limitation in manually annotating datasets for user opinions is the time-consuming nature of the process, as annotators must read entire conversations to label each user's post, making data acquisition costly. Additionally, manual annotation often faces challenges related to quality, as accurately labeling opinions requires understanding demographic details and domain-specific knowledge. Given these limitations, achieving a comprehensive and accurate set of user opinions corresponding to posts about a topic often requires multiple annotators or iterative rounds of annotation. Since users could change their opinion (often times with subtle transitions and not with explicit statements) within a conversation, tracking such changes across multiple users manually becomes very cumbersome.

Recently, large language models (LLMs), especially those built on Transformer architectures (Vaswani et al., 2017) and pretrained on large datasets, have resulted in state-of-the-art accuracies on several complex natural language processing (NLP) tasks (Brown et al., 2020; Chung et al., 2024). LLMs are also being frequently used for dialog response generation (Zhang et al., 2020; Bao et al., 2019; Roller et al., 2021; Adiwardana et al., 2020). Given the complex and cumbersome nature of conversation understanding, we hypothesize that LLMs can be effective in capturing nuances involved in understanding user opinions and their shifts in multi-user conversational contexts. Also, since these models possess long-range memory capabilities, we believe that they can reason over extended conversational threads involving numerous participants, as good as human annotators, if not better.

In this work, we leverage LLMs like Mistral Large and GPT-4 to perform two tasks: i) User Stance classification, which involves labeling a user's stance of a post in a conversation on a five-point

[1] `https://anonymous.4open.science/r/USDC-0F7F`

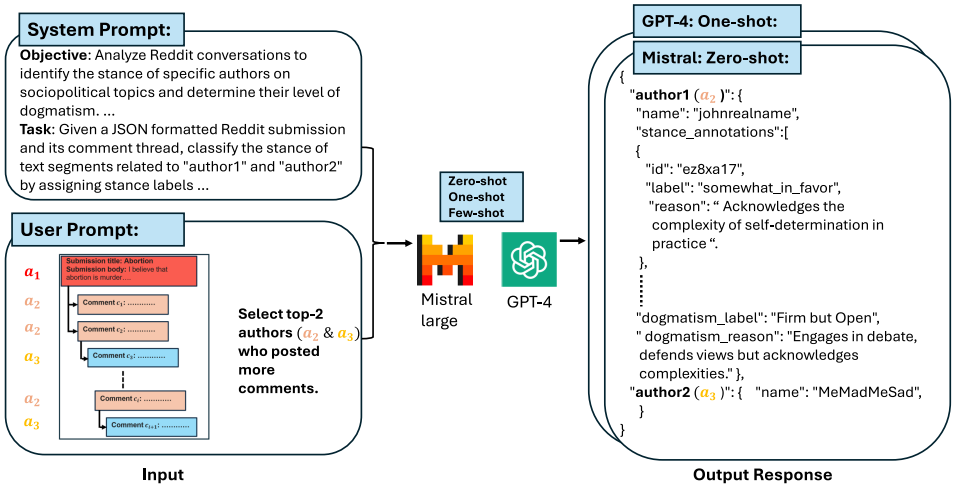


Figure 2: Generating annotations using LLMs: We pass the entire conversation for each Reddit thread in JSON format. The JSON highlights the top two authors who posted the most comments, alongside annotation guidelines for stance and dogmatism labels in the system prompt.

scale; ii) User Dogmatism classification, which deals with labeling a user's overall opinion in the conversation on a four-point scale. Besides the class labels, we also obtain reasoning behind these labels from these LLMs. We experiment with these two models as human-like annotators to generate user opinions in full-length, multi-user Reddit conversations in a zero-shot, one-shot as well as few-shot setup. Thus, overall for every sample, we obtain six annotations ({Mistral Large, GPT-4}$\times${zero-shot, one-shot, few-shot}). Fig. 2 presents our LLM-based annotation pipeline for user-level Stance and Dogmatism tasks. We consider majority voting over these six as our final annotations. Overall, this helps us curate our USDC (a dataset of user stance and dogmatism in conversations) dataset, which consists of 764 multi-user conversations from 22 subreddits, including 1,528 user-level dogmatism samples and 9,618 stance samples across all posts from selected users. Overall, the annotations in the dataset highlight specific user opinions in each post related to stance, track opinion fluctuations leading to a dogmatic nature, and provide reasoning about why users hold specific opinions.

USDC addresses several weaknesses of existing post-level stance and dogmatism datasets. First, the full-length multi-user conversation aspect of USDC enables it to capture contextual and opinion shifts of multiple users. This feature allows it to serve as both an instruction-tuning user opinion dataset and an evaluation benchmark. We believe that the ability to perform instruction tuning for user opinions at a large scale can bridge the gap between open-source and commercial user trait understanding models. Additionally, the in-context learning annotations using state-of-the-art LLMs in USDC make it a more comprehensive measure of how current LLMs understand complex tasks like capturing opinions. This aspect makes it a valuable resource, especially for social media agents seeking deeper insights into user behavior.

In this work, we utilize our USDC dataset to finetune as well as instruction-tune open-source LLMs for generating stance and dogmatism labels for users. We experiment with three pretrained small language models (SLMs) like LLaMA-2-7B, LLaMA-3-8B, and Falcon-7B. We also experiment with four instruction-tuned SLMs like LLaMA-2-chat-7B, LLaMA-3-8B-instruct, Vicuna-7B-v.1.5, and Falcon-7B-instruct. We report weighted F1 scores obtained using these models for both the tasks.

We make the following contributions: 1) We contribute USDC (a dataset of user stance and dogmatism in conversations) dataset consisting of 764 multi-user conversations labeled with 1,528 user-level dogmatism samples and 9,618 stance samples. 2) We report initial results for the stance and dogmatism detection tasks using seven small language models for the UDSC dataset. We find that stance detection performance improves with instruction-tuning (F1-score of 56.2) compared to fine-tuning (F1-score of 54.9). However, dogmatism detection performs worse with instruction-tuning (F1-score of 49.2) compared to fine-tuning (F1-score of 51.4), highlighting the complexity of this task. 3) We make the code and dataset publicly available[1]. Also, the finetuned and instruction-tuned models are made available as well.

2 Related Work

Opinion fluctuations in user conversations. Our work is closely related to previous studies (Fast & Horvitz, 2016; Sakketou et al., 2022), which explore Stance and Dogmatism at the post level, where posts are randomly sampled from conversation threads. Fast & Horvitz (2016) predicted user dogmatism on randomly sampled Reddit posts from conversations, with each post limited to 200-300 characters. One major limitation of this work is the unavailability of a public dataset and missing annotator demographic details. Sakketou et al. (2022) created the post-level Stance dataset, SPINOS, where each post is considered independent, and submission posts are missing while annotators label the data. Additionally, the quality of the dataset is not validated due to missing demographic details of these annotators. Our work overcomes the limitations of previous studies and presents Stance detection for posts and Dogmatism labels of users in conversations, considering the entire context, while preserving submission IDs. Hence, our dataset provides clear user-level posts and dogmatism data, which are useful for modeling dynamic user representations.

Generating annotations for NLP tasks using Large Language Models Our work also relates to a growing body of literature suggesting that large language models can perform similarly to human annotators in labeling complex NLP tasks (Zhou et al., 2022; Zhang et al., 2023; Bansal & Sharma, 2023; Lowmanstone et al., 2023; Wadhwa et al., 2023; Honovich et al., 2023; Zheng et al., 2024; Ye et al., 2022a; Meng et al., 2022). Several studies have explored LLM-based annotation generation in zero-shot or few-shot task settings (Ye et al., 2022a; Meng et al., 2022; Ye et al., 2022b), while others have compared pairs of language models to assess the quality of annotations generated by these LLMs (Zheng et al., 2024). However, these studies focused on generating annotations for NLP tasks such as sentiment analysis, natural language inference (Gilardi et al., 2023; Alizadeh et al., 2023), or creating synthetic dialogues, but only for dyadic conversations (Lee et al., 2023). Our approach complements these previous studies by focusing on generating annotations of user opinions in complex multi-user conversations.

3 USDC Dataset Curation

In this section, we will discuss three main things: 1) Collection of Reddit conversations, 2) Obtaining LLM annotations, and 3) Inter-annotator agreement with LLMs as annotators.

3.1 Collection of Reddit Conversation Threads

Initial crawl. We crawl an year (2022) worth of multi-user conversation data from 22 subreddits of Reddit [2] using praw API [3]. This dataset includes submissions and all associated user comments. Each submission, which serves as the initial message of the conversation, contains a title and content body. This is followed by comments and replies to the submission or other comments. Overall, we crawled 3,619 Reddit conversations across the 22 subreddits. A sample Reddit conversation is displayed in Fig. 1.

Quality filtering of conversations. Since submission content on Reddit can sometimes include videos, we perform the following filtering steps. 1) We only consider submissions where the content is text. 2) We remove conversations with [deleted] tags and empty content. 3) We exclude conversations where the posts were either discarded by users or removed by moderators.

Reddit user conversations can be very long and we observed up to 591 comments in a single crawled conversation data. Considering the maximum sequence length allowed by various language models, we retained only those conversations that contain at least 20 and at most 70 comments. Considering conversations with fewer than 20 comments results in too few comments to accurately gauge user opinions based on small samples. Further, we ensure that at least two users covering $\sim$50% of the comments in the conversations. We did not remove any comments or reduce the post length in the selected conversations. Out of the initial 3,619 conversations, these filtering steps result into 764 conversations getting selected. Table. 4 in the Appendix shows detailed subreddit level statistics.

