USDC: A DATASET OF <u>U</u>SER <u>S</u>TANCE AND <u>D</u>OGMATISM IN LONG <u>C</u>ONVERSATIONS

Anonymous authors

Paper under double-blind review

ABSTRACT

Although prior studies have explored Stance and Dogmatism in user conversations, their datasets are constructed at the post level, treating each post as independent and randomly sampling posts from conversation threads. Consequently, these datasets cannot capture users' opinion fluctuations expressed throughout the entire conversation context. Identifying these user's opinion fluctuations in long conversation threads on various topics is extremely critical for various applications, including enhanced personalization, market research, political campaigns, customer service, targeted advertising, and content moderation. Therefore, training language models to automate this task is essential. However, gathering manual annotations to train such models presents multiple challenges: 1) It is time-consuming and costly; 2) Conversation threads could be very long, increasing the 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) in complex natural language processing tasks, we leverage Mistral Large and GPT-4 to automate the human annotation process on two tasks while also providing reasoning: i) User Stance classification, which involves labeling a user's stance in a post within a conversation on a five-point scale; ii) User Dogmatism classification, which involves labeling a user's overall opinion in the conversation on a four-point scale. By applying the Majority voting on zero-shot, one-shot, and few-shot annotations from these two LLMs on 764 multi-user Reddit conversations, we curate the USDC dataset. USDC is then used to finetune and instruction-tune multiple deployable small language models for the 5-class stance and 4-class dogmatism classification tasks. Additionally, human annotations on 200 test conversations achieved inter-annotator agreement scores of 0.49 for stance and 0.50 for dogmatism, indicating a reasonable level of consistency between human and LLM annotations. We make the code and dataset publicly available 1 .

037

000

001

002 003 004

006

008 009

010 011

012

013

014

015

016

017

018

019

021

024

025

026

027

028

029

031

032

034

1 INTRODUCTION

Understanding fluctuations in a user's (or author's) opinions during a conversation is a fundamental 040 to successful interpersonal interactions. It is essential for developing better communication skills, 041 fostering empathy, and making informed decisions. This understanding is particularly relevant in the 042 context of dogmatism—a phenomenon observed in areas such as politics, religion, culture, intellect, 043 and science-where rigid adherence to beliefs often hinders open-mindedness and empathy (Rokeach, 044 1954). By aligning with the opinions and stances of potential customers, advertisers can target their 045 campaigns more effectively. Companies can leverage this information for market research, tailoring 046 products and services to meet consumer needs and preferences. Similarly, political groups can gauge 047 public reactions to policies and campaigns, adjusting their strategies accordingly. Identifying differing 048 opinions can facilitate conflict resolution by helping to understand the perspectives of all parties. 049 By recognizing and respecting diverse opinions, society can promote tolerance and maintain social harmony. 050

Fig. 1 illustrates a sample Reddit conversation on the topic of *Capitalism vs. Socialism*. In this context, an author's initial post—comprising the title and body—is referred to as a submission. Multiple

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



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 socialism and examines how the authors position their opinions regarding socialism vs. capitalism.

071 authors can then share their opinions as comments on this submission. Specifically, this example contains 6 comments $\{c_i\}_{i=1}^6$ from 3 authors $\{a_j\}_{j=1}^3$. We also display stance and dogmatism predictions from two LLMs: Mistral Large and GPT-4. Some authors, like a_1 , change their views during 073 074 the discussion based on the beliefs or opinions of others. At the beginning of the dialogue, author 075 a_1 somewhat favors socialism (in submission and c_2). However, after considering the viewpoints 076 of author a_2 in comments c_1 and c_3 , a_1 shifts their stance to somewhat favoring capitalism (in c_4), 077 illustrating a firm yet open-minded approach. On the other hand, author a_3 seems very flexible based on their comment c_5 . Conversely, author a_3 appears very flexible based on their comment c_5 . Understanding such conversations requires comprehending the fine-grained topics being discussed 079 and the dynamic viewpoints of 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 several prior studies have explored
Stance and Dogmatism at the post level, and numerous datasets exist for analyzing individual user
posts (Fast & Horvitz, 2016; Sakketou et al., 2022; Villa-Cox et al., 2020; Li et al., 2023; Niu et al.,
2024), these typically involve random subsampling or selecting posts with a limited number of tokens,
treating each post as independent. Consequently, the comprehensive exploration of a specific user's
opinion fluctuations within an entire conversational thread remains underexplored.

Crowdsourcing is one possible approach to address the need for a suitable dataset. However, manually annotating datasets for user opinions is time-consuming and costly, as annotators must read entire conversations to label each user's posts. 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 can change their opinion (often 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) (Touvron et al., 2023a;b; Jiang et al., 2023; Zhang et al., 2023b), 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 frequently used for synthetic 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 effectively capture the nuances involved in understanding user opinions and their shifts in multi-user conversational contexts. Furthermore, since these models possess long-range memory capabilities, we believe they can reason over extended conversational threads involving numerous participants, as good as human annotators, if not better.

106

In this work, we leverage LLMs like Mistral Large (Jiang et al., 2023) and GPT-4 (OpenAI, 2023) to perform two tasks: i) User Stance classification, which involves labeling a user's stance of a



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

post in a conversation on a five-point scale; ii) User Dogmatism classification, which deals with 126 labeling a user's overall opinion in the conversation on a four-point scale. Besides the class labels, 127 we also obtain the reasoning behind these labels from LLMs. We experiment with these two models 128 as human-like annotators, generating user opinions in full-length, multi-user Reddit conversations 129 in zero-shot, one-shot, and few-shot setups. Thus, for every sample, we obtain annotations in six 130 settings ({Mistral Large, GPT-4}×{zero-shot, one-shot, few-shot}). Fig. 2 presents our LLM-based 131 annotation pipeline for user-level Stance and Dogmatism tasks. We consider majority voting over these six settings as our final annotations. This approach enables us to curate our USDC (a dataset of 132 user stance and dogmatism in conversations) dataset, which consists of 764 multi-user conversations 133 from 22 subreddits, including 1.528 user-level dogmatism samples and 9.618 stance samples across 134 all posts from selected users. The annotations in the dataset highlight specific user opinions in each 135 post related to stance, track opinion fluctuations leading to a dogmatic nature, and provide reasoning 136 about why users hold specific opinions. 137

138 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 139 of multiple users. This feature allows it to serve as both an instruction-tuning user opinion dataset and 140 an evaluation benchmark. We believe that the ability to perform instruction-tuning for user opinions 141 at a large scale can bridge the gap between open-source and commercial user trait understanding 142 models. Additionally, the in-context learning annotations using state-of-the-art LLMs in USDC make 143 it a more comprehensive measure of how current LLMs understand complex tasks like capturing 144 opinions. Further, the USDC dataset offers several use cases that extend its value in various domains, 145 including, (i) Improving moderation tools, (ii) Analyzing public opinion dynamics, (iii) Enhancing 146 dialogue systems and (iv) Creating dynamic contextual user representations. These aspects make it a 147 valuable resource, especially for social media agents seeking deeper insights into user behavior.

148 To demonstrate the utility of USDC, we utilize our dataset to fine-tune and instruction-tune open-149 source LLMs for generating stance and dogmatism labels for users. We experiment with three 150 pretrained small language models (SLMs) like LLaMA-2-7B, LLaMA-3-8B (Touvron et al., 2023b), 151 and Falcon-7B (Almazrouei et al., 2023). We also experiment with four instruction-tuned SLMs like 152 LLaMA-2-chat-7B, LLaMA-3-8B-instruct, Vicuna-7B-v.1.5, and Falcon-7B-instruct. Additionally, 153 we conduct a comprehensive evaluation of the USDC dataset by incorporating human annotations on 154 200 test conversations and measuring inter-annotator agreement between LLM and human annotations. 155 When comparing the LLM-generated annotations with human annotations, it becomes evident that the "lost in the middle" phenomenon (Liu et al., 2024) is marginal in LLMs, whereas human annotators 156 maintain a steady understanding and agreement throughout the conversation, regardless of its length 157 or complexity. Additonally, the "recency bias" phenomenon (Peysakhovich & Lerer, 2023) shows 158 that human annotators rely heavily on the full context to maintain better inter annotator agreement 159 with LLMs. 160

161

108

109

110

111

112

113

114

115

116

117

118 119 120

121

122

123

162 We make the following contributions: 1) We introduce USDC, a dataset of user stance and dogmatism 163 in conversations dataset. 2) We benchmark initial results for the stance and dogmatism tasks using 164 seven SLMs for the UDSC dataset. We find that stance performance improves with instruction-tuning 165 (F1-score of 56.2) compared to finetuning (F1-score of 54.9). However, dogmatism performs worse 166 with instruction-tuning (F1-score of 49.2) compared to finetuning (F1-score of 51.4), highlighting the complexity of this task. 3) We provide human annotations on 200 test conversations, achieving inter-167 annotator agreement scores of 0.49 for stance and 0.50 for dogmatism, indicating a reasonable level 168 of consistency between human and LLM annotations. 4) We apply transfer learning by fine-tuning SLMs on the USDC dataset and assess the model's performance on existing post level stance datasets, 170 including SPINOS, MT-CDS, and the Twitter-stance. We find that our transfer learning results are 171 either comparable to or outperform prior studies. 5) We make the code, models and dataset publicly 172 available¹.

173 174

2 RELATED WORK

175 176

Post level stance and dogmatism. Previous stance detection studies have primarily focused on 177 evaluating stances within individual posts of users or through multi-party discussions on some 178 specific topic in social media interactions (Villa-Cox et al., 2020; Sakketou et al., 2022; Li et al., 179 2023; Niu et al., 2024). Sakketou et al. (2022) introduced the post level Stance dataset, SPINOS, 180 where each post is considered independently, without including submission posts for context, which 181 affects the labeling by annotators. Recently, the MT-CSD dataset, introduced by Niu et al. (2024), 182 addresses stance detection in multi-turn conversations with multiple targets, addressing different 183 aspects of stance detection while the focus is on the multi-party discussions. In contrast to these 184 two studies, Villa-Cox et al. (2020) specifically focus on extracting stances (denying vs. supporting 185 opinions) from replies and quotes on controversial issues in Twitter conversations. Li et al. (2023) focus on target-specific stance detection, where the goal is to classify individual posts or comments 187 into a stance class related to a specific issue, such as COVID-19 vaccination. From the above studies, we clearly observe that these works focus more on stance detection at the post level, while our 188 work emphasizes user-level opinion fluctuations. Additionally, the prior studies are limited in scope, 189 targeting specific issues (5 topics in (Villa-Cox et al., 2020), 1 topic in (Li et al., 2023)), whereas 190 USDC covers a broader range of general subreddits across 22 different topics. 191

Similar to post level stance datasets, 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 the treatment of each post as independent. Overall, all these prior studies contrast with the USDC dataset, which focuses on tracking user-level opinions across long, multi-user conversations, capturing the evolution of stance and dogmatism over extended discussions rather than just on a specific target issues.