[2] https://www.reddit.com/

[3] https://github.com/praw-dev/praw

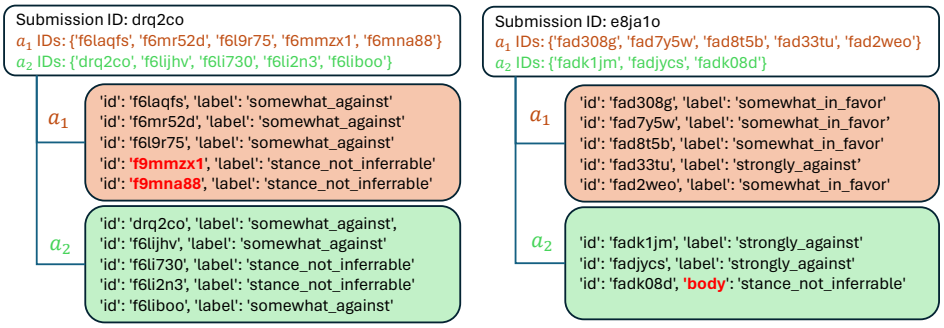


Figure 3: Failure cases of LLMs: Mistral Large few-shot output (left), the ids ("f6mmzx1","f6mna88") were mismatched with generated ids ("f9mmzx1","f9mna88"), GPT-4 zero-shot output (right), the key **"label"** was mismatched with generated key "body".

3.2 Obtaining LLM Annotations

Representing Reddit conversations in JSON format.

To create the prompt, we follow the nested hierarchical structure of Reddit conversations to maintain the context. Specifically, we maintain a JSON structure for each conversation, where each author has their post IDs, and comments or replies are available in the body section. An example of a Reddit conversation in JSON format is provided in Appendix D. Note that the JSON explicitly includes the top-2 authors who posted the most comments in the conversation, as well as their respective post IDs. Our emphasis on these top-2 users (covering 47% posts of total posts on average) aimed at accurately assigning Stance and Dogmatism labels, acknowledging the challenge of modeling a user's opinion belief based on a very number of posts within a conversation.

Using LLMs as human-like annotators. To annotate the position (or Stance) of a user towards a subreddit topic at each post and the overall opinion (or Dogmatism level) of a user in a conversation, we employ two well-known commercialized API-based LLMs: GPT-4 (OpenAI, 2023) and Mistral Large (Jiang et al., 2024). OpenAI GPT-4 is a decoder-based language model which features a context window of 32k to 128k tokens. Mistral Large features a context window of 32k tokens. Additionally, we also examined other versions of these models, such as GPT-3.5 and Mistral-small and medium, but found that these models failed to produce annotations in the desired format. We briefly discuss these limitations in Section 6.

For both GPT-4 and Mistral Large, we supplied a system prompt that contains the definition of Stance and Dogmatism, guidelines for annotating each user conversation, and the necessary labels for Stance and Dogmatism, as shown in Fig 2. The system prompt is detailed in the Appendix B. Along with the system prompt, we provided a user prompt comprising the entire user conversation in a structured JSON format, as discussed above. Additionally, we prompted the model to generate reasoning for each label, explaining why the LLMs assigned a particular label to a specific user post. We used zero-shot, one-shot, and few-shot settings to get the LLM-based annotations. For the few-shot setting, we added two examples in the prompt. Samples of generated outputs using GPT-4 in zero-shot, one-shot, and few-shot settings are shown in Appendix E.1, E.2, E.3 respectively. Similarly, samples of generated outputs using Mistral Large in zero-shot, one-shot, and few-shot settings are shown in Appendix E.4, E.5, E.6 respectively.

Annotation tasks. We prompt the LLMs to perform two annotation tasks: 1) Stance detection, which determines if a user comment or post is *Strongly In Favor*, *Strongly Against*, *Stance Not Inferrable*, *Somewhat In Favor*, or *Somewhat Against* towards specific subreddit submission content; 2) Dogmatism identification, which evaluates the user's overall opinion in conversation and categorizes them into one of four categories: *Firm but Open*, *Open to Dialogue*, *Flexible* or *Deeply Rooted*. This assessment reveals whether a user is open to changing their beliefs or remains steadfast in their opinions based on interactions with other users.

Addressing LLM response and JSON parsing failures. Sometimes the LLMs got confused with the author IDs and missed Stance labels for some author IDs (Fig. 3 (left)). Sometimes, there were minor errors in key naming ('label' vs 'body' in Fig. 3 (right)). For each LLM setting, we observed

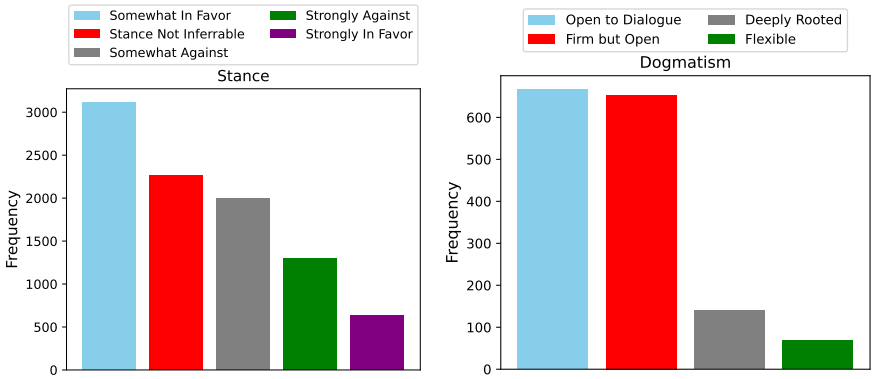


Figure 4: Distribution of class labels for Stance (left) and Dogmatism (right) tasks. These class labels are determined by majority voting across GPT-4 and Mistral Large models.

such errors in around 15 cases on average. We manually fixed such JSON parse errors and missing Stance labels for some author IDs.

Majority voting. After obtaining six annotations ({Mistral Large, GPT-4}$\times${zero-shot, one-shot, few-shot}) for each sample, we aggregate using majority voting to determine the final gold annotations for the Stance and Dogmatism tasks. Fig. 4 presents the class distributions for both the annotation tasks. Additionally, we present the class distributions obtained from each model with the three settings (zero-shot, one-shot and few-shot) for two tasks in Appendix Figs. 5 and 6 respectively.

3.3 Inter-annotator Agreement with LLMs as Annotators

As the quality of labeling on subjective tasks is challenging, we validated the inter-annotator agreement (IAA) between the six LLMs (GPT-4 Zero-shot, GPT-4 One-shot, GPT-4 Few-shot, Mistral Large Zero-shot, Mistral Large One-shot, and Mistral Large Few-shot) for the Stance as well as Dogmatism tasks. We perform IAA using two approaches: i) Cohen's kappa score (Cohen, 1960) and ii) Fleiss' kappa score (Fleiss, 1971). Cohen's kappa measures the agreement between two raters, while Fleiss' kappa extends this to multiple raters. Hence, we employed Cohen's kappa for pairwise comparisons and Fleiss' kappa for overall agreement across all models.

Fig. 7 in Appendix shows the pairwise Cohen's kappa values for both Stance and Dogmatism tasks. We observe that Cohen's kappa values range from 0.36 to 0.72 for Stance and 0.31 to 0.61 for dogmatism, indicating moderate agreement between the models. Broadly kappa values are higher for model pairs within a family (GPT-4 or Mistral large). Thus, the large variance in the kappa scores is not due to the various in-context learning settings (ZS, OS, FS) but rather due to architectural differences.

The overall Fleiss' kappa value was calculated as 0.485 for Stance and 0.435 for Dogmatism, suggesting moderate agreement among all six models. Comparing LLM IAA with previous studies, we observe that for dogmatism, the LLM IAA of 0.435 matches with 0.44 as mentioned in Fast & Horvitz (2016). Similarly, for Stance, the LLM IAA of 0.485 is much higher than 0.34 as reported in Sakketou et al. (2022). It is important to note that previous studies on Stance and Dogmatism datasets were created on post-level data with limited token lengths, whereas our work focuses on entire user conversations. This suggests that LLMs can be considered as competent annotators for complex subjective tasks. However, the moderate agreement levels indicate potential areas for improvement and align with the observed performance variations among the models.

4 Training Small Language Models

In this section, we briefly discuss the small language models that we experiment with. We also discuss their finetuning and instruction tuning details.

4.1 Small Language Models

we train three pretrained small language models (LLaMA-2-7B, LLaMA-3-8B, Falcon-7B) and four instruction-tuned small language models (LLaMA-2-chat-7B, LLaMA-3-8B-instruct, Vicuna-7B-v.1.5, and Falcon-7B-instruct). We finetune as well as instruction tune these models using the proposed USDC dataset. We use pretrained models checkpoints from Hugging Face. All of these LLMs have context length of 4096 tokens.

LLaMA models (Touvron et al., 2023a) are decoder-only LLMs trained on 1.6 trillion tokens from a mixture of corpora including C4, English CommonCrawl, Wikipedia, Github, and more. We use two versions of models in our study: LLaMa-2-7B (Touvron et al., 2023b) and LLaMa-3-8B and their instruction tuned variants.

Falcon models (Almazrouei et al., 2023) are decoder-only LLMs trained on $\geq$ 1 trillion tokens of text, with a particular emphasis on the RefinedWeb corpus. For Falcon, we use both the pretrained and instruction tuned 7B parameter variants in our study.

Vicuna model (Chiang et al., 2023) is finetuned from the LLaMA 7B model on approximately 70K user-shared conversations gathered from ShareGPT.com and we used the 7B parameter variants.

4.2 Experimental Setup

Train-test setup. We conducted both finetuning and instruction-tuning of small language models. For this purpose, we divided the dataset of 764 conversations into train ($\sim$ 75%) and test splits ($\sim$ 25%). The training dataset comprised 564 conversations, including 1128 samples of Dogmatism labels and 7520 samples of Stance labels. Conversely, the testing dataset consisted of 200 conversations, with 400 samples of Dogmatism labels and 1831 samples of Stance labels across two authors posts.