Generating annotations for NLP tasks using LLMs. Our work also relates to a growing body of literature suggesting that LLMs can perform similarly to human annotators in labeling complex 199 NLP tasks (Zhou et al., 2022; Zhang et al., 2023a; Bansal & Sharma, 2023; Lowmanstone et al., 200 2023; Wadhwa et al., 2023; Honovich et al., 2023; Zheng et al., 2024; Ye et al., 2022a; Meng et al., 201 2022). Several studies have explored LLM-based annotation generation in zero-shot or few-shot task 202 settings (Ye et al., 2022a; Meng et al., 2022; Ye et al., 2022b), while others have compared pairs of 203 language models to assess the quality of annotations generated by these LLMs (Zheng et al., 2024). 204 However, these studies focused on generating annotations for NLP tasks such as sentiment analysis, 205 natural language inference (Gilardi et al., 2023; Alizadeh et al., 2023), or creating synthetic dialogues, 206 but only for dyadic conversations (Lee et al., 2023). Our approach complements these previous 207 studies by focusing on generating annotations of user opinions in complex multi-user conversations.

208 209

210

215

3 USDC DATASET CURATION

2113.1Collection of Reddit Conversation Threads

Initial crawl. We crawl a year (2019) of multi-user conversation data from 22 subreddits of Reddit
 using praw API². This dataset includes submissions and all associated user comments. Each

²https://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".

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 shown 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 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, as shorter conversations (fewer than 20 comments) are insufficient for accurately gauging user opinions. 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. 3 in the Appendix B shows detailed subreddit level statistics.

244 245

226

227

228 229

230

231

232

3.2 Obtaining LLM Annotations

246 247

Representing Reddit conversations in JSON format. To create the prompt, we follow the nested 248 hierarchical structure of Reddit conversations to maintain the context. Specifically, we maintain a 249 JSON structure for each conversation, where each author has their post IDs, and comments or replies 250 are available in the body section. An example of a Reddit conversation in JSON format is provided in 251 Appendix E. Note that the JSON explicitly includes the top-2 authors who posted the most comments in the conversation, and their respective post IDs. Our emphasis on these top-2 users (covering 253 47% posts of total posts on average) aimed at accurately assigning Stance and Dogmatism labels, 254 acknowledging the challenge of modeling a user's opinion belief based on a very limited number of posts within a conversation. 255

256 Using LLMs as human-like annotators. To annotate the stance of a user towards a submission at 257 each individual post and to assess the overall level of dogmatism expressed by the user throughout 258 the conversation, we employ two well-known commercialized API-based LLMs: GPT-4 (OpenAI, 259 2023) and Mistral Large (Jiang et al., 2024). OpenAI GPT-4 is a decoder-based language model with 260 a context window of 32k to 128k tokens. Mistral Large features a context window of 32k tokens. 261 Additionally, we 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 262 discuss these limitations, along with the situations where LLMs are prone to errors, in Appendix F. 263

For both GPT-4 and Mistral Large, we supplied a system prompt that contains the definitions 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 C.
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
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 G.1, G.2, G.3 respectively.
Similarly, samples of generated outputs using Mistral Large in zero, one, and few-shot settings are
shown in Appendix G.4, G.5, G.6 respectively.

274 Annotation tasks. We prompt the LLMs to perform two annotation tasks: 1) Stance detection, which 275 determines if a user comment or post is Strongly In Favor, Strongly Against, Stance Not Inferrable, 276 Somewhat In Favor, or Somewhat Against towards specific subreddit submission content; Our 5-class 277 stance detection scheme is inspired by the SPINOS dataset proposed by Sakketou et al. (2022). 278 These labels provide a fine-grained analysis similar to sentiment labels, allowing for a more detailed 279 understanding of user opinions. 2) Dogmatism identification, which evaluates the user's overall 280 opinion in conversation and categorizes them into one of four categories: Firm but Open, Open to Dialogue, Flexible or Deeply Rooted. Our 4-class dogmatism task is inspired by Fast & Horvitz 281 (2016), where the authors reported ratings that correspond to each level of dogmatism. We have 282 adopted similar definitions for dogmatism labels and incorporated them into our system prompts to 283 ensure consistency and accuracy in our annotations. This assessment reveals whether users are open 284 to changing their beliefs or remain steadfast in their opinions based on interactions with other users. 285

- Addressing failures in JSON parsing of LLM response. Sometimes, LLMs get confused with the author IDs and miss their Stance labels (Fig. 3 (left)). Sometimes, there were minor errors in key naming ('label' vs 'body' in Fig. 3 (right)). We observed such errors in ~15 cases across LLM setting. We manually fixed JSON parsing errors and corrected author IDs for associated Stance labels.
- 290 **Majority voting conflict.** After obtaining six annotations ({Mistral Large, GPT-4}×{zero, one, 291 and few-shot}) for each sample, we follow the two step process to obtain final gold annotations. (i) 292 Majority voting: we aggregate using majority voting (i.e label that appears most frequently across 293 models) to determine the final gold annotations for the Stance and Dogmatism tasks. (ii) Handling situations with no clear majority: when generating annotations using both GPT-4 and Mistral Large, 294 it is possible that the two models might provide different annotations for the same conversation. In 295 these cases, we use the annotation provided by GPT-4 in the few-shot setting as the deciding factor or 296 "gold standard". We chose to prioritize GPT-4 few-shot annotations because human annotations have 297 better IAA agreement with GPT-4 few-shot. Further, few-shot models, which are fine-tuned with a 298 small amount of task-specific data, often provide more accurate and contextually relevant annotations. 299

Class distributions for stance task is as follows: 3117 (somewhat in favour), 2266 (stance not inferrable), 1998 (somewhat against), 1303 (strongly against) and 640 (strongly in favor). For dogmatism task, the distribution is as follows: 666 (open to dialogue), 653 (firm but open), 140 (deeply rooted), and 69 (flexible). We present the class distributions obtained from each model with the 3 settings (zero, one, and few-shot) for both the tasks in Figs. 4 and 5, respectively, in Appendix B.

305 306

3.3 INTER-ANNOTATOR AGREEMENT WITH LLMS AS ANNOTATORS

As the quality of labeling on subjective tasks is challenging, we validate the inter-annotator agreement (IAA) between the two LLMs in three settings (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 and 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. 6 in the Appendix B shows pairwise Cohen's kappa values for both 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 settings. 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). This suggests that LLMs can be considered as competent annotators for complex subjective tasks.

324 3.4 USDC TEST DATASET EVALUATION WITH HUMAN LABELS

Due to the time-consuming nature of the manual annotation process, we perform human annotations on a set of 200 test conversations. In the forms for human annotations, we displayed the top 2 authors Reddit posts from the conversation, along with the submission title and content. We also provided a link to the original Reddit URL so that annotators could look at the full conversation. We provided detailed annotation guidelines (similar to the ones mentioned in the prompt in Appendix C) to instruct human annotators in carrying out these tasks.

332 With three human annotators on a sample of 200 conversations, as shown in Appendix O Fig. 20, we 333 achieved an inter-annotator agreement score of 0.49 for the stance detection and 0.50 for dogmatism tasks, indicating a reasonable level of consistency between human and LLM annotations. The 334 annotators included two males and one female, affiliated with academia and industry, aged between 335 20 and 40, who were very familiar with Reddit topics. We calculated the inter-annotator agreement 336 among the three human annotators themselves. Tables 12 and 13 in Appendix P report the IAA scores 337 for both stance and dogmatism tasks among the human annotators. The results showed an agreement 338 of 0.57 for the stance and 0.52 for the dogmatism. These findings demonstrate the level of consistency 339 among human annotators, providing a more comprehensive understanding of the alignment between 340 LLM-generated labels and human judgments.

341 342 343

344

4 TRAINING SMALL LANGUAGE MODELS (SLMS)

In this section, we briefly discuss the small language models that we experiment with. We also discuss their finetuning and instruction-tuning details. 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, 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 and instruction-tune these models using the proposed USDC dataset. We use pretrained model checkpoints from Hugging Face (Wolf et al., 2020). All of these LLMs have a context length of 4096 tokens. Model details and hyper-parameter settings are in Appendix H.

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 (~ 75%) and test splits (~ 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.

Finetuning of SLMs. For Stance classification, we treat each user post as an independent sample. In
contrast, for the 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, and we finetune the models by apply the LoRA technique (Hu et al.,
2021), with SFTT before saving the finetuned model. For finetuning, we used prompt for Stance
classification as shown in Fig. 7 (see Appendix D). Similarly, Fig. 8 (see Appendix D) 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 C). Similar to
 finetuning, we use same train-test splits for instruction-tuning.

368 369

370

5 Results

Baseline (un-fine-tuned) model performance and what constitutes a "reasonable" F1 score?
To establish a reasonable F1-score benchmark for fine-tuning and instruction-tuning (discussed in
the next subsections), we evaluated the un-fine-tuned SLMs, GPT-4 and Mistral Large, in few-shot
settings. This evaluation includes both stance and dogmatism tasks, using majority voting to enhance
reliability. The results are summarized in the Tables 4, 5, 6 and 7 in Appendix I. We make the
following observations: (i) Majority Voting generally provides a slight improvement over individual
few-shot configurations, which suggests the value of combining predictions from multiple models.
(ii) The difference between GPT-4 and Mistral Large in un-fine-tuned few-shot settings is relatively

		Stance Classification						Dogmatism Classification							
	Model		GPT-4		Mis	Mistral Large		Majority	GPT-4			Mistral Large			Majority
	Wouch	ZS	OS	FS	ZS	OS	FS	majority	ZS	OS	FS	ZS	OS	FS	majority
	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
0.0	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
nin	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
Stu	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
ine	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
щ	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
ğ	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
ini	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
ion-tu	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
nct	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
str	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
Ц	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

378	Table 1: Finetuning and instruction-tuning results: weighted F1-score for Stance and Dogmatism
379	classification using SLMs on USDC test set. ZS: Zero-shot, OS: One-shot, FS: Few-shot.

small, indicating that both models are fairly comparable in performance on these tasks when using the LLaMa-3-8B model.

As shown by the un-fine-tuned model's performance for stance classification in Table 6, an overall accuracy of 0.311 and F1 scores as low as 0.06 for certain classes, the baseline for this task is relatively low. Similarly, for dogmatism in Table 7, an overall accuracy of 0.40 and F1 scores as low as 0.00 for certain classes. In this context, an F1 score that significantly improves upon this baseline—especially if it approaches or exceeds 50%—could be considered reasonable.

400 Do SLMs finetuned with task-specific LLM annotations accurately perform Stance and Dogma-401 tism tasks on user opinions?

402 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 report AUC scores and 403 other qualitative analysis in Appendix J (Fig. 10 and 11). We make the following observations 404 from these results: 1) Compared to the baseline, while the un-fine-tuned models show moderate 405 performance, the fine-tuned models nearly double their F1 scores, particularly for the Stance task. 406 Even for dogmatism tasks, we saw better improvement in F1-score after fine tuning. 2) For both 407 tasks when finetuning, the majority voting labels as ground truth has a relatively high performance, 408 scoring above 50% weighted F1-score across several (7/7 for stance and 2/7 for dogmatism) models. 409 3) Finetuned LLaMa-3 models (LLaMA-3-8B and LLaMA-3-8B-instruct) perform better across both 410 tasks. 4) For GPT-4 annotations, in most cases, SLMs finetuned with few-shot annotations outperform 411 those trained with zero and one-shot annotations. For Mistral Large annotations, SLMs finetuned with 412 one-shot annotations perform the best. 5) Specifically, for the stance detection task, Vicuna-7B-v.1.5 413 finetuned using few-shot annotations is the best model trained with GPT-4 annotations. Similarly, 414 LLaMA-3-8B-instruct finetuned with one-shot annotations is the best model trained with Mistral Large annotations. 6) For the dogmatism detection task, LLaMA-3-8B-instruct finetuned using 415 few-shot annotations is the best model trained with GPT-4 annotations. Similarly, LLaMA-2-chat-7B 416 finetuned with one-shot annotations is the best model trained with Mistral Large annotations. 7) 417 Overall, we observe that instruction-tuned SLMs perform better than the pretrained SLMs. 418

419 Do SLMs instruction-tuned with task-specific LLM annotations perform better than SLMs 420 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 421 stance and dogmatism detection tasks on the USDC test set in Table 1. We report AUC scores and 422 other qualitative analysis in Appendix K (see Fig. 12). We make the following observations from 423 these results: 1) SLMs with instruction-tuning result in higher weighted F1-scores than SLMs with 424 finetuning for stance detection, while SLMs with finetuning outperform SLMs with instruction-tuning 425 in dogmatism detection. 2) Contrary to finetuning results, instruction-tuning results demonstrate 426 that using majority voting labels as ground truth, SLM instruction-tuning yields relatively high 427 performance only for the stance detection task, but not for the dogmatism detection. 3) Similar 428 to finetuning results, LLaMA-3 models (LLaMA-3-8B and LLaMA-3-8B-instruct) perform better 429 across both tasks. Additionally, GPT-4 annotations yield the best results in the few-shot setting, while 430 Mistral Large annotations perform best in the one-shot setting.

431

380 381 382

393

394

395

396

397

398

Overall, we draw the following conclusions when comparing SLM finetuning and instruction-tuning:
(1) Since dogmatism detection is inherently a more complex and varied 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.

438

439 440 QUALITATIVE ANALYSIS

441 Error Analysis. Table 2 illustrates the confusion matrix for stance detection for LLaMa-3-8B 442 finetuning and instruction-tuning. We make the following observations from this table: 1) For both finetuning and instruction-tuning, there is a significant misclassification between "Somewhat Against" 443 and "Somewhat In Favor," as well as between "Somewhat In Favor" and "Stance Not Inferrable." 444 These overlaps suggest challenges distinguishing moderate stances, indicating a need for enhanced 445 feature representation and clearer class definitions to improve model performance. We report the 446 confusion matrix for dogmatism detection task in Fig. 9 in the Appendix. It shows significant 447 misclassifications, especially for the "Deeply Rooted" and "Flexible" labels, with zero accuracy 448 and F1-scores. On the other hand, the model performs moderately better for "Firm but Open" and 449 "Open to Dialogue" classes with accuracies of 48.7% and 64.4%, respectively. The confusion matrix 450 also indicates substantial confusion to distinguish between intermediate levels of dogmatism, such 451 as "Firm but Open" and "Open to Dialogue". The area under the ROC curve (AUC) measures the 452 model's ability to distinguish between classes. Hence, we further report the ROC curve, which shows 453 the trade-off between the true positive rate (TPR) and false positive rate (FPR) for each class for stance and dogmatism tasks, see Figs. 10 and. 11 in Appendix J. 454

- 455 Lost in the Middle. To analyze the "lost in the middle" (Liu et al., 2024) phenomenon in our 456 LLM-based user-stance annotations, for a given user, we divided the data into time segments and 457 calculated inter-annotator agreement (IAA) using Cohen's Kappa scores across different models and 458 settings. The data was segmented based on the submission_id, author_id, and stance_id_timestamp. 459 For each group (i.e., each combination of submission_id and author_id), the timestamps were divided 460 into equal segments. The number of entries for each group was divided by the desired number of segments (3), and the division was done as evenly as possible, with each segment containing a roughly 461 equal number of time-stamped entries. Fig. 13 in Appendix reports the comparison statistics of IAA 462 scores for the stance detection task across initial, middle, and later time stamps. From Fig. 13, we 463 observe that the analysis across different time segments, especially when divided into three segments, 464 clearly demonstrates that the "lost in the middle" phenomenon is marginal. 465
- The partial decrease in inter-annotator agreement during the middle parts of the conversations suggests 466 that as conversations progress, models might face challenges in maintaining consistent agreement; 467 however, the decrease in agreement scores is minimal. The recovery in agreement towards the 468 final segments could indicate that as conversations start to conclude, they become more focused, 469 or that the models are better able to align on concluding statements. This trend underscores the 470 importance of considering segment-based analysis when evaluating model performance over long-471 form conversations. When comparing the model-generated annotations with human annotations, it 472 becomes evident that we do not encounter the "lost in the middle" problem. The human annotations 473 demonstrate a consistent level of inter-annotator agreement (IAA) across all three segments-initial, 474 middle, and final. This suggests that human annotators maintain a steady understanding and agreement 475 throughout the conversation, regardless of its length or complexity.
- 476 Recency Bias Phenomenon (Prior Context vs. Full Context). To investigate the impact of recency 477 bias (Peysakhovich & Lerer, 2023) on LLM performance in user-stance annotations, we focused on 478 verifying model annotations by examining the prior context for a given user, rather than considering 479 the entire conversation. The goal was to determine whether assessing each response within its 480 immediate context, followed by aggregation, would yield different results compared to analyzing 481 the full conversation context. Further details about the prior context annotations using LLMs are 482 discussed in Appendix O. Fig. 14 in the Appendix reports IAA scores, which contains a matrix of 483 Cohen's Kappa scores across different models and settings, including GPT-4 Few-Shot (FS), Mistral Large FS, Majority Voting, as well as GPT-4 FS PC and Mistral Large FS PC (here, PC denotes prior 484 context). From the figure, we observe that The agreement between GPT-4 FS and Majority Voting is 485 higher when the full conversation is considered (0.75) compared to when only prior context is used.

Table 2: 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.

			Pr	edicted	1					Pr	edicted	1	
		SOA	SOIF	SNI	SGA	SIF			SOA	SOIF	SNI	SGA	SIF
	SOA	151	132	34	44	2		SOA	143	125	37	54	4
	SOIF	93	537	113	17	14		SOIF	82	543	106	27	16
Actual	SNI	23	78	259	5	0	Actual	SNI	22	82	253	6	2
	SGA	52	35	13	115	17		SGA	41	35	11	131	14
	SIF	18	50	12	25	27		SIF	16	53	10	23	30

The agreement between GPT-4 FS PC and Mistral Large FS PC (both based on prior context) is lower than when using the full context, indicating that prior context alone may not capture all the necessary nuances for consistent annotation.

Human Agreement. The comparison of human annotations with models like GPT-4 FS and Mistral 498 Large FS shows that human annotators also rely heavily on the full conversation context to maintain 499 agreement. The results from this additional experiment, supported by the data in Fig. 14 in Appendix, 500 suggest that while prior context can provide some useful insights, it is not as effective as considering 501 the entire conversation context for maintaining high inter-annotator agreement. In summary, the 502 experiment highlights the importance of full context in LLM-based annotations and suggests that while 503 recency can influence model performance, it should be supplemented with the entire conversation 504 context to ensure higher accuracy and agreement. 505

Transfer Learning Evaluation of Models trained on USDC. To evaluate the quality of LLM-generated annotations, the annotators labeled 200 conversations and transfer learning is applied by fine-tuning the SLMs on the USDC dataset. We subsequently tested the model's performance on several existing stance datasets, including SPINOS (Sakketou et al., 2022), MT-CDS (Niu et al., 2024), and the Twitter stance dataset (Villa-Cox et al., 2020). We observe that performance of models trained using USDC is better or comparable to that of models trained using individual datasets themselves. Detailed results and analysis of results for the three datasets are reported in Appendix N.

513

514 6 DISCUSSION & CONCLUSION

515 We introduced USDC, a large-scale dataset of user stance and dogmatism in conversations, leveraging 516 LLMs as human-like annotators. This dataset is used for various applications, including analyzing 517 public opinions, enhancing dialogue systems, improving content moderation tools by identifying 518 and flagging dogmatic or polarizing users in online discussions, and generating dynamic contextual 519 user representations. The full-length multi-user conversation aspect of USDC allows it to capture 520 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 521 between SLMs and commercial LLMs for understanding user traits. While finetuning SLMs shows 522 good F1-score on both stance and dogmatism tasks, the F1-score remains below 60% (54.9% for 523 stance and 51.4% for dogmatism). On the other hand, instruction-tuning of SLMs only improves 524 F1-score performance on stance, not the dogmatism task. Further, the performance still falls short 525 of 60%, with weighted F1-scores of 56.2% for stance and 49.2% for dogmatism. These findings 526 indicate that there is still significant room for improvement in understanding user opinions from a text 527 segment. Human evaluation showed an agreement of 0.57 for the stance and 0.52 for the dogmatism 528 tasks between LLM and human annotations. This indicates that LLM-generated annotations in USDC 529 are close to human labels. Transfer-learning on 3 datasets also showed positive results.

530 **Limitations.** We plan to extend this work along the following directions in the future. 1) We would 531 like to extend this work to multi-lingual conversations and verify how accurately SLMs and LLMs 532 perform on the stance and dogmatism tasks in the multi-lingual scenario. 2) We analyzed user 533 dogmatism based on their posts within a single conversation. This approach could be extended to 534 include posts across multiple conversations and utilize similar profile information if available. 3) We analyzed dogmatism information for only the top two authors. Users with fewer comments often do 536 not provide enough information to accurately assess their stance or dogmatism, as many contribute 537 only one or two comments, which is insufficient to determine their overall opinion or dogmatic nature. Therefore, our study prioritizes the two most active users, who contribute approximately 50% of 538 the comments in each conversation, to better capture opinion fluctuations and provide a more robust analysis of stance and dogmatism.

540 REFERENCES 541

577

589

590

542	Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoory Kulshreshtha, Gauray Nemade, Yifeng Lu, et al. Towards a human-like open-
543 544	domain chatbot. <i>arXiv preprint arXiv:2001.09977</i> , 2020.
545	
546	Meysam Alizadeh, Mael Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Ko-
547	and approach chatget in text-appointed tasks arXiv preprint arXiv:2307.02179, 2023
548	
549	Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojo-
550	caru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al.
551	The falcon series of open language models. arXiv preprint arXiv:2311.16867, 2023.
552	Parikshit Bansal and Amit Sharma. Large language models as annotators: Enhancing generalization
553	of nlp models at minimal cost. arXiv preprint arXiv:2306.15766, 2023.
554	
555	Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. Plato: Pre-trained dialogue generation
556	model with discrete latent variable. arXiv preprint arXiv:1910.07931, 2019.
557	Tom Brown, Beniamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
550	Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are
560	few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.
561	Wai Lin Chiang, Zhuahan Li, Zi Lin, Ying Shang, Zhanghao Wu, Hao Zhang, Lianmin Zhang,
562	Sivuan Zhuang, Zhuohan Zhuang, Joseph F, Gonzalez, Jon Stoica, and Fric P Xing. Vicuna: An
563	open-source chatbot impressing gpt-4 with 90%* chatgpt quality. March 2023. URL https:
564	//lmsys.org/blog/2023-03-30-vicuna/.
565	
566	Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuazhi Wang, Mostafa Dahahani, Siddhartha Brahma, et al. Scaling instruction finatuned language
567	models <i>Journal of Machine Learning Research</i> 25(70):1–53, 2024
568	nodels. sournar of machine Learning Research, 25(10).1 55, 2021.
569	Jacob Cohen. A coefficient of agreement for nominal scales. Educational and psychological
570	<i>measurement</i> , 20(1):37–46, 1960.
571	Ethan East and Eric Horvitz Identifying dogmatism in social media: Signals and models In
572	Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp.
574	690–699, 2016.
575	
576	Joseph L Fielss. Measuring nominal scale agreement among many raters. <i>Psychological bulletin</i> , 76 (5):378–1071
	$(J_{1}, J_{1}, 0, 17/1$

- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for 578 text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30):e2305016120, 579 2023. 580
- 581 Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning 582 language models with (almost) no human labor. In Proceedings of the 61st Annual Meeting of the 583 Association for Computational Linguistics (Volume 1: Long Papers), pp. 14409–14428, 2023. 584
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, 585 et al. Lora: Low-rank adaptation of large language models. In International Conference on 586 Learning Representations, 2021. 587
- 588 Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. arXiv preprint arXiv:2310.06825, 2023.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris 592 Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 593 Mixtral of experts. arXiv preprint arXiv:2401.04088, 2024.