Implementation details for reproducibility. All experiments were conducted on a machine equipped with an NVIDIA A100 GPU with 80 GB of GPU RAM, partitioned into two devices of 40 GB each. We employed 4-bit quantization with normalized floating precision (nf4) from the bitsandbytes library [4]. Additionally, we utilized LoRA (Hu et al., 2021) with a rank of 64 and an alpha value of 16 during task-based instruction tuning. Finally, we use PEFT (Parameter Efficient Finetuning) [5] library to train large language models with SFTT (Supervised Finetuning Trainer) setting. To further enhance performance, we divided the training dataset into a validation set, comprising a randomly chosen 10% subset from the training set, used exclusively for hyperparameter tuning. More details about bitsandbytes, PEFT and SFTT parameters are reported in Appendix.

4.3 Finetuning and Instruction Tuning of Small Language Models (SLMs)

Finetuning of SLMs. For Stance classification, we treat each user post as an independent sample. In contrast, for Dogmatism classification, we consider the entire user conversation as a single sample by concatenating all the threads from a user in that conversation. To load the pretrained SLMs, we perform 4-bit quantization, apply the LoRA technique (Hu et al., 2021), and fine-tune the models with SFTT before saving the fine-tuned model. For finetuning, we used prompt for Stance classification as shown in Fig. 8 (see Appendix). Similarly, Fig. 9 (see Appendix) displays prompt for Dogmatism identification.

Instruction tuning of SLMs. We instruction tune the SLMs on user conversations along with their gold labels from the training part of the USDC dataset. For instruction tuning, we use the same prompt as used for LLMs to generate the USDC dataset (also shown in Appendix B). Similar to finetuning, we use same train-test splits for instruction tuning.

5 Results

Do SLMs finetuned with task-specific LLM annotations accurately perform Stance and Dogmatism tasks on user opinions? We show the weighted F1 of various SLMs finetuned with task-specific LLM annotations on the stance and dogmatism detection tasks on the USDC test set in Table 1. We

[4] `https://pypi.org/project/bitsandbytes/`

[5] `https://github.com/huggingface/peft`

Table 1: Finetuning: weighted F1 score for Stance classification using SLMs on USDC test set. ZS: Zero-shot, OS: One-shot, FS: Few-shot.

Model	Stance Classification							Dogmatism Classification						
	GPT-4			Mistral Large			Majority	GPT-4			Mistral Large			Majority
	ZS	OS	FS	ZS	OS	FS		ZS	OS	FS	ZS	OS	FS	
LLaMA-2-7B	51.8	52.9	52.7	35.1	49.2	46.0	54.0	42.1	44.2	45.2	39.3	47.6	43.7	43.4
LLaMA-2-chat-7B	52.8	51.4	51.8	34.7	47.5	46.5	51.3	42.1	42.5	48.8	41.1	49.7	45.5	48.3
LLaMA-3-8B	51.3	52.2	52.9	34.9	48.5	47.0	**54.9**	42.0	47.8	45.3	39.9	47.4	36.3	**51.4**
LLaMA-3-8B-instruct	51.2	52.6	52.7	33.9	49.5	45.6	54.5	44.8	46.2	49.7	46.1	45.8	46.1	50.8
Falcon-7B	50.7	51.1	51.6	34.9	47.2	43.9	53.2	41.5	42.1	43.3	36.5	38.4	37.5	40.1
Falcon-7B-instruct	51.2	51.5	51.6	35.1	47.7	44.2	51.0	41.7	42.1	42.9	36.8	38.5	36.9	39.7
Vicuna-7B-v.1.5	51.0	53.0	53.2	35.1	48.5	45.8	54.7	42.9	48.3	40.8	45.9	42.6	46.2	42.3

Table 2: Instruction-tuning: weighted F1 score for Stance classification using SLMs on USDC test set. ZS: Zero-shot, OS: One-shot, FS: Few-shot.

Model	Stance Classification							Dogmatism Classification						
	GPT-4			Mistral Large			Majority	GPT-4			Mistral Large			Majority
	ZS	OS	FS	ZS	OS	FS		ZS	OS	FS	ZS	OS	FS	
LLaMA-2-7B	53.2	54.0	54.5	36.8	50.3	47.2	55.5	43.0	45.0	46.3	40.6	48.2	45.0	44.0
LLaMA-2-chat-7B	54.0	54.5	55.0	36.5	50.7	47.6	54.0	43.2	45.5	47.0	40.8	48.5	45.5	43.8
LLaMA-3-8B	53.5	54.8	55.5	37.0	50.5	48.0	**56.2**	43.5	46.0	47.5	41.0	48.8	45.8	45.1
LLaMA-3-8B-instruct	53.0	54.2	55.0	36.0	50.0	47.0	55.5	43.8	46.5	47.8	41.5	**49.2**	46.0	44.8
Falcon-7B	52.8	53.4	54.0	36.5	49.5	46.5	54.8	42.5	44.6	45.8	39.8	47.0	44.0	43.8
Falcon-7B-instruct	53.0	53.8	54.2	36.8	49.8	46.8	54.5	42.8	44.8	46.0	40.0	47.2	44.2	43.0
Vicuna-7B-v.1.5	53.3	54.5	55.2	37.0	50.2	47.8	55.2	43.7	46.8	47.2	41.2	48.2	46.5	44.8

report AUC scores and other qualitative analysis in Appendix F (Fig. 11 and 12). We make the following observations from these results: 1) For both tasks, the majority voting labels as ground truth, has a relatively high performance, scoring above 50% weighted F1-score across several models. 2) LLaMa-3 models (LLaMA-3-8B and LLaMA-3-8B-instruct) perform better across both the tasks. 3) For GPT-4 annotations, in most cases, SLMs finetuned with few-shot annotations outperform those trained with zero and one-shot annotations. For Mistral Large annotations, typically SLMs finetuned with one-shot annotations performs the best. 4) Specifically, for Stance detection task, Vicuna-7B-v.1.5 finetuned using few-shot annotations is the best model trained with GPT-4 annotations. Similarly, LLaMA-3-8B-instruct finetuned with one-shot annotations is the best model trained with Mistral Large annotations. 5) For the Dogmatism detection task, LLaMA-3-8B-instruct finetuned using few-shot annotations is the best model trained with GPT-4 annotations. Similarly, LLaMA-2-chat-7B finetuned with one-shot annotations is the best model trained with Mistral Large annotations. 6) Overall, we observe that instruction tuned SLMs perform better than the pretrained SLMs.

Do SLMs instruction-tuned with task-specific LLM annotations perform better than SLMs finetuned with task-specific LLM annotations for the Stance and Dogmatism tasks? We show the weighted F1 of various SLMs instruction-tuned with task-specific LLM annotations, on the stance and dogmatism detection tasks on the USDC test set in Table 2. We report AUC scores and other qualitative analysis in Appendix F (see Fig. 13). We make the following observations from these results: 1) SLMs with instruction-tuning result in higher weighted F1-scores than SLMs with finetuning for stance detection, while SLMs with finetuning outperform SLMs with instruction-tuning in dogmatism detection. 2) Contrary to results in Table 1, Table 2 demonstrates that using majority voting labels as ground truth, SLM instruction-tuning yields relatively high performance only for the stance detection task, but not for the dogmatism detection. 3) Similar to results in Table 1, LLaMA-3 models (LLaMA-3-8B and LLaMA-3-8B-instruct) perform better across both tasks. Additionally, GPT-4 annotations yield the best results in the few-shot setting, while Mistral Large annotations perform best in the one-shot setting.

Overall, we draw the following conclusions when comparing SLM finetuning and instruction-tuning: (1) Since dogmatism detection is inherently a more complex and varied task than stance detection, the model might struggle to generalize from the instructional data. (2) The system prompt used in finetuning is much simpler than the original system prompt for instruction-tuning, making it challenging to handle the context length for longer conversations. We perform an error analysis to further analyze the results in the next subsection.

Error Analysis Table 3 illustrates the confusion matrix for stance detection for LLaMa-3-8B finetuning and instruction-tuning. We make the following observations this table: 1) For both finetuning and instruction-tuning, there is a significant misclassification between "Somewhat Against" and "Somewhat In Favor," as well as between "Somewhat In Favor" and "Stance Not Inferrable." These overlaps suggest challenges in distinguishing moderate stances, indicating a need for enhanced

		Predicted				
		SOA	SOIF	SNI	SGA	SIF
Actual	SOA	151	132	34	44	2
	SOIF	93	537	113	17	14
	SNI	23	78	259	5	0
	SGA	52	35	13	115	17
	SIF	18	50	12	25	27

		Predicted				
		SOA	SOIF	SNI	SGA	SIF
Actual	SOA	143	125	37	54	4
	SOIF	82	543	106	27	16
	SNI	22	82	253	6	2
	SGA	41	35	11	131	14
	SIF	16	53	10	23	30

Table 3: Confusion matrix for LLaMa-3-8B Stance detection models on USDC test set: finetuning (left) and instruction-tuning (right). SOA: Somewhat Against, SOIF: Somewhat In Favor, SNI: Stance Not Inferrable, SGA: Strongly Against, SIF: Strongly In Favor.

feature representation and clearer class definitions to improve model performance. We report the confusion matrix for dogmatism detection task in Appendix Fig. 10. Fig. 10 shows significant misclassifications, especially for the "Deeply Rooted" and "Flexible" labels, with both having zero accuracy and F1-scores. On the other hand, the model performs moderately better for "Firm but Open" and "Open to Dialogue" classes with accuracies of 48.7% and 64.4% respectively. The confusion matrix also indicates substantial confusion to distinguish between intermediate levels of dogmatism, such as "Firm but Open" and "Open to Dialogue". The area under the ROC curve (AUC) is a measure of the model's ability to distinguish between classes. Hence, we further report the ROC curve which shows the trade-off between the true positive rate (TPR) and false positive rate (FPR) for each class for stance and dogmatism tasks, see Figs. 11 and. 12 in Appendix F.