594 595 596	Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen White, and Sujay Jauhar. Making large language models better data creators. In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pp. 15349–15360, 2023.
598 599 600	Yupeng Li, Haorui He, Shaonan Wang, Francis CM Lau, and Yunya Song. Improved target-specific stance detection on social media platforms by delving into conversation threads. <i>IEEE Transactions on Computational Social Systems</i> , 2023.
601 602 603	Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. <i>Transactions of the Association for Computational Linguistics</i> , 12:157–173, 2024.
604 605 606 607	London Lowmanstone, Ruyuan Wan, Risako Owan, Jaehyung Kim, and Dongyeop Kang. Annota- tion imputation to individualize predictions: Initial studies on distribution dynamics and model predictions. <i>arXiv preprint arXiv:2305.15070</i> , 2023.
608 609 610	Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. <i>Advances in Neural Information Processing Systems</i> , 35:462–477, 2022.
611 612 613	Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. A challenge dataset and effective models for conversational stance detection. <i>arXiv preprint arXiv:2403.11145</i> , 2024.
614 615	R OpenAI. Gpt-4 technical report. arxiv 2303.08774. View in Article, 2(5), 2023.
616	Alexander Devenkhowich and Adam Lerer. Attention sorting combate recency bias in long context
617	language models. arXiv preprint arXiv:2310.01427, 2023.
618 619	Milton Rokeach. The nature and meaning of dogmatism. <i>Psychological Review</i> , 61(3), 1954.
620	Stephen Boller, Emily Dinen, Nemen Coyal, De Ju, Mery Williamson, Vinhen Liu, Jing Yu, Mula
621	Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain
622	chatbot. In Proceedings of the 16th Conference of the European Chapter of the Association for Com-
623	putational Linguistics: Main Volume, pp. 300–325, Online, 2021. Association for Computational
624	Linguistics.
626 627 628	Flora Sakketou, Allison Lahnala, Liane Vogel, and Lucie Flek. Investigating user radicaliza- tion: A novel dataset for identifying fine-grained temporal shifts in opinion. <i>arXiv preprint</i> <i>arXiv:2204.10190</i> , 2022.
629 630 631	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. <i>arXiv preprint arXiv:2302.13971</i> , 2023a.
632 633 634	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. <i>arXiv preprint arXiv:2307.09288</i> , 2023b.
635 636 637 638	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. <i>Advances in neural information processing systems</i> , 30, 2017.
639 640 641	Ramon Villa-Cox, Sumeet Kumar, Matthew Babcock, and Kathleen M Carley. Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. <i>arXiv preprint arXiv:2006.00691</i> , 2020.
642 643 644	Manya Wadhwa, Jifan Chen, Junyi Jessy Li, and Greg Durrett. Using natural language explanations to rescale human judgments. <i>arXiv preprint arXiv:2305.14770</i> , 2023.
645 646 647	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In <i>Proceedings of the 2020 Conference on Empirical Methods in</i> <i>Natural Language Processing: System Demonstrations</i> , pp. 38–45, 2020.

648 649 650	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 <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pp. 11653–11669, 2022a.
652 653	Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Progen: Progressive zero-shot dataset generation via in-context feedback. In <i>Findings of the Association</i> for Computational Linguistics: EMNLP 2022, pp. 3671–3683, 2022b
654 655	Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. Llmaaa: Making large language
657	models as active annotators. arXiv preprint arXiv:2510.19596, 2025a.
658 659 660	Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. <i>arXiv</i> preprint arXiv:2308.10792, 2023b.
661 662 663 664 665	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 <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations</i> , 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.
667 668 669	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. <i>Advances in Neural Information Processing Systems</i> , 36, 2024.
670	Vangahaa Zhau Andrai Jaan Muraganu Ziwan Han Vairan Dagtar Silviy Ditig Harris Chan
671 672	and Jimmy Ba. Large language models are human-level prompt engineers. <i>arXiv preprint</i>
672	arXiv:2211.01910, 2022.
674	
675	
676	
677	
678	
679	
680	
681	
682	
683	
684	
685	
686	
687	
688	
689	
690	
691	
692	
693	
694	
695	
696	
697	
698	
699	
700	
701	

702	А	OVERVIEW OF APPENDIX SECTIONS
703		Section R: Datailed Statistics of the USDC Dataset
705		• Section B. Detailed Statistics of the USDC Dataset
706		Section C: System Prompt for LLM Annotation
707		Section D: Prompts for Finetuning SLMs
708 709		Section E: Sample of User Input Prompt
710		Section F: Situations Leading to LLM Annotation Errors and Inconsistencies
711		Section G: Samples of JSON Outputs from LLMs
712		Section H: Details of small language models and Hyper-parameter settings
714		• Section I: Baseline (un-fine-tuned) model performance
715		• Section I: SI M Einstuning: AUC (Area Under the Curve) Analysis
716		• Section J. SLW Finetuning. AUC (Alea Under the Curve) Analysis
717		 Section K: SLM instruction-tuning: AUC (Area Under the Curve) analysis
718 719		Section L: "Lost in the Middle" Analysis
720		Section M: Recency Bias Analysis
721		Section N: SLM finetuning: Transfer Learning Performance
722		• Section O: Individual user responses within their specific context vs. entire conversation at
724		once for stance and dogmatism
725		• Section P: Inter-Annotator Agreement (IAA) between human annotators
726		• Section O Robustness analysis of Human-LLM Annotations
727		
728 729		• Section R Qualitative examples demonstrating cases with high, moderate, and low inter- annotator agreement (IAA)
730		
731	п	
732	В	DETAILED STATISTICS OF THE USDU DATASET

Table 3 shows the detailed statistics of our USDC dataset at the subreddit level. Fig. 4 shows the distribution of stance labels across LLM annotations across zero-shot, one-shot, and few-shot settings.
Fig. 5 shows the distribution of dogmatism labels across LLM annotations across zero-shot, one-shot, and few-shot settings.

739				
740	subreddit	num_conversations	min_total_token_count	max_total_token_count
741	DebateCommunism	73	529	11557
	Abortiondebate	70	1271	7401
742	CapitalismVSocialism	61	665	16927
743	prochoice	60	582	7278
	brexit	56	637	4553
744	climateskeptics	56	734	7550
745	prolife	54	672	13342
= 4.0	gunpolitics	52	683	7889
746	MensRights	52	623	5774
747	climatechange	49	520	7427
	nuclear	41	572	5282
748	progun	39	436	3632
749	NuclearPower	23	629	4589
	Vegetarianism	22	627	3958
750	AntiVegan	20	351	5052
751	climate	13	701	4678
	Egalitarianism	10	665	4060
752	VeganActivism	8	460	3685
753	Veganism	2	1332	1738
	AnimalRights	1	845	845
/54	animalwelfare	1	1363	1363
755	GunsAreCool	1	2945	2945



Figure 4: Distribution of Stance labels across LLM annotations in six settings: GPT-4, Mistral Large×Zero-shot, One-shot, Few-shot. Somewhat In Favor is the most frequent class across all six settings, while Strongly In Favor is the least frequent.



Figure 5: Distribution of dogmatism labels across LLM annotations in six settings: GPT-4, Mistral Large×Zero-shot, One-shot, Few-shot. Open to Dialogue is the most frequent class across all six settings, while Flexible is the least frequent.



Figure 6: Inter-annotator agreement (IAA): Cohen's Kappa score across six different settings (2 models×3 settings) for Stance (left) and Dogmatism (right) tasks.

C SYSTEM PROMPT FOR LLM ANNOTATION

We used the following system prompt as annotation guidelines both to obtain annotations from LLMs and for the instruction-tuning of SLMs.

832	""
833	### Introduction
834	**Objective**: Analyze Reddit conversations to identify the stance
835	of specific authors on sociopolitical topics and determine their
836	level of dogmatism.
837	author's standpoint and judgement towards a given topic.
838	**Dogmatism Definition**: Dogmatism is an opinion strongly believed
839	as a fact to support a stance without a question or allowance
840	for conversation.
841	**Task**: Given a JSON formatted Reddit submission and its comment
842	'author1' and 'author2' by assigning one of the following
843	five predefined stance labels: 'strongly_against',
844	`somewhat_against', `somewhat_in_favor', `strongly_in_favor',
845	'stance_not_inferrable'. Also, assign a dogmatism label for each
846	author by assigning one of the following four predefined labels: 'Deeply Rooted', 'Firm but Open', 'Open to Dialogue', 'Flexible'
847	beepry nooced, film but open, open to blatogue, fickible.
848	### Description of Stance Labels:
849	1. **strongly_against / strongly_in_favor**: Marks text showing
850	strong opinions, emotional expressions, or argumentative tones.
851	openness to discussion, less certainty, or showing interest in
852	different viewpoints.
853	<pre>3. **stance_not_inferrable**: Use for texts that are neutral,</pre>
854	support both stances, or where the stance is unclear despite
855	being on-topic.
856	### Description of Dogmatism Labels:
857	1. **Deeply Rooted**: Reflects a strong, unchangeable belief. This
858	label conveys the idea of someone who is firm in their opinion
859	and unlikely to be swayed.
860	2. **Firm but Open**: Indicates a person who is not likely to
861	authoritatively. It captures the essence of being steadfast in
862	one's beliefs without being dismissive of others.
863	3. **Open to Dialogue**: Describes someone who holds a certain opinion but is genuinely interested in considering other

```
864
             viewpoints. This label suggests a willingness to engage in
865
             meaningful conversation about differing perspectives.
866
         4. **Flexible**: Denotes a person who is not firmly committed to
867
             their stance and is open to changing their opinion. This label
             is indicative of flexibility and openness to new information or
868
             arguments.
869
870
         ### Input Data Format
871
         The input data will be in JSON format and will include several key
872
             elements to represent a Reddit submission and its associated
             comments. Each element provides specific information as
873
             described below:
874
875
         - 'id': This is the unique identifier for the Reddit submission.
876
         - 'title': The title of the post. This is what users see first and
             often summarizes or hints at the content of the submission.
877
            'content': The main post's detailed description. This text
878
             segment provides the core message or information the author
879
             wishes to communicate with the Reddit community. It may include
880
             narratives, questions, or any information relevant to the title.
881
           'comments': An array (list) of comments related to the Reddit
882
             submission. Each comment in this array includes the following
             fields:
883
            - 'id': The unique identifier for the comment, allowing for
884
                identification and reference within the dataset.
885
              'author1' or 'author2': The username of the comment's author,
886
                if it is made by one of our focus authors. This helps in
887
                tracking contributions by specific individuals.
            - 'body': The text of the comment. This is the main content of
888
                the comment where the author responds to the post or another
                comment, providing insights, opinions, or further information.
890
            - 'replies': An array of comments that are direct responses to
891
                this comment. The structure of each reply follows the same
                format as the initial comment, including 'id', 'author1' or
892
                 'author2' (if applicable), 'body', and potentially more
893
                 'replies'.
894
895
         ### Output Data Format
896
         Submit your annotations in JSON format, grouping all stance
             annotations under the key ``stance_annotations''. Each entry
897
             should be a dictionary containing the segment's ``id'', your
898
             ``label'', and the ``reason'' for your choice. Include the
899
             dogmatism label and its justification under ``dogmatism_label''
900
             and ''dogmatism_reason'' keys, respectively.
901
         The output should follow this structure:
902
          ```json
903
904
 "author1": {
905
 "stance_annotations": [
906
 {
 "id": "[segment_id]",
907
 "label": "[chosen_label]",
908
 "reason": "[Justification in <50 words]"
909
 },
910
 • • •
911
 1.
 "dogmatism_label": "[chosen_dogmatism_label]",
912
 "dogmatism_reason": "[Justification in <50 words]"
913
 },
914
 "author2": {
915
 "stance_annotations": [
916
 "id": "[segment_id]",
917
 "label": "[chosen_label]",
```

```
918
 "reason": "[Justification in <50 words]"
919
 },
920
 . . .
],
921
 "dogmatism_label": "[chosen_dogmatism_label]",
922
 "dogmatism_reason": "[Justification in <50 words]"
923
 }
924
 }
 , , ,
925
 ### Instructions for Effective Annotation
926
927
 1. **Labeling Stance**: For each segment (including the original
928
 Reddit submission, comments, or replies) where "author1" or
929
 "author2" is mentioned, assign a stance label that best
930
 represents the stance expressed towards the discussed topic in
 the submission. This comprehensive approach ensures no relevant
931
 contribution by "author1" or "author2" is overlooked. Evaluate
932
 the stance based on the content's tone, argumentation, and
933
 engagement level with the topic.
934
 2. **Providing Justification**: For each label assigned, include a
935
 concise reason, aiming for less than 50 words. Focus on the
 stance and argumentative indicators present in the text.
936
 3. **Dogmatism Assessment**: After reviewing all segments from
937
 "author1" and "author2", assign a single dogmatism label
938
 reflecting the overall tone and approach in their contributions.
939

940
```

# D PROMPTS FOR FINETUNING SLMS

Fig. 7 and 8 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 7: Prompt for stance classification, for finetuning SLMs.

941 942

943 944

945

946 947

948

949

950

951 952

953

954

955

956 957

958

959 960 961

#### 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 8: Prompt for dogmatism classification, for finetuning SLMs.

# E SAMPLE OF USER INPUT PROMPT

```
.....
991
 Now complete the given task for the respective authors i.e., author1
992
 respective ids are ['dhoxyz', 'f3pghji', 'f3tywb4', 'f3uomn2'].
993
 author2 respective ids are ['f3rt0bf', 'f3rqu2u'] for the data in
994
 json format
995
 "id":"dhoxyz",
996
 "title":"This sub should encourage anti vs. pro-gun discussions
997
 instead of shutting them down instantly",
998
 "content":"Honesly, I followed this sub especifically to take part in
999
 these discussions, but everytime I see a comment that even
1000
 remotely suggests anti gun ideals or a discussion on the subject
 just gets ignored and downvoted to hell. Kind of expecting this to
1001
 go the same way (my karma anus is ready, downvotes) , but I have
1002
 to hope for healthy discussions on the subject.",
1003
 "comments":[
1004
 {
 "id":"f3p9n2c",
1005
 "body":"I think the problem now is the two sides are at an
1006
 impasse. Everytime there is a "compromise" pro gun loses
 something. Now days pro gun is interpreting the Constitution
1008
 more literal, which leaves even the most mild policies of
1009
 anti gun as infringements. To further compound this anti gun
1010
 is only considering the most extreme measures. "Assault
 Weapons" bans, mandatory buybacks, red flag laws, etc.. I
1011
 think at this point there is just nothing left to talk about
1012
 . The middle ground is gone.",
1013
 "replies":[
1014
 {
 "id":"f3pati9",
1015
 "replies":[
1016
 {
1017
 "id":"f3pdu44",
1018
 "body":"You are exactly right. I'm done with the
1019
 idea that there can be real compromise. We
1020
 should have at least gotten national reciprocity
 and shall-issue in every state in exchange for
1021
 what we've given up. Now you have to be a
1022
 goddamn lawyer to exercise your rights without
1023
 violating the law."
1024
 },
1025
 "id":"f3rt0bf",
```

```
1026
 "body":"I am prepared for UBCs, if they do this:
1027
 1. Lower the age to buy handguns to 18, nationwide.
1028
 2. Repeal the Hughes Amendment:
 3. A FOPA-like ban on assault weapon bans (what the
1029
 FOPA did with a registry)
1030
 4. The punishment for violation is a monetary fine
1031
 only
1032
 5. A repeal of the GCA ban on foreign NFA weapons
1033
 6. A repeal of the National Minimum Drinking Age Act
1034
 of 1984"
 }
1035
]
1036
 },
1037
 {
1038
 "id":"f3pd55z",
 "body":"Everytime there is a "compromise" pro gun loses
1039
 something. That and today's compromise is tomorrow's
1040
 loophole to be closed. All such compromises do is push
1041
 that policy off until the next round."
1042
 }
1043
 1
1044
 },
 {
1045
 "id":"f3paf0j",
1046
 "body": "Yeah this sub it's not conducive to conversion. Its
1047
 quickly devolving to little more than "Boogaloo" memes and
1048
 shouting "SHALL. NOT." at each other. However, as far as I
1049
 know, the mods won't delete your thread and ban you from the
 sub for trying to have a good faith discussion, like some
1050
 of the gun control subs will.",
1051
 "replies":[
1052
 {
1053
 "id":"f3pusbm",
 "body":"Unfortunately this sub's mod team takes a very
1054
 passive approach to moderation. With very little
1055
 effort they could make this sub into a quality progun
1056
 meeting ground *without having to resort to
1057
 censorship*. Instead they promote low-effort memes and
1058
 endless duplication of posts through their inaction.
1059
 whubbard has the chops to resurrect this sub. Let's
 see if he's up to the challenge.",
1060
 "replies":[
1061
 {
1062
 "id":"f3q8xj6",
1063
 "body":"We voted to ban memes last week. All about
1064
 rolling it out now.",
 "replies":[
1065
 {
1066
 "id":"f3qn4p8",
1067
 "body": "Damn I might have to eat some crow
1068
 here then..."
1069
 1
1070
 }
1071
]
1072
 }
1073
 1
1074
 },
1075
 "id":"f3pafqa",
1076
 "body":"Found the gun grabber!!",
1077
 "replies":[
1078
 {
 "id":"f3pcw4h",
1079
 "body":"Witch hunter."
```

1080	
1081	}
1000	
1002	} <b>r</b>
1003	1 "id"."f3pa151"
1084	"body":"I see people have discussions when it makes sense to.
1085	Not much reason to spend time responding to the same gun
1086	control measures over and over though."
1087	},
1088	{
1089	"id":"f3paw3h",
1090	"body":"I get where you're coming from, but people's ability to
1091	protect themselves and own their own property isn't
1092	cease to own their own property. It's like breathing, there
1093	can be nothing less than total ability to breath when and
1094	how someone wants. It's just that simple."
1095	},
1096	{
1007	"id":"f3pax9m", "bod" """
1000	"body": "My take on this, what kind of open discussion is
1000	, not to be infringed upon? They're making all these
1100	unlawful laws to portray it as it's somehow legitimate. They
1100	are not, We are at an apex, to which both political
1101	spectrums and even us to a degree are liable for.\nI
1102	certainly believe both sides are waiting for this to boil
1103	over so each can finger point. I just speculate it's going
1104	to be the neri humanity been whispering about but never
1105	<pre>}.</pre>
1106	{
1107	"id":"f3pb6ny",
1108	"body":"The time for discussion is over."
1109	}, ,
1110	l "id"."f3pfawa"
1111	"body":"I don't know what you're talking about. Sure people
1112	downvote, but they also talk. We get "why do you need guns"
1113	posts at least weekly, and several people will engage in
1114	actual conversation with them, citing facts, clearing up
1115	statistics, and telling stories to illustrate why this is
1116	important to them, but they are usually met with "you stupid
1117	People who come here to discuss and learn will be treated
1118	well. People who are just trolling are treated like trolls
1119	• " ,
1120	"replies":[
1121	
1122	"ld":"I3pghjl", "body":"I mada thig negt bagayaa I'm alwaya gooing
1103	rational conversation seeking comments getting blown
1123	to downvote hell.",
1125	"replies":[
1125	{
1120	"id":"f3pi9xv",
1127	"body":"[Like this one?](https://www.reddit.com/r/progun
1120	/comments/uncusz/yup/isp/stg/)> One smart man in a sub full of welp "strong opinions" You start
1129	off with arrogance, as the sole arbiter of what
1130	constitutes a "smart man". Then you back it up with
1131	a dismissive swipe at what you term "strong opinions
1132	".> Every other country can see that PROPER gun
1133	control reduces gun violence by a ton, More
	arrogance. False equivalence. Unsupported claims.>

1134	but the US refuses to let go of it's antique laws In
1135	a shocking turn of events, more arrogance. > Fully
1136	aware that this is a fully progun sub, willing to
1137	take the downvotes in order to spark a discussion
1138	and crack some heads. You aren't the first arrogant
1120	asshole to grace this sub with posts like this. Try
1133	bringing something other than your own self-
1140	importance to the discussion. Edit: And then there's
1141	[this gem](https://www.reddit.com/r/
1142	unpopularopinion/comments/d3w5z1/
1143	people_living_in_the_us_are_living_in_one_of_the/
1144	IU6735g/.> Wanna Ieel like you could be shot at
1145	every single moment: Move to the US, it if prob
1146	shot by a random citizen (sometimes police)."
1147	},
1148	{
1149	"id":"f3pj8k0",
1150	"body":"As is tradition. We're done with that
1150	condescending bullshit from antis, you dont come
1101	here for good faith discussion and whether you
1152	get a reasonable response or not, nothing ever
1153	because we get the same treatment anytime we
1154	attempt to speak out in anti subs."
1155	},
1156	{
1157	"id":"f3plgf4",
1158	"body":"If downvotes hurt your feelings, you shouldn
1159	't be on reddit. People tend to downvote
1160	anything they disagree with (which is why some
1161	subs specifically ask you to only downvole
1162	) It's a had habit but that's the way it is
1162	People downvote and *still* enage. You want to
1167	post a view contrary to the prevailing view of
1104	the sub, take your lumps and participate in what
C011	conversation you are offered. But if you're
1166	only here to preach about how stupid, misguided,
1167	unevolved, uneducated, irrational, and/or
1168	VIOLENC WE ALE, don t expect a polite response.
1169	
1170	"id":"f3tcgf1",
1171	"body":"An arrogant Israeli trying to tell another
1172	nation how they should be run. You're just a
1173	walking stereotype aren't you? And before you
1174	say anything, I popped into your comment history
1175	. Inal's where the calling you israeli comes
1176	"replies".[
1177	
1178	"id":"f3tywb4",
1179	"body":"I thought that trying to tell other
1180	nations how they should run was your guys'
1101	s stereotype.",
1101	"replies":[
1182	נ "id"•"לאוטעאמ"
1183	"body":"No we do in and try to make them
1184	work our way."
1185	}
1186	]
1187	}