Verification using Human Interaction. Due to the time-consuming nature of the manual annotation process, we perform human annotations on the set of 200 test conversations. In the forms for human annotations, we displayed the top 2 author's Reddit posts from the conversation, along with the submission title and content. We also provided a link to the original Reddit URL for annotators to look at the full conversation. We provided detailed annotation guidelines (similar to the ones mentioned in the prompt in Appendix B) to instruct human annotators in carrying out these tasks. Here is a sample Google form[6]. With three human annotators on a sample of 10 conversations, the agreement of majority labels (i.e., USDC test set labels) with human labels is 0.56 for the stance detection task and 0.45 for the dogmatism task. The annotators included two males and one female, affiliated with both academia and industry, aged between 20 and 40, and were very familiar with Reddit topics.

6 Conclusion

In this paper, we focused on the problems of 5-class stance and 4-class dogmatism classification in long conversations. Using LLMs as human-like annotators, we introduced USDC, a large-scale dataset of user stance and dogmatism in conversations. This is achieved by providing detailed annotation guidelines in the system prompt and full-length conversation as user prompt. Commercialized API-based LLMs generate author-level stance and dogmatism labels via zero, one and few-shot settings. The full-length multi-user conversation aspect of USDC allows it to capture the contextual and opinion shifts of multiple users in a conversation. We believe that the ability to perform finetuning or instruction tuning SLMs for user opinions at a large scale can bridge the gap between SLMs and commercial LLMs for understanding user traits. While finetuning SLMs shows F1-score on both stance and dogmatism tasks, the F1-score remains below 60% (54.9% for Stance and 51.4% for Dogmatism). On the other hand, instruction tuning of SLMs only improves F1-score performance on stance, not the dogmatism task. Further, the performance still falls short of 60%, with weighted F1-scores of 56.2% for stance and 49.2% for dogmatism. These findings indicate that there is still significant room for improvement in understanding user opinions from a text segment.

Limitations. We plan to extend this work along the following directions in the future. 1) We performed this work on English conversations only. It would be nice to extend this to multi-lingual conversations and verify how accurately SLMs and LLMs perform on the Stance and Dogmatism tasks in the multi-lingual scenario. 2) We analyzed user dogmatism based on their posts within a single conversation. This approach could be extended to include posts across multiple conversations and utilize similar profile information if available. 3) We analyzed dogmatism information for only the top two authors. This was mainly because considering more authors increases the output generation length, and we were constrained by our budget. This implies that our current models have not been evaluated for authors who do not post frequently.

[6] `https://forms.gle/dbPQBsNyfNJjvUeR9`

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179*, 2023.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

Parikshit Bansal and Amit Sharma. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766*, 2023.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931*, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Ethan Fast and Eric Horvitz. Identifying dogmatism in social media: Signals and models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 690–699, 2016.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76 (5):378, 1971.

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

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14409–14428, 2023.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen White, and Sujay Jauhar. Making large language models better data creators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15349–15360, 2023.

London Lowmanstone, Ruyuan Wan, Risako Owan, Jaehyung Kim, and Dongyeop Kang. Annotation imputation to individualize predictions: Initial studies on distribution dynamics and model predictions. *arXiv preprint arXiv:2305.15070*, 2023.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477, 2022.

R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5), 2023.

Milton Rokeach. The nature and meaning of dogmatism. *Psychological Review*, 61(3), 1954.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 300–325, Online, 2021. Association for Computational Linguistics.

Flora Sakketou, Allison Lahnala, Liane Vogel, and Lucie Flek. Investigating user radicalization: A novel dataset for identifying fine-grained temporal shifts in opinion. *arXiv preprint arXiv:2204.10190*, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Manya Wadhwa, Jifan Chen, Junyi Jessy Li, and Greg Durrett. Using natural language explanations to rescale human judgments. *arXiv preprint arXiv:2305.14770*, 2023.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11653–11669, 2022a.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Progen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3671–3683, 2022b.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. Llmaaa: Making large language models as active annotators. *arXiv preprint arXiv:2310.19596*, 2023.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 270–278, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.30. URL `https://aclanthology.org/2020.acl-demos.30`.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have ensured that the main claims made in the abstract and introduction are directly correlating to the research findings and the methods we have employed.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the main limitations of the work performed by the authors in the discussion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our paper does not require any explicit theorems and proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper has delineated all the information related to the experimental setup in the experimental setup section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have released the code and dataset, making the dataset publicly available under a license.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provided all the training and testing details in the experimental setup.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We conducted our experiments on all LLM-generated annotations across zero-shot, one-shot, and few-shot settings, using majority voting as labels. Our reported results represent the average performance across all test samples.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have included the specifications of the hardware and software environments to ensure the reproducibility of our results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper fully conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper explores how advancements and applications of our findings could benefit society by capturing opinions of users in conversation benefit interpersonal skills. Specifically, we investigate the effectiveness of current state-of-the-art large language models in this context.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our research does not pose any risks for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have explicitly cited the crawled websites, code and models used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We open-source the code and the new USDC dataset, and we provide complete documentation on how the dataset was created.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provide full instructions on how we surveyed our LLM generated annotations using human participants in the Results section.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We use publicly available Reddit user conversations to create the USDC dataset, and we do not collect any new data that would require IRB approval.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

A Detailed Statistics of the USDC Dataset

Table 4 shows the detailed statistics of our USDC dataset at the sub-reddit level. Fig. 5 shows the distribution of stance labels across LLM annotations across zero-shot, one-shot and few-shot settings. Fig. 6 shows the distribution of dogmatism labels across LLM annotations across zero-shot, one-shot and few-shot settings.

Table 4: Statistics of the User Conversation Dataset.

subreddit	num_conversations	min_total_token_count	max_total_token_count
DebateCommunism	73	529	11557
Abortiondebate	70	1271	7401
CapitalismVSocialism	61	665	16927
prochoice	60	582	7278
brexit	56	637	4553
climateskeptics	56	734	7550
prolife	54	672	13342
gunpolitics	52	683	7889
MensRights	52	623	5774
climatechange	49	520	7427
nuclear	41	572	5282
progun	39	436	3632
NuclearPower	23	629	4589
Vegetarianism	22	627	3958
AntiVegan	20	351	5052
climate	13	701	4678
Egalitarianism	10	665	4060
VeganActivism	8	460	3685
Veganism	2	1332	1738
AnimalRights	1	845	845
animalwelfare	1	1363	1363
GunsAreCool	1	2945	2945

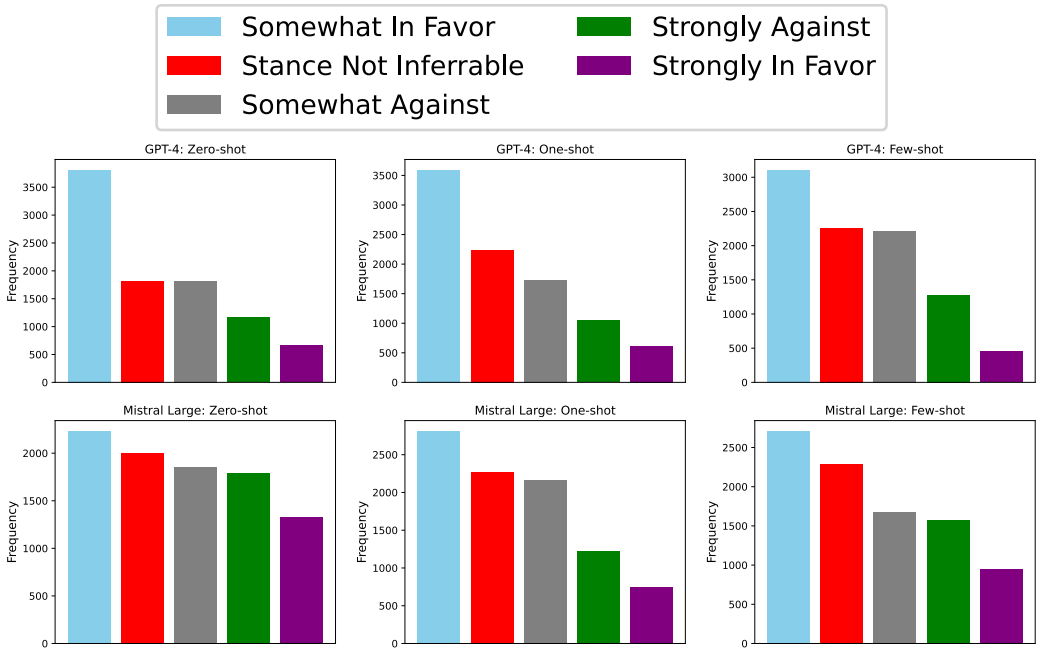


Figure 5: Distribution of Stance labels across LLM annotations.

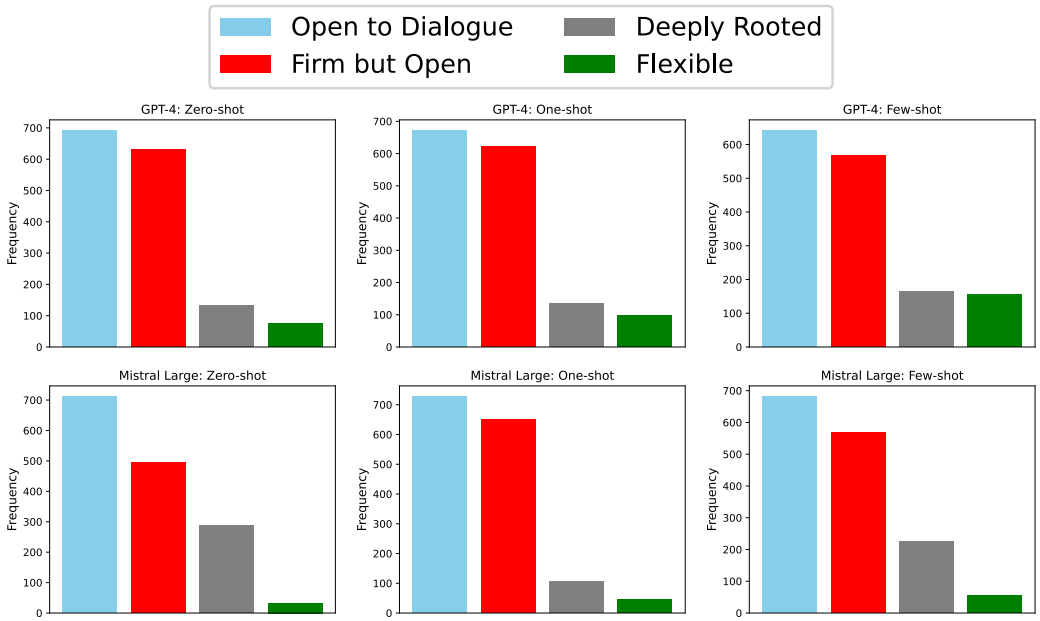


Figure 6: Distribution of dogmatism labels across LLM annotations.