90 91		
92	},	
93	{	
94		"id":"f3pzseh", "Nadu":"Ittla a littla unfauturata hut the muchbaus who same an
95		bere tend to be intellectually disbonest and/or uninformed
96		There was some Australian post a few days ago that pretty
97		much asked why we like our guns more than children. No
98		discussion to be had there. There's also some posts that
99		clearly demonstrate the poster should inform himself or
00	}.	neiseil a little.
01	{	
02		"id":"f3rqu2u",
03		"body":"Actually, do that. It shows everyone that they tend to
04		be crazy, unstable, ignorant, stereotyping, arrogant
05	},	bastaids who hate black people with a half thigger.
06	{	
07		"id":"f3t7tgg",
08		"body":"Welcome to reddit, home of every single safe place for
09		better than r/politics"
10	},	Social chan 1, polloios
11	{	
12		"id":"f3unt9z",
13		"body":"This isn't r/gundebate. This is a pro gun subreddit.
14		.",
15		"replies":[
16		{
17		"id":"f3uomn2", "body","Seelly the sourcester is mostly double "
18		"body":"Sadiy tho, r/gundebate is pretty dead"
19		]
20	},	
21	{	
22		"id":"f4dip6o",
23	ı	"body":"Anything else you want to give away for free?"
24	1	
25	}	

# F SITUATIONS LEADING TO LLM ANNOTATION ERRORS AND INCONSISTENCIES

1229

1230 1231

1236

1237

1238

1239

Before proceeding with LLM annotation using larger models, we first tested other versions of GPT
and Mistral models, such as GPT-3.5 and Mistral-small and medium. However, we found that these
models failed to produce annotations in the desired format. Below are some specific situations where
LLMs were prone to errors:

- System Prompt Clarity: The importance of a clear and precise system prompt cannot be overstated. When the prompt lacked clarity, LLMs often generated annotations for unspecified authors, indicating confusion about the task requirements.
- Understanding Conversation Structure: Without providing a clear example of the conversation structure, none of the LLMs were able to understand the task properly. This demonstrates the need for explicit guidance when dealing with complex conversation data.

1242

1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254 1255

1256 1257

1258

1259

- **Interface Issues:** Using an interface to facilitate LLM annotation proved problematic. After processing 2 to 3 examples, LLMs began providing annotations for previous user IDs, even when presented with new conversations. This suggests that the model lost track of the task and context.
  - **Consistency in Annotations:** For smaller conversations, different LLMs tended to produce similar annotations. However, as the conversations grew longer, the annotations became inconsistent across different models, indicating challenges in maintaining accuracy over extended discourse.
    - **Confusion with Author IDs:** Occasionally, LLMs confuse author IDs, resulting in missed stance labels for certain authors (as shown in Fig. 3 (left) in the main paper). Additionally, there were minor errors in key naming (e.g., 'label' vs. 'body' as shown in Fig. 3 (right) in the main paper), which further highlighted the model's limitations.
- G SAMPLES OF JSON OUTPUTS FROM LLMS

#### G.1 SAMPLE JSON OUTPUT FROM LLM: GPT-4 ZERO-SHOT

```
1260
 {
1261
 "author1": {
1262
 "stance_annotations": [
1263
 {
1264
 "id": "dhoxyz",
 "label": "somewhat_in_favor",
1265
 "reason": "Advocates for open discussions, showing interest in
1266
 dialogue."
1267
 },
1268
 {
1269
 "id": "f3pghji",
 "label": "somewhat_in_favor",
1270
 "reason": "Expresses desire for rational discussions, despite
1271
 downvotes."
1272
 },
1273
 {
1274
 "id": "f3tywb4",
 "label": "somewhat_against",
1275
 "reason": "Sarcastic retort, but lacks strong emotional language
1276
 . "
1277
 },
1278
 {
1279
 "id": "f3uomn2",
1280
 "label": "stance_not_inferrable",
 "reason": "Comment is factual, no clear stance on the debate
1281
 itself."
1282
 }
1283
 1,
1284
 "dogmatism_label": "Open to Dialogue",
 "dogmatism_reason": "Shows interest in discussion, acknowledges
1285
 downvotes."
1286
 },
 "author2": {
1288
 "stance_annotations": [
1289
 {
 "id": "f3rt0bf",
1290
 "label": "somewhat_in_favor",
1291
 "reason": "Proposes conditions for UBCs, indicating some openness
1292
 . "
1293
 },
1294
 {
 "id": "f3rqu2u",
1295
 "label": "strongly_against",
```

```
1296
 "reason": "Uses strong negative descriptors for opponents,
1297
 indicating firm stance."
1298
 }
1299
],
 "dogmatism_label": "Firm but Open",
1300
 "dogmatism_reason": "Mixed; proposes solutions but also dismisses
1301
 opposition."
1302
 }
1303
```

### 1306 G.2 SAMPLE JSON OUTPUT FROM LLM: GPT-4 ONE-SHOT

1304 1305

```
1308
 {
1309
 "author1": {
 "stance_annotations": [
1310
 {
1311
 "id": "dhoxyz",
1312
 "label": "somewhat_in_favor",
 "reason": "Advocates for open discussions, indicating a pro-
1313
1314
 discussion stance."
 },
1315
 {
1316
 "id": "f3pghji",
1317
 "label": "somewhat_in_favor",
1318
 "reason": "Expresses desire for rational conversation, showing a
1319
 supportive stance for dialogue."
1320
 },
 {
1321
 "id": "f3tywb4",
1322
 "label": "somewhat_against",
1323
 "reason": "Uses sarcasm to counter a stereotype, indicating a
1324
 stance against the opposing view."
 },
1325
 {
1326
 "id": "f3uomn2",
1327
 "label": "somewhat_in_favor",
1328
 "reason": "Expresses disappointment in lack of active debate,
 supporting open discussion."
1329
 }
1330
 1,
1331
 "dogmatism_label": "Open to Dialogue",
1332
 "dogmatism_reason": "Shows a consistent preference for discussion and
1333
 engagement, without dogmatic assertions."
1334
 },
 "author2": {
1335
 "stance_annotations": [
1336
 {
1337
 "id": "f3rt0bf",
1338
 "label": "somewhat_in_favor",
 "reason": "Lists conditions for agreement, showing openness to
1339
 compromise."
1340
 },
1341
1342
 "id": "f3rqu2u",
1343
 "label": "strongly_against",
 "reason": "Uses strong language to describe opponents, indicating
1344
 a firm stance against them."
1345
 }
1346
],
1347
 "dogmatism_label": "Firm but Open",
1348
 "dogmatism_reason": "While showing strong opinions, also lists
1349
 specific conditions for compromise."
```

```
1350
 }
1351
1352
1353
 G.3 SAMPLE JSON OUTPUT FROM LLM: GPT-4 LARGE FEW-SHOT
1354
1355
 {
 "author1": {
1356
 "stance_annotations": [
1357
 {
1358
 "id": "dhoxyz",
1359
 "label": "somewhat_in_favor",
 "reason": "Expresses desire for open discussions, indicating a
1360
 stance in favor of dialogue."
1361
 },
1362
 {
1363
 "id": "f3pghji",
1364
 "label": "somewhat_in_favor",
1365
 "reason": "Advocates for rational conversation despite observing
 negative reactions."
1366
 },
1367
 {
1368
 "id": "f3tywb4",
1369
 "label": "somewhat_against",
1370
 "reason": "Uses sarcasm to counter a stereotype, showing a stance
 against it."
1371
 },
1372
 {
1373
 "id": "f3uomn2",
1374
 "label": "stance_not_inferrable",
 "reason": "Comment on the state of another subreddit does not
1375
 reveal a stance on the gun debate."
1376
 }
1377
],
1378
 "dogmatism_label": "Open to Dialogue",
1379
 "dogmatism_reason": "Shows a consistent
1380
 preference for discussion over confrontation."
1381
 },
 "author2": {
1382
 "stance_annotations": [
1383
 {
1384
 "id": 'f3rt0bf",
 "labe"': 'somewhat_in_favor",
1385
 "reason": 'Lists conditions for accepting UBCs, indicating some
1386
 openness to compromise."
1387
 },
1388
1389
 "id": 'f3rqu2u",
1390
 "label": "strongly_against",
 "reason": "Uses derogatory language to describe opponents,
1391
 showing strong opposition."
1392
 }
1393
 1,
1394
 "dogmatism_label": "Firm but Open",
1395
 "dogmatism_reason": "While showing strong opinions, also suggests
 conditions for policy acceptance."
1396
 }
1397
1398
1399
1400
 G.4 SAMPLE JSON OUTPUT FROM LLM: MISTRAL LARGE ZERO-SHOT
1401
```

"author1": { "stance\_annotations": [

1402

1403

{

```
1404
 {
1405
 "id": "dhoxyz",
1406
 "label": "somewhat_in_favor",
 "reason": "Author expresses desire for open discussions on anti-
1407
 gun ideals."
1408
 },
1409
 {
1410
 "id": "f3pghji",
1411
 "label": "stance_not_inferrable",
 "reason": "Author is commenting on downvotes, not expressing
1412
 stance."
1413
 },
1414
 {
1415
 "id": "f3tywb4",
1416
 "label": "stance_not_inferrable",
 "reason": "Author is making a stereotypical comment, not
1417
 expressing stance."
1418
 },
1419
 {
1420
 "id": "f3uomn2",
1421
 "label": "stance_not_inferrable",
 "reason": "Author is commenting on the state of another subreddit
1422
 , not expressing stance."
1423
 }
1424
],
1425
 "dogmatism_label": "Open to Dialogue",
1426
 "dogmatism_reason": "Author expresses desire for open discussions and
1427
 does not dismiss opposing views."
 },
1428
 "author2": {
1429
 "stance_annotations": [
1430
 {
1431
 "id": "f3rt0bf",
 "label": "strongly_in_favor",
1432
 "reason": "Author lists conditions for accepting UBCs, showing
1433
 strong pro-gun stance."
1434
 },
1435
 {
1436
 "id": "f3rqu2u",
 "label": "strongly_in_favor",
1437
 "reason": "Author"s comment is dismissive of anti-gun advocates,
1438
 showing strong pro-gun stance."
1439
 }
1440
 1,
1441
 "dogmatism_label": "Deeply Rooted",
 "dogmatism_reason": "Author"s comments are dismissive of opposing
1442
 views and show a strong, unchangeable belief in pro-gun ideals."
1443
 }
1444
1445
1446
1447
 G.5 SAMPLE JSON OUTPUT FROM LLM: MISTRAL LARGE ONE-SHOT
1448
```

```
1449
 {
1450
 "author1": {
1451
 "stance_annotations": [
1452
 {
 "id": "dhoxyz",
1453
 "label": "somewhat_in_favor",
1454
 "reason": "Expresses desire for open discussions on gun control,
1455
 but follows pro-gun subreddit."
1456
 },
1457
 {
 "id": "f3pghji",
```

```
1458
 "label": "stance_not_inferrable",
1459
 "reason": "Comments on downvotes, but no clear stance on gun
1460
 control."
1461
 },
 {
1462
 "id": "f3tywb4",
1463
 "label": "stance_not_inferrable",
1464
 "reason": "Comment about stereotypes, but no clear stance on gun
1465
 control."
1466
 },
 {
1467
 "id": "f3uomn2",
1468
 "label": "stance_not_inferrable",
1469
 "reason": "Mentions r/gundebate being dead, but no clear stance
1470
 on gun control."
 }
1471
],
1472
 "dogmatism_label": "Open to Dialogue",
1473
 "dogmatism_reason": "Seeks open discussions and does not express
1474
 strong dogmatic views."
1475
 },
 "author2": {
1476
 "stance_annotations": [
1477
 {
1478
 "id": "f3rt0bf",
1479
 "label": "somewhat_in_favor",
1480
 "reason": "Lists conditions for supporting UBCs, showing openness
1481
 to compromise."
 },
1482
 {
1483
 "id": "f3rqu2u",
1484
 "label": "strongly_in_favor",
1485
 "reason": "Strong negative language against gun control advocates
1486
 , suggesting strong pro-gun stance."
 }
1487
 1,
1488
 "dogmatism_label": "Firm but Open",
1489
 "dogmatism_reason": "Expresses strong pro-gun views but open to
1490
 compromise under certain conditions."
1491
 }
 }
1492
```

G.6 SAMPLE JSON OUTPUT FROM LLM: MISTRAL LARGE FEW-SHOT

1493 1494

```
1496
 {
 "author1": {
1497
 "stance_annotations": [
1498
 {
1499
 "id': "dhoxyz",
1500
 "label": "somewhat_in_favor",
 "reason": "Expresses desire for open discussions on anti vs. pro-
1501
 qun debates."
1502
 },
1503
 {
1504
 "id": "f3pghji",
1505
 "label": "stance_not_inferrable",
 "reason": "Comments on downvotes, but no clear stance on the
1506
 topic."
1507
 },
1508
 {
1509
 "id": "f3tywb4",
1510
 "label": "stance_not_inferrable",
 "reason": "Makes a stereotypical comment, but no clear stance on
1511
 the topic."
```

```
1512
 },
1513
 {
1514
 "id": "f3uomn2",
 "label": "stance_not_inferrable",
1515
 "reason": "Mentions r/gundebate being dead, but no clear stance
1516
 on the topic."
1517
 }
1518
],
1519
 "dogmatism_label": "Open to Dialogue",
 "dogmatism_reason": "Seeks open discussions and engages in
1520
 conversation without strong dogmatic language."
1521
 },
1522
 "author2": {
 "stance_annotations": [
1524
 {
 "id": "f3rt0bf",
1525
 "label": "somewhat_in_favor",
1526
 "reason": "Lists conditions for supporting UBCs, showing openness
1527
 to discussion."
1528
 },
1529
 {
 "id": "f3rqu2u",
1530
 "label": "strongly_against",
1531
 "reason": "Uses derogatory language to express strong opposition
1532
 to 'grabbers'."
1533
 }
1534
 1,
 "dogmatism_label": "Firm but Open",
1535
 "dogmatism_reason": "Expresses strong opinions but also shows
1536
 willingness to consider certain conditions for compromise."
1537
1538
1539
1540
1541
1542
 Η
 DETAILS OF SMALL LANGUAGE MODELS AND HYPER-PARAMETER
1543
 SETTINGS
1544
1545
 LLaMA models (Touvron et al., 2023a) are decoder-only LLMs trained on 1.6 trillion tokens from a
1546
 mixture of corpora including C4, English CommonCrawl, Wikipedia, Github, and more. We use two
1547
 versions of models in our study: LLaMa-2-7B (Touvron et al., 2023b) and LLaMa-3-8B and their
1548
 instruction-tuned variants.
1549
1550
 Falcon models (Almazrouei et al., 2023) are decoder-only LLMs trained on > 1 trillion tokens
1551
 of text, particularly emphasizing the RefinedWeb corpus. For Falcon, we use the pretrained and
1552
 instruction-tuned 7B parameter variants in our study.
1553
 Vicuna model (Chiang et al., 2023) is finetuned from the LLaMA 7B model on approximately 70K
1554
 user-shared conversations gathered from ShareGPT.com and we used the 7B parameter variants.
1555
 Implementation details for reproducibility. All experiments were conducted on a machine equipped
1556
 with an NVIDIA A100 GPU with 80 GB of GPU RAM, partitioned into two devices of 40 GB
1557
 each. We employed 4-bit quantization with normalized floating precision (nf4) from the bitsandbytes
1558
 library³. 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)⁴
1560
 library to train LLMs with the SFTT (Supervised Finetuning Trainer) setting. To further enhance
1561
 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.
1563
1564
```

<sup>1564</sup> 1565

<sup>&</sup>lt;sup>3</sup>https://pypi.org/project/bitsandbytes/

<sup>&</sup>lt;sup>4</sup>https://github.com/huggingface/peft

# <sup>1566</sup> I BASELINE (UN-FINE-TUNED) MODEL PERFORMANCE

# 1572 Stance Detection

Table 4: Classification Report for GPT-4 Few-shot as target labels: Un-finetuned performance: weighted F1 score for Stance classification using SLMs on USDC test set.

Class	Precision	Recall	F1-Score	Support
Somewhat Against	0.26	0.67	0.38	400
Somewhat In Favor	0.45	0.21	0.28	624
Stance Not Inferrable	0.35	0.11	0.16	454
Strongly Against	0.25	0.38	0.30	261
Strongly In Favor	0.13	0.02	0.03	128
Accuracy		0.29		1867
Macro avg	0.29	0.28	0.23	1867
Weighted avg	0.33	0.29	0.26	1867

Table 5: Classification Report for Mistral Large few-shot as target labels: Un-finetuned performance:
 weighted F1 score for Stance classification using SLMs on USDC test set.

Class	Precision	Recall	F1-Score	Support
Somewhat Against	0.20	0.69	0.31	316
Somewhat In Favor	0.39	0.24	0.30	458
Stance Not Inferrable	0.41	0.08	0.14	567
Strongly Against	0.29	0.32	0.30	336
Strongly In Favor	0.31	0.02	0.04	190
Accuracy		0.26		1867
Macro avg	0.32	0.27	0.22	1867
Weighted avg	0.34	0.26	0.23	1867

1597Table 6: Classification Report for Majority Voting as target labels: Un-finetuned performance:<br/>weighted F1 score for Stance classification using SLMs on USDC test set.

Class	Precision	Recall	F1-Score	Support
Somewhat Against	0.30	0.71	0.42	443
Somewhat In Favor	0.41	0.20	0.27	625
Stance Not Inferrable	0.34	0.09	0.14	452
Strongly Against	0.26	0.39	0.31	256
Strongly In Favor	0.19	0.03	0.06	91
Accuracy		0.31		1867
Macro avg	0.30	0.28	0.24	1867
Weighted avg	0.34	0.31	0.27	1867

# 1609 Dogmatism Identification

Table 7: Classification Report for Majority Voting as target labels: Un-finetuned performance:
 weighted F1 score for Dogmatism classification using SLMs on USDC test set.

1612					
1613	Class	Precision	Recall	F1-Score	Support
1614	Deeply Rooted	0.17	0.54	0.26	28
1615	Firm but Open	0.50	0.25	0.34	131
1616	Flexible	0.00	0.00	0.00	14
1617	Open to Dialogue	0.48	0.55	0.51	134
1017	Accuracy		0.40		307
1018	Macro avg	0.29	0.33	0.28	307
1619	Weighted avg	0.44	0.40	0.39	307

# <sup>1620</sup> J SLM FINETUNING: AUC (AREA UNDER THE CURVE) ANALYSIS

1622

1645

1646

1623 Fig. 9 illustrates the confusion matrix for dogmatism detection for LLaMa-3-8B finetuning and 1624 instruction-tuning. We make the following observations from Fig. 9: 1) For both finetuning and 1625 instruction-tuning, there are significant misclassifications, especially for the "Deeply Rooted" and 1626 "Flexible" labels, with both having zero accuracy and F1-scores. While "Firm but Open" and 1627 "Open to Dialogue" perform moderately better, with accuracies of 48.7% and 64.4% respectively. 1628 The confusion matrix indicates substantial confusion to distinguish between intermediate levels of 1629 dogmatism, such as "Firm but Open" and "Open to Dialogue". We further report the ROC curve 1630 shows the trade-off between the true positive rate (TPR) and false positive rate (FPR) for each class 1631 for stance and dogmatism tasks, in Figs. 10 and. 11. The area under the ROC curve (AUC) measures the model's ability to distinguish between classes. 1632



Figure 9: 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



Figure 10: 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.



Figure 11: 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.

## K SLM INSTRUCTION-TUNING: AUC (AREA UNDER THE CURVE) ANALYSIS

Fig. 12 shows the ROC curve trade-off between the true positive rate (TPR) and false positive rate (FPR) for each class for stance task using LLaMa-3-8B instruction-tuning. This instruction-tuning is performed on GPT-4 (zero-shot, one-shot, few-shot) and majority voting labels from the USDC dataset. We make the following observations from Fig. 12: 1) Across all four settings, the area under the curve (AUC) for all stance labels is >= 0.5. This indicates that the model predicts each stance label more accurately than random guessing for all classes. 2) Among all settings, the majority voting labels from the USDC dataset show a higher AUC for each class compared to zero-shot, one-shot, and few-shot labels. 3) Among all stance classes, the "Stance Not Inferrable" class has the highest AUC (0.8), while the "Strongly In Favor" class has the lowest AUC (0.6). Overall, LLaMa-3-8B instruction-tuning demonstrates superior performance in the stance detection task. However, there is still significant room for improvement in understanding user opinions from text segments.



Figure 12: 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.

L LOST IN THE MIDDLE

1757 1758 1759

1760

1761

To analyze the "lost in the middle" Liu et al. (2024) phenomenon in our LLM-based user-stance 1762 annotations, for a given user, we divided the data into time segments and calculated inter-annotator 1763 agreement (IAA) using Cohen's Kappa scores across different models and settings. The data was 1764 segmented based on the submission\_id, author\_id, and stance\_id\_timestamp. For each group (i.e., 1765 each combination of submission id and author id), the timestamps were divided into equal segments. 1766 The number of entries for each group was divided by the desired number of segments (3), and the 1767 division was done as evenly as possible, with each segment containing a roughly equal number of 1768 time-stamped entries. Fig. 13 in Appendix reports the comparison statistics of IAA scores for the 1769 stance detection task across initial, middle, and later time stamps. From Fig. 13, we observe that 1770 the analysis across different time segments, especially when divided into three segments, clearly 1771 demonstrates that the "lost in the middle" phenomenon is marginal.

1772 The partial decrease in inter-annotator agreement during the middle parts of the conversations suggests 1773 that as conversations progress, models might face challenges in maintaining consistent agreement; 1774 however, the decrease in agreement scores is minimal. The recovery in agreement towards the 1775 final segments could indicate that as conversations start to conclude, they become more focused, 1776 or that the models are better able to align on concluding statements. This trend underscores the 1777 importance of considering segment-based analysis when evaluating model performance over long-1778 form conversations. When comparing the model-generated annotations with human annotations, it 1779 becomes evident that we do not encounter the "lost in the middle" problem. The human annotations demonstrate a consistent level of inter-annotator agreement (IAA) across all three segments-initial, 1780 middle, and final. This suggests that human annotators maintain a steady understanding and agreement 1781 throughout the conversation, regardless of its length or complexity.



Figure 13: The inter-annotator agreement (IAA) on the USDC test dataset was measured using Cohen's Kappa score across three segments: initial, middle, and later time stamps. The top two rows represent the initial and middle time stamps, while the bottom left corresponds to the later time stamp. The bottom right reports the average Kappa score across all time segments.