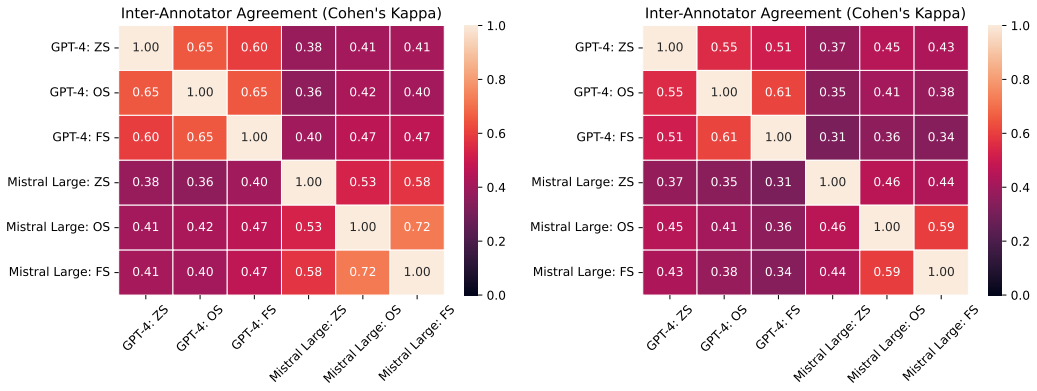


Figure 7: Inter-annotator agreement (IAA): Cohen's Kappa score across six different models (2 models×3 settings) for Stance (left) and Dogmatism (right) tasks.

B System Prompt for LLM Annotation

We used the following prompt for getting annotations from LLMs as well as for instruction-tuning of SLMs.

```
"""
### Introduction
**Objective**: Analyze Reddit conversations to identify the stance of
    specific authors on sociopolitical topics and determine their level of
    dogmatism.
**Stance Definition**: Stance is defined as the expression of the author's
    standpoint and judgement towards a given topic.
**Dogmatism Definition**: Dogmatism is an opinion strongly believed as a fact
    to support a stance without a question or allowance for conversation.
**Task**: Given a JSON formatted Reddit submission and its comment thread,
    classify the stance of text segments related to ''author1'' and
    ''author2'' by assigning one of the following five predefined stance
    labels: 'strongly_against', 'somewhat_against', 'somewhat_in_favor',
    'strongly_in_favor', 'stance_not_inferrable'. Also, assign a dogmatism
```

```
    label for each author by assigning one of the following four predefined
    labels: ‘Deeply Rooted’, ‘Firm but Open’, ‘Open to Dialogue’, ‘Flexible’.

### Description of Stance Labels:
1. **strongly_against / strongly_in_favor**: Marks text showing strong
    opinions, emotional expressions, or argumentative tones.
2. **somewhat_against / somewhat_in_favor**: Identifies texts with openness
    to discussion, less certainty, or showing interest in different
    viewpoints.
3. **stance_not_inferrable**: Use for texts that are neutral, support both
    stances, or where the stance is unclear despite being on-topic.

### Description of Dogmatism Labels:
1. **Deeply Rooted**: Reflects a strong, unchangeable belief. This label
    conveys the idea of someone who is firm in their opinion and unlikely to
    be swayed.
2. **Firm but Open**: Indicates a person who is not likely to change their
    mind but does not impose their views authoritatively. It captures the
    essence of being steadfast in one’s beliefs without being dismissive of
    others.
3. **Open to Dialogue**: Describes someone who holds a certain opinion but is
    genuinely interested in considering other viewpoints. This label suggests
    a willingness to engage in meaningful conversation about differing
    perspectives.
4. **Flexible**: Denotes a person who is not firmly committed to their stance
    and is open to changing their opinion. This label is indicative of
    flexibility and openness to new information or arguments.

### Input Data Format
The input data will be in JSON format and will include several key elements
    to represent a Reddit submission and its associated comments. Each
    element provides specific information as described below:

- ‘id’: This is the unique identifier for the Reddit submission.
- ‘title’: The title of the post. This is what users see first and often
    summarizes or hints at the content of the submission.
- ‘content’: The main post’s detailed description. This text segment provides
    the core message or information the author wishes to communicate with the
    Reddit community. It may include narratives, questions, or any
    information relevant to the title.
- ‘author1’ or ‘author2’: The username of our focus author. This field is
    applicable if the post or comment is made by one of the specific authors
    we are tracking in the dataset.
- ‘comments’: An array (list) of comments related to the Reddit submission.
    Each comment in this array includes the following fields:
    - ‘id’: The unique identifier for the comment, allowing for identification
        and reference within the dataset.
    - ‘author1’ or ‘author2’: The username of the comment’s author, if it is
        made by one of our focus authors. This helps in tracking contributions
        by specific individuals.
    - ‘body’: The text of the comment. This is the main content of the comment
        where the author responds to the post or another comment, providing
        insights, opinions, or further information.
    - ‘replies’: An array of comments that are direct responses to this
        comment. The structure of each reply follows the same format as the
        initial comment, including ‘id’, ‘author1’ or ‘author2’ (if
        applicable), ‘body’, and potentially more ‘replies’.

### Output Data Format
Submit your annotations in JSON format, grouping all stance annotations under
    the key ‘‘stance_annotations’’. Each entry should be a dictionary
    containing the segment’s ‘‘id’’, your ‘‘label’’, and the ‘‘reason’’ for
    your choice. Include the dogmatism label and its justification under
    ‘‘dogmatism_label’’ and ‘‘dogmatism_reason’’ keys, respectively.
```

```
The output should follow this structure:
‘‘‘json
{
  "author1": {
    "name": "[author_name]",
    "stance_annotations": [
      {
        "id": "[segment_id]",
        "label": "[chosen_label]",
        "reason": "[Justification in <50 words]"
      },
      ...
    ],
    "dogmatism_label": "[chosen_dogmatism_label]",
    "dogmatism_reason": "[Justification in <50 words]"
  },
  "author2": {
    "name": "[author_name]",
    "stance_annotations": [
      {
        "id": "[segment_id]",
        "label": "[chosen_label]",
        "reason": "[Justification in <50 words]"
      },
      ...
    ],
    "dogmatism_label": "[chosen_dogmatism_label]",
    "dogmatism_reason": "[Justification in <50 words]"
  }
}
’’’
### Instructions for Effective Annotation

1. **Labeling Stance**: For each segment (including the original Reddit
    submission, comments, or replies) where "author1" or "author2" is
    mentioned, assign a stance label that best represents the stance
    expressed towards the discussed topic in the submission. This
    comprehensive approach ensures no relevant contribution by "author1" or
    "author2" is overlooked. Evaluate the stance based on the content’s tone,
    argumentation, and engagement level with the topic.
2. **Providing Justification**: For each label assigned, include a concise
    reason, aiming for less than 50 words. Focus on the stance and
    argumentative indicators present in the text.
3. **Dogmatism Assessment**: After reviewing all segments from "author1" and
    "author2", assign a single dogmatism label reflecting the overall tone
    and approach in their contributions.
"""
```

C Prompts for Finetuning SLMs

Fig. 8 and 9 shows the prompts used for finetuning SLMs for the stance and dogmatism classification tasks respectively.

Stance Classification

Analyze the stance of the post enclosed in square brackets.
Categorize each post into one of the following categories based on its stance:

- Somewhat In Favor
- Somewhat Against
- Stance Not inferrable
- Strongly In Favor
- Strongly Against

and return the answer as one of the corresponding stance labels.

`[{data_point["stance_id_comment"]}]`

Figure 8: Prompt for stance classification, for finetuning SLMs.

User Dogmatism Identification

Analyze the comments of a user in conversation enclosed in square brackets.
Categorize the opinion fluctuation of the user into one of the following categories based on its change:

- Open to Dialogue
- Firm but Open
- Deeply Rooted
- Flexible

Return the answer as one of the corresponding dogmatism labels.

`[{data_point["comments_string_for_dogmatism"]}]`

Figure 9: Prompt for dogmatism classification, for finetuning SLMs.

D Sample of User Input Prompt

```
"""
### User Prompt
Now complete the given task for the respective authors i.e., author1
    name is "rookerin0" and respective ids are ['dhoxyz', 'f3pghji', '
    f3tywb4', 'f3uomn2']. author2 name is "MikeWillTerminate" and
    respective ids are ['f3rt0bf', 'f3rqu2u'] for the data in json
    format
{
   "id":"dhoxyz",
   "author1":"rookerin0",
   "title":"This sub should encourage anti vs. pro-gun discussions
       instead of shutting them down instantly",
   "content":"Honesly, I followed this sub especifically to take part
       in these discussions, but everytime I see a comment that even
       remotely suggests anti gun ideals or a discussion on the
       subject just gets ignored and downvoted to hell. Kind of
       expecting this to go the same way (my karma anus is ready,
       downvotes) , but I have to hope for healthy discussions on the
       subject.",
   "comments":[
      {
         "id":"f3p9n2c",
         "body":"I think the problem now is the two sides are at an
             impasse. Everytime there is a "compromise" pro gun loses
```

```
            something. Now days pro gun is interpreting the
            Constitution more literal, which leaves even the most
            mild policies of anti gun as infringements. To further
            compound this anti gun is only considering the most
            extreme measures. "Assault Weapons" bans, mandatory
            buybacks, red flag laws, etc.. I think at this point
            there is just nothing left to talk about. The middle
            ground is gone.",
         "replies":[
            {
               "id":"f3pati9",
               "replies":[
                  {
                     "id":"f3pdu44",
                     "body":"You are exactly right. I'm done with the
                         idea that there can be real compromise. We
                         should have at least gotten national
                         reciprocity and shall-issue in every state in
                          exchange for what we've given up. Now you
                         have to be a goddamn lawyer to exercise your
                         rights without violating the law."
                  },
                  {
                     "author2":"MikeWillTerminate",
                     "id":"f3rt0bf",
                     "body":"I am prepared for UBCs, if they do this:
                     1. Lower the age to buy handguns to 18,
                         nationwide.
                     2. Repeal the Hughes Amendment:
                     3. A FOPA-like ban on assault weapon bans (what
                         the FOPA did with a registry)
                     4. The punishment for violation is a monetary
                         fine only
                     5. A repeal of the GCA ban on foreign NFA weapons
                     6. A repeal of the National Minimum Drinking Age
                         Act of 1984"
                  }
               ]
            },
            {
               "id":"f3pd55z",
               "body":"Everytime there is a "compromise" pro gun loses
                    something. That and today's compromise is tomorrow
                   's loophole to be closed. All such compromises do
                   is push that policy off until the next round."
            }
         ]
      },
      {
         "id":"f3paf0j",
         "body":"Yeah this sub it's not conducive to conversion. Its
            quickly devolving to little more than "Boogaloo" memes
            and shouting "SHALL. NOT." at each other. However,  as
            far as I know, the mods won't delete your thread and ban
            you from the sub for trying to have a good faith
            discussion,  like some of the gun control subs will.",
         "replies":[
            {
               "id":"f3pusbm",
               "body":"Unfortunately this sub's mod team takes a very
                   passive approach to moderation. With very little
                   effort they could make this sub into a  quality
                   progun meeting ground *without having to resort to
                   censorship*. Instead they promote low-effort memes
                   and endless duplication of posts through their
```

```
                    inaction. whubbard has the chops to resurrect this
                    sub. Let's see if he's up to the challenge.",
                "replies":[
                   {
                      "id":"f3q8xj6",
                      "body":"We voted to ban memes last week. All
                          about rolling it out now.",
                      "replies":[
                         {
                            "id":"f3qn4p8",
                            "body":"Damn I might have to eat some crow
                                here then..."
                         }
                      ]
                   }
                ]
             }
          ]
       },
       {
          "id":"f3pafqa",
          "body":"Found the gun grabber!!",
          "replies":[
             {
                "id":"f3pcw4h",
                "body":"Witch hunter."
             }
          ]
       },
       {
          "id":"f3pal5l",
          "body":"I see people have discussions when it makes sense to.
               Not much reason to spend time responding to the same gun
               control measures over and over though."
       },
       {
          "id":"f3paw3h",
          "body":"I get where you're coming from, but people's ability
              to protect themselves and own their own property isn't
              something that is compromisable. Anything less, and they
              cease to own their own property. It's like breathing,
              there can be nothing less than total ability to breath
              when and how someone wants. It's just that simple."
       },
       {
          "id":"f3pax9m",
          "body":"My take on this, What kind of open discussion is
              possible for a right that is guaranteed and most
              importantly, not to be infringed upon? They're making all
               these unlawful laws to portray it as it's somehow
              legitimate. They are not, We are at an apex, to which
              both political spectrums and even us to a degree are
              liable for.\nI certainly believe both sides are waiting
              for this to boil over so each can finger point. I just
              speculate it's going to be the hell humanity been
              whispering about but never thought it would ever occur."
       },
       {
          "id":"f3pb6ny",
          "body":"The time for discussion is over."
       },
       {
          "id":"f3pfqwq",
          "body":"I don't know what you're talking about. Sure people
              downvote, but they also talk. We get "why do you need
```

```
            guns" posts at least weekly, and several people will
            engage in actual conversation with them, citing facts,
            clearing up statistics, and telling stories to illustrate
             why this is important to them, but they are usually met
            with "you stupid @#$%, you think you're Rambo" or
            something equally clever. People who come here to discuss
             and learn will be treated well. People who are just
            trolling are treated like trolls.",
         "replies":[
            {
               "author1":"rookerin0",
               "id":"f3pghji",
               "body":"I made this post because I'm always seeing
                   rational, conversation seeking comments getting
                   blown to downvote hell.",
               "replies":[
                  {
                  "id":"f3pi9xv",
                  "body":"[Like this one?](https://www.reddit.com/r/
                     progun/comments/dhcu92/yup/f3p75tg/)> One smart
                     man in a sub full of... welp... "strong opinions
                     ". You start off with arrogance, as the sole
                     arbiter of what constitutes a "smart man". Then
                     you back it up with a dismissive swipe at what
                     you term "strong opinions".> Every other country
                     can see that PROPER gun control reduces gun
                     violence by a ton, More arrogance. False
                     equivalence. Unsupported claims.> but the US
                     refuses to let go of it's antique laws In a
                     shocking turn of events, more arrogance.> Fully
                     aware that this is a fully pro gun sub, willing
                     to take the downvotes in order to spark a
                     discussion and crack some heads. You aren't the
                     first arrogant asshole to grace this sub with
                     posts like this. Try bringing something other
                     than your own self-importance to the discussion.
                     Edit: And then there's [this gem](https://www.
                     reddit.com/r/unpopularopinion/comments/d3w5z1/
                     people_living_in_the_us_are_living_in_one_of_the/
                     f06r3sg/.> Wanna feel like you could be shot at
                     every single moment? Move to the US, it'll prob
                     happen to you either as a bystander, or you'd be
                     shot by a random citizen (sometimes police)."
                  },
                  {
                     "id":"f3pj8k0",
                     "body":"As is tradition. We're done with that
                         condescending bullshit from antis, you dont
                         come here for good faith discussion and
                         whether you get a reasonable response or not,
                          nothing ever changes, easier to downvote you
                          and move on because we get the same
                         treatment anytime we attempt to speak out in
                         anti subs."
                  },
                  {
                     "id":"f3plgf4",
                     "body":"If downvotes hurt your feelings, you
                         shouldn't be on reddit. People tend to
                         downvote anything they disagree with (which
                         is why some subs specifically ask you to only
                          downvote things that contribute nothing to
                         the discussion). It's a bad habit, but that's
                          the way it is. People downvote and *still*
                         enage. You want to post a view contrary to
```

```
                        the prevailing view of the sub, take your
                        lumps and participate in what conversation
                        you are offered. But if you're only here to
                        preach about how stupid, misguided, unevolved
                        , uneducated, irrational, and/or violent we
                        are, don't expect a polite response."
                  },
                  {
                     "id":"f3tcgf1",
                     "body":"An arrogant Israeli trying to tell
                         another nation how they should be run. You're
                          just a walking stereotype aren't you? And
                         before you say anything, I popped into your
                         comment history. That's where the calling you
                          Israeli comes from.",
                     "replies":[
                        {
                           "author1":"rookerin0",
                           "id":"f3tywb4",
                           "body":"I thought that trying to tell other
                               nations how they should run was your
                              guys's stereotype.",
                           "replies":[
                              {
                                 "id":"f3u0vkq",
                                 "body":"No we go in and try to make
                                     them work our way."
                              }
                           ]
                        }
                     ]
                  }
               ]
            }
         ]
      },
      {
         "id":"f3pzseh",
         "body":"It's a little unfortunate but the grabbers who come
             on here tend to be intellectually dishonest and/or
             uninformed. There was some Australian post a few days ago
              that pretty much asked why we like our guns more than
             children. No discussion to be had there. There's also
             some posts that clearly demonstrate the poster should
             inform himself or herself a little."
      },
      {
         "author2":"MikeWillTerminate",
         "id":"f3rqu2u",
         "body":"Actually, do that. It shows everyone that they tend
             to be crazy, unstable, ignorant, stereotyping, arrogant
             bastards who hate black people with a hair trigger."
      },
      {
         "id":"f3t7tgg",
         "body":"Welcome to reddit, home of every single safe place
             for anything that doesnt violate the TOS. At least its
             slightly better than r/politics"
      },
      {
         "id":"f3unt9z",
         "body":"This isn't r/gundebate. This is a pro gun subreddit.
              That said, we do allow some debate provided it remains
             civil.",
         "replies":[
```

```
            {
                "author1":"rookerin0",
                "id":"f3uomn2",
                "body":"Sadly tho, r/gundebate is pretty dead..."
            }
        ]
    },
    {
        "id":"f4dip6o",
        "body":"Anything else you want to give away for free?"
    }
  ]
}

 #### Response
```

E Samples of JSON Outputs from LLMs

E.1 Sample JSON Output from LLM: GPT-4 Zero-shot

```
{
    "author1": {
    "name": "rookerin0",
    "stance_annotations": [
    {
        "id": "dhoxyz",
        "label": "somewhat_in_favor",
        "reason": "Advocates for open discussions, showing interest in
            dialogue."
    },
    {
        "id": "f3pghji",
        "label": "somewhat_in_favor",
        "reason": "Expresses desire for rational discussions, despite
           downvotes."
    },
    {
        "id": "f3tywb4",
        "label": "somewhat_against",
        "reason": "Sarcastic retort, but lacks strong emotional
           language."
    },
    {
        "id": "f3uomn2",
        "label": "stance_not_inferrable",
        "reason": "Comment is factual, no clear stance on the debate
           itself."
    }
    ],
    "dogmatism_label": "Open to Dialogue",
    "dogmatism_reason": "Shows interest in discussion, acknowledges
       downvotes."
    },
    "author2": {
    "name": "MikeWillTerminate",
    "stance_annotations": [
    {
        "id": "f3rt0bf",
        "label": "somewhat_in_favor",
        "reason": "Proposes conditions for UBCs, indicating some
           openness."
    },
```

```
    {
        "id": "f3rqu2u",
        "label": "strongly_against",
        "reason": "Uses strong negative descriptors for opponents,
            indicating firm stance."
    }
    ],
    "dogmatism_label": "Firm but Open",
    "dogmatism_reason": "Mixed; proposes solutions but also dismisses
        opposition."
    }
}
```


E.2 Sample JSON Output from LLM: GPT-4 One-shot


```
{
    "author1": {
    "name": "rookerin0",
    "stance_annotations": [
    {
        "id": "dhoxyz",
        "label": "somewhat_in_favor",
        "reason": "Advocates for open discussions, indicating a pro-
            discussion stance."
    },
    {
        "id": "f3pghji",
        "label": "somewhat_in_favor",
        "reason": "Expresses desire for rational conversation, showing
             a supportive stance for dialogue."
    },
    {
        "id": "f3tywb4",
        "label": "somewhat_against",
        "reason": "Uses sarcasm to counter a stereotype, indicating a
            stance against the opposing view."
    },
    {
        "id": "f3uomn2",
        "label": "somewhat_in_favor",
        "reason": "Expresses disappointment in lack of active debate,
            supporting open discussion."
    }
    ],
    "dogmatism_label": "Open to Dialogue",
    "dogmatism_reason": "Shows a consistent preference for discussion
        and engagement, without dogmatic assertions."
    },
    "author2": {
    "name": "MikeWillTerminate",
    "stance_annotations": [
    {
        "id": "f3rt0bf",
        "label": "somewhat_in_favor",
        "reason": "Lists conditions for agreement, showing openness to
             compromise."
    },
    {
        "id": "f3rqu2u",
        "label": "strongly_against",
        "reason": "Uses strong language to describe opponents,
            indicating a firm stance against them."
    }
```

```
    ],
        "dogmatism_label": "Firm but Open",
        "dogmatism_reason": "While showing strong opinions, also lists
            specific conditions for compromise."
    }
}
```

E.3 Sample JSON Output from LLM: GPT-4 large Few-shot

```
{
    "author1": {
    "name": "rookerin0",
    "stance_annotations": [
    {
        "id": "dhoxyz",
        "label": "somewhat_in_favor",
        "reason": "Expresses desire for open discussions, indicating a
            stance in favor of dialogue."
    },
   {
        "id": "f3pghji",
        "label": "somewhat_in_favor",
        "reason": "Advocates for rational conversation despite
           observing negative reactions."
    },
   {
        "id": "f3tywb4",
        "label": "somewhat_against",
        "reason": "Uses sarcasm to counter a stereotype, showing a
           stance against it."
    },
   {
        "id": "f3uomn2",
        "label": "stance_not_inferrable",
        "reason": "Comment on the state of another subreddit does not
           reveal a stance on the gun debate."
    }
    ],
    "dogmatism_label":  "Open to Dialogue",
    "dogmatism_reason": "Shows a consistent
    preference for discussion over confrontation."
    },
    "author2": {
    "name": 'MikeWillTerminate",
    "stance_annotations": [
    {
        "id": 'f3rt0bf",
        "labe"': 'somewhat_in_favor",
        "reason": 'Lists conditions for accepting UBCs, indicating
           some openness to compromise."
    },
    {
        "id": 'f3rqu2u",
        "label": "strongly_against",
        "reason": "Uses derogatory language to describe opponents,
           showing strong opposition."
    }
    ],
    "dogmatism_label": "Firm but Open",
    "dogmatism_reason": "While showing strong opinions, also suggests
       conditions for policy acceptance."
    }
}
```

E.4 Sample JSON Output from LLM: Mistral Large Zero-shot


```
{
    "author1": {
    "name": "rookerin0",
    "stance_annotations": [
    {
        "id": "dhoxyz",
        "label": "somewhat_in_favor",
        "reason": "Author expresses desire for open discussions on
           anti-gun ideals."
    },
    {
        "id": "f3pghji",
        "label": "stance_not_inferrable",
        "reason": "Author is commenting on downvotes, not expressing
           stance."
    },
    {
        "id": "f3tywb4",
        "label": "stance_not_inferrable",
        "reason": "Author is making a stereotypical comment, not
           expressing stance."
    },
    {
        "id": "f3uomn2",
        "label": "stance_not_inferrable",
        "reason": "Author is commenting on the state of another
           subreddit, not expressing stance."
    }
    ],
    "dogmatism_label": "Open to Dialogue",
    "dogmatism_reason": "Author expresses desire for open discussions
       and does not dismiss opposing views."
    },
    "author2": {
    "name": "MikeWillTerminate",
    "stance_annotations": [
    {
        "id": "f3rt0bf",
        "label": "strongly_in_favor",
        "reason": "Author lists conditions for accepting UBCs, showing
            strong pro-gun stance."
    },
    {
        "id": "f3rqu2u",
        "label": "strongly_in_favor",
        "reason": "Author"s comment is dismissive of anti-gun
           advocates, showing strong pro-gun stance."
    }
    ],
    "dogmatism_label": "Deeply Rooted",
    "dogmatism_reason": "Author"s comments are dismissive of opposing
       views and show a strong, unchangeable belief in pro-gun ideals
       ."
    }
}
```


E.5 Sample JSON Output from LLM: Mistral Large One-shot


```
{
    "author1": {
    "name": "rookerin0",
```

```
    "stance_annotations": [
    {
        "id": "dhoxyz",
        "label": "somewhat_in_favor",
        "reason": "Expresses desire for open discussions on gun
           control, but follows pro-gun subreddit."
    },
    {
        "id": "f3pghji",
        "label": "stance_not_inferrable",
        "reason": "Comments on downvotes, but no clear stance on gun
           control."
    },
    {
        "id": "f3tywb4",
        "label": "stance_not_inferrable",
        "reason": "Comment about stereotypes, but no clear stance on
           gun control."
    },
    {
        "id": "f3uomn2",
        "label": "stance_not_inferrable",
        "reason": "Mentions r/gundebate being dead, but no clear
           stance on gun control."
    }
    ],
    "dogmatism_label": "Open to Dialogue",
    "dogmatism_reason": "Seeks open discussions and does not express
       strong dogmatic views."
    },
    "author2": {
    "name": "MikeWillTerminate",
    "stance_annotations": [
    {
        "id": "f3rt0bf",
        "label": "somewhat_in_favor",
        "reason": "Lists conditions for supporting UBCs, showing
           openness to compromise."
    },
    {
        "id": "f3rqu2u",
        "label": "strongly_in_favor",
        "reason": "Strong negative language against gun control
           advocates, suggesting strong pro-gun stance."
    }
    ],
    "dogmatism_label": "Firm but Open",
    "dogmatism_reason": "Expresses strong pro-gun views but open to
       compromise under certain conditions."
    }
}
```


E.6 Sample JSON Output from LLM: Mistral Large Few-shot


```
{
    "author1": {
    "name": "rookerin0",
    "stance_annotations": [
    {
        'id': "dhoxyz",
        "label": "somewhat_in_favor",
        "reason": "Expresses desire for open discussions on anti vs.
           pro-gun debates."
```

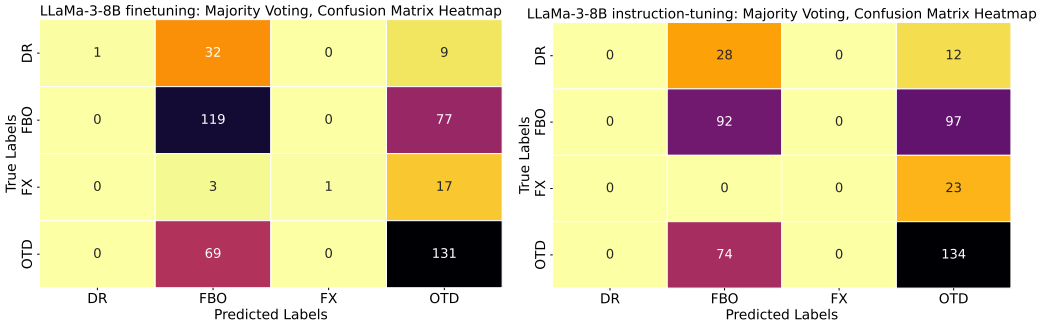


Figure 10: Confusion matrix for LLaMa-3-8B Dogmatism detection models on USDC test set: finetuning (left) and instruction-tuning (right). Here, DR: Deeply Rooted, FX: Flexible, FBO: Firm but Open, OTD: Open to Dialogue


```
    },
    {
        "id": "f3pghji",
        "label": "stance_not_inferrable",
        "reason": "Comments on downvotes, but no clear stance on the
            topic."
    },
    {
        "id": "f3tywb4",
        "label": "stance_not_inferrable",
        "reason": "Makes a stereotypical comment, but no clear stance
            on the topic."
    },
    {
        "id": "f3uomn2",
        "label": "stance_not_inferrable",
        "reason": "Mentions r/gundebate being dead, but no clear
            stance on the topic."
    }
    ],
    "dogmatism_label": "Open to Dialogue",
    "dogmatism_reason": "Seeks open discussions and engages in
        conversation without strong dogmatic language."
    },
    "author2": {
    "name": "MikeWillTerminate",
    "stance_annotations": [
    {
        "id": "f3rt0bf",
        "label": "somewhat_in_favor",
        "reason": "Lists conditions for supporting UBCs, showing
            openness to discussion."
    },
    {
        "id": "f3rqu2u",
        "label": "strongly_against",
        "reason": "Uses derogatory language to express strong
            opposition to 'grabbers'."
    }
    ],
    "dogmatism_label": "Firm but Open",
    "dogmatism_reason": "Expresses strong opinions but also shows
        willingness to consider certain conditions for compromise."
    }
}
```

F SLM finetuning: AUC (Area Under the Curve) analysis

Fig. 10 illustrates the confusion matrix for dogmatism detection for LLaMa-3-8B finetuning and instruction-tuning. We make the following observations from Fig. 10: 1) For both finetuning and instruction-tuning, there is significant misclassifications, especially for the "Deeply Rooted" and "Flexible" labels, with both having zero accuracy and F1-scores. While "Firm but Open" and "Open to Dialogue" perform moderately better, with accuracies of 48.7% and 64.4% respectively. The confusion matrix indicates substantial confusion to distinguish between intermediate levels of dogmatism, such as "Firm but Open" and "Open to Dialogue. We further reports the ROC curve shows the trade-off between the true positive rate (TPR) and false positive rate (FPR) for each class for stance and dogmatism tasks, in Figs. 11 and. 12. The area under the ROC curve (AUC) is a measure of the model's ability to distinguish between classes.

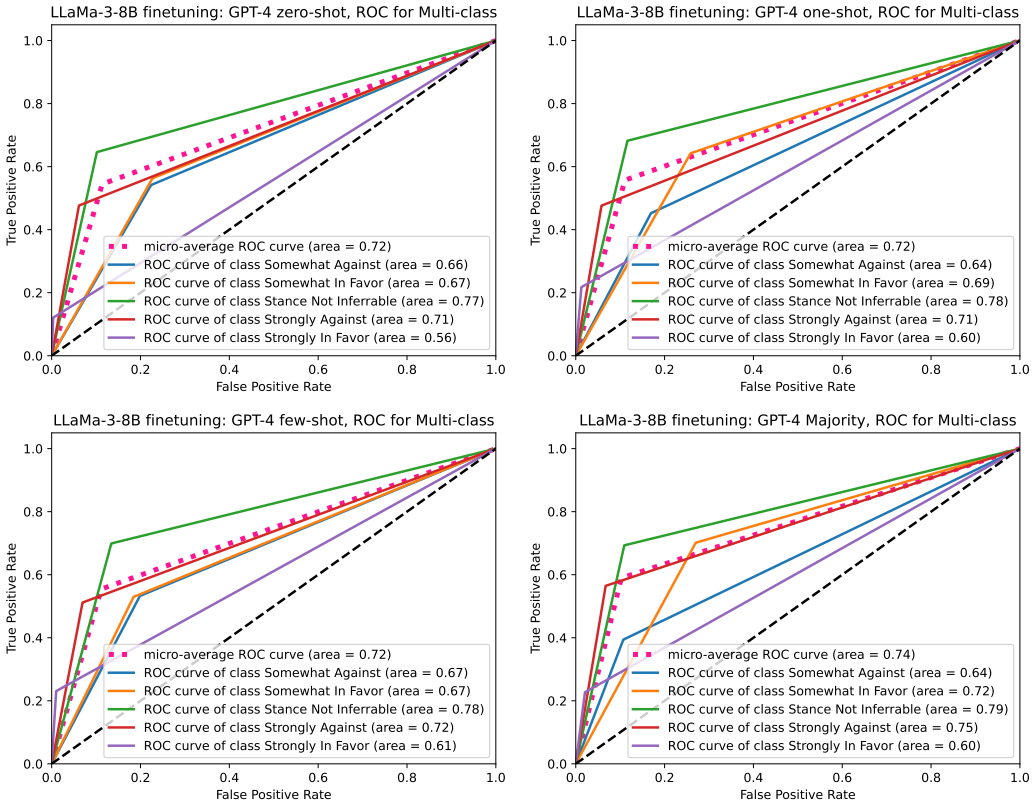


Figure 11: LLaMa-3-8B finetuning for stance detection task: Visualize the ROC curves for each class along with their AUC values for GPT-4 Annotations across zero-shot, one-shot, few-shot and majority labels.

G SLM instruction-tuning: AUC (Area Under the Curve) analysis

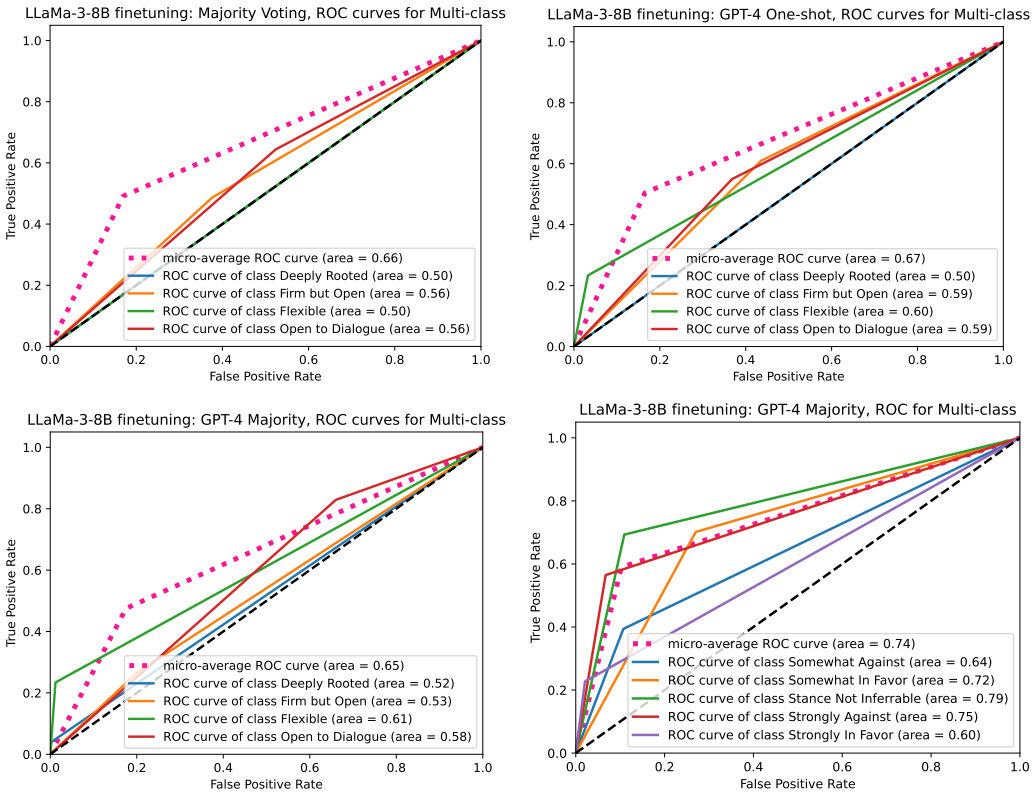


Figure 12: LLaMa-3-8B finetuning for dogmatism task: Visualize the ROC curves for each class along with their AUC values for GPT-4 Annotations across zero-shot, one-shot, few-shot and majority labels.

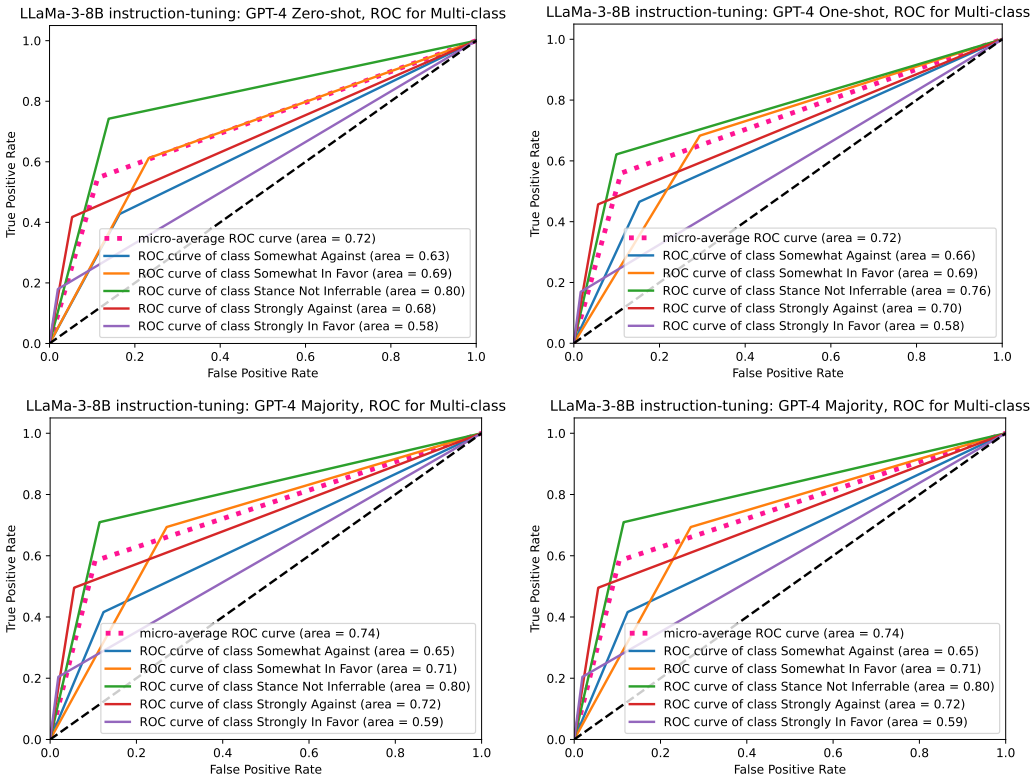


Figure 13: LLaMa-3-8B instruction-tuning for stance detection task: Visualize the ROC curves for each class along with their AUC values for GPT-4 Annotations across zero-shot, one-shot, few-shot and majority labels.