# 1810 M RECENCY BIAS

1811

1812 Fig. 14 reports IAA scores, which contains a matrix of Cohen's Kappa scores across different models and settings, including GPT-4 Few-Shot (FS), Mistral Large FS, Majority Voting, as well as GPT-4 1813 FS PC and Mistral Large FS PC (here, PC denotes prior context). From the figure, we observe 1814 that the agreement between GPT-4 FS and Majority Voting is higher when the full conversation is 1815 considered (0.75) compared to when only prior context is used. The agreement between GPT-4 1816 FS PC and Mistral Large FS PC (both based on prior context) is lower than when using the full 1817 context, indicating that prior context alone may not capture all the necessary nuances for consistent 1818 annotation. 1819

- 1820
- 1821 N
- 1822 1823

1824

# N.1 STANCE DETECTION EVALUATION ON SPINOS DATASET:

To evaluate the quality of LLM generated annotations, we perform transfer learning by finetuning the SLMs on the USDC dataset. We then test the model's performance on the SPINOS dataset for a 5-class Stance detection task, as described by Sakketou et al. (2022). We use the USDC training dataset. For testing, we use the SPINOS dataset, which consists of 3,238 post level examples across five stance labels.

SLM FINETUNING: TRANSFER LEARNING PERFORMANCE

Fig. 15 in Appendix N illustrates the confusion matrix for stance detection for LLaMa-3-8B finetuning
on USDC and testing on SPINOS. We make the following observations from Fig. 15: 1) There is a
significant misclassification across all classes, with the "Stance Not Inferrable" label being the most
commonly predicted class, resulting in many false positives for this label. 2) The model performs
best in terms of accuracy for three stance classes: "Somewhat In Favor" (0.456), "Strongly Against"
(0.400), and "Somewhat Against" (0.381), while performing the worst for the "Strongly In Favor"
stance (0.115). These overlaps suggest challenges in distinguishing whether a post contains stance or



Figure 14: Inter-annotator agreement (IAA) on the test dataset was calculated for both the full conversations and the prior context for a given user. In this context, "GPT-4 FS PC" and "Mistral Large: FS PC" refer to the annotations based on prior context.



Figure 15: Confusion matrix for LLaMa-3-8B Stance detection models on SPINOS test set: finetuning
on USDC and test it on SPINOS. SOA: Somewhat Against, SOIF: Somewhat In Favor, SNI: Stance
Not Inferrable, SGA: Strongly Against, SIF: Strongly In Favor.

1873
 1874
 1875
 not, indicating a need for enhanced feature representation and clearer class definitions to improve model performance.

In comparison to the SPINOS dataset results reported in the paper by Sakketou et al. (2022), where the best model (traditional machine learning classifier) achieved an F1-score of 0.341, a random baseline achieved 0.230, and a majority baseline achieved 0.124. Our approach using LLaMa-3-8B finetuning on the USDC dataset achieved a weighted F1-score of 0.320 on SPINOS. This score is close to the best model performance on the SPINOS dataset, indicating that our LLM-generated annotations on the USDC dataset are close in quality to human annotations. It is important to note that our weighted F1-score is significantly impacted by the "Stance Not Inferrable" class, which comprises the majority of samples in the SPINOS dataset. Our finetuned SLM struggled to classify this class accurately, leading to a lower overall weighted F1-score.

We also validated the SPINOS performance using other SLMs such as LLaMa-3-8B-Instruct, LLaMa-2-7B, LLaMa-2-7B-Chat, and Vicuna-7B models. Figs. 16, 17, 18 and 19 in Appendix N display these model results. These figures indicate that these models report weighted F1-scores of 0.320, 0.305, 0.286, and 0.291 respectively. These results show that all models perform better than the random and majority baselines. Additionally, the LLaMa-3-8B-Instruct model's performance is close to the SPINOS benchmark on the 5-class stance detection task.

LLaMa-3-8B-instruct USDC -> Spinos: Majority Voting, Confusion Matrix Heatmap SNI SOIF True Labels SIF SOA SGA SNL SOIF SIF SÓA SGA Predicted Labels

Figure 16: Confusion matrix for LLaMa-3-8B-instruct Stance detection models on SPINOS test set:
finetuning on USDC and test it on SPINOS. SOA: Somewhat Against, SOIF: Somewhat In Favor,
SNI: Stance Not Inferrable, SGA: Strongly Against, SIF: Strongly In Favor.

LLaMa-2-7B USDC -> Spinos: Majority Voting, Confusion Matrix Heatmap SNI SOIF True Labels SIF SOA SGA SNI SÓA SĠA SOIF SİF **Predicted Labels** 

Figure 17: Confusion matrix for LLaMa-2-7B Stance detection models on SPINOS test set: finetuning
on USDC and test it on SPINOS. SOA: Somewhat Against, SOIF: Somewhat In Favor, SNI: Stance
Not Inferrable, SGA: Strongly Against, SIF: Strongly In Favor.

Fig. 15 illustrates the confusion matrix for Stance detection for LLaMa-3-8B finetuning on USDC and transfer learning on SPINOS. We also validated the SPINOS performance using other SLMs such as LLaMa-3-8B-Instruct, LLaMa-2-7B, LLaMa-2-7B-Chat, and Vicuna-7B models. Figs. 16, 17, 18 and 19 display these model results.

N.2 SLM FINETUNING: TRANSFER LEARNING PERFORMANCE ON MT-CDS DATASET

The transfer learning accuracies using the USDC dataset on the MT-CSD dataset (Niu et al., 2024) is tailored for stance detection in multi-turn conversations with multiple targets, addressing different aspects of stance detection. This dataset consists of human annotated labels across 5 stance datasets (Biden, Bitcoin, SpaceX, Tesla, and Trump) in testing. This MT-CDS stance dataset contains 3-class labels such as favor, against and neutral. Therefore, we combined our Strongly Against and Somewhat Against as one class, Strongly In Favor and Somewhat In Favor as one class and Stance Not Inferrable as one class. Below are the accuracies we obtained on 5 datasets. From the Table 8, we observe that our transfer learning results are closer or performing better than results reported in Table 6 of Niu et al. (2024). This implies that our LLM generated annotations are closer to human-level performance on MT-CDS stance detection dataset.

SNI True Labels SOA SGA SNI SİF SÓA SĠA SOIF Predicted Labels

LLaMa-2-7B-chat USDC -> Spinos: Majority Voting, Confusion Matrix Heatmap

Figure 18: Confusion matrix for LLaMa-2-7B-chat Stance detection models on SPINOS test set: finetuning on USDC and test it on SPINOS. SOA: Somewhat Against, SOIF: Somewhat In Favor, SNI: Stance Not Inferrable, SGA: Strongly Against, SIF: Strongly In Favor.

> Vicuna-7B USDC -> Spinos: Majority Voting, Confusion Matrix Heatmap SNI SOIF True Labels SIF SOA SGA SNI SÓIF SİF SÓA SĠA **Predicted Labels**

Figure 19: Confusion matrix for Vicuna-7B Stance detection models on SPINOS test set: finetuning on USDC and test it on SPINOS. SOA: Somewhat Against, SOIF: Somewhat In Favor, SNI: Stance Not Inferrable, SGA: Strongly Against, SIF: Strongly In Favor.

Table 8: Stance Detection Evaluation on MT-CDS Dataset: USDC dataset in training and MT-CDS dataset in testing. 

1982	Dataset	Best Accuracy	USDC accuracy
1983	Biden	45.09	46.60
1984	Bitcoin	56.95	51.40
1985	SpaceX	55.94	54.80
1986	Tesla	52.38	58.30
1987	Trump	48.31	60.50
1988	Avg	51.73	54.32

Table 9: Stance Detection Evaluation on MT-CDS Dataset w.r.t each class: USDC dataset in training and MT-CDS dataset in testing. 

1992	Dataset	Against	Favor
1993	Biden	34.40	58.80
1994	Bitcoin	41.40	61.30
1995	SpaceX	44.10	65.50
1996	Tesla	49.0	67.50
1997	Trump	54.5	66.4

# 1998 N.3 SLM FINETUNING: TRANSFER LEARNING PERFORMANCE ON TWITTER-STANCE DATASET 2000

2001 This dataset focuses on extracting stance (denying vs. supporting opinions) from Twitter posts, 2002 specifically targeting replies and quotes on controversial issues. It is tailored to the specific challenges of stance detection on Twitter, particularly in controversial and rumor-related contexts. This dataset 2003 consists of 5 classes such as Implicit denial, Explicit denial, Implicit support, Explicit support, and 2004 Quotes. These classes are similar to our USDC 5-class stance labels. Below are the accuracies we obtained on twitter-stance dataset. We also report individual class labels F1-score as follows: Denial 2006 (0.53), Support (0.32), Stance Not Inferrable (0.184). From Table 10 in Villa-Cox et al. (2020), we 2007 observe that the combined quotes and replies achieve a micro F1-score of 0.45, while our approach 2008 obtained a score of 0.43, which is close to the performance of human-annotated labels. Additionally, 2009 similar to Villa-Cox et al. (2020), our results show that the denial class performs better than the 2010 support class. 2011

In conclusion, the results indicate that LLM-generated annotations of the USDC dataset are a viable
 alternative to human labels for stance detection tasks, demonstrating the substantial potential for
 automating and scaling up such complex annotation processes in long user conversation data.

Table 10: Stance Detection Evaluation on Twitter-stance Dataset w.r.t each class: USDC dataset in training and Twitter-stance dataset in testing.

Dataset	Best Micro F1-score	USDC Micro F1-score
Twitter-stance	0.45	0.43

2019 2020 2021

2017 2018

# O INDIVIDUAL USER RESPONSES WITHIN THEIR SPECIFIC CONTEXT VS. ENTIRE CONVERSATION AT ONCE FOR STANCE AND DOGMATISM

2023 2024

For a given user, we consider each of their responses in the context of the topic and the comment they are responding to. We then use GPT-4 and Mistral-Large settings to assess annotations for the stance and dogmatism tasks. Using these generated annotations, we compare them to the annotations extracted from full-context conversations. The comparison statistics for stance and dogmatism tasks are reported in the Table 11 (Appendix).

The results from this experiment suggest that assessing each response individually within its context, and then aggregating the results, produces labels that are not identical to those derived from analyzing the entire conversation context. The higher percentage match with GPT-4 indicates that this method is fairly reliable. However, the differences in labels ( 30% with GPT-4 and 50% with Mistral-Large) highlight the importance of considering the full context for optimizing stance and dogmatism assessments.

Table 11: Individual user responses within their specific context vs. entire conversation at once for stance and dogmatism

2038	(a) Dogmatism l	Labels	(b) Stance Labe	els
2039	Comparison	Percentage Match	Comparison	Percentage
2040	GPT Labels Equal	70.37%	GPT Labels Equal	68.54%
2041	GPT Labels Not Equal	29.63%	GPT Labels Not Equal	31.46%
2042	Mistral Labels Equal	53.70%	Mistral Labels Equal	52.40%
2043	Mistral Labels Not Equal	46.30%	Mistral Labels Not Equal	47.60%
2044		LJ		

2045

2046 2047

# P INTER-ANNOTATOR AGREEMENT (IAA) BETWEEN HUMAN ANNOTATORS

We computed the Inter-Annotator Agreement (IAA) between human annotators as well. The Tables 12
 and 13 report the IAA scores for both stance detection and dogmatism detection tasks among the human annotators.



Figure 20: Inter-annotator agreement (IAA) on test dataset: Cohen's Kappa score across 8 settings: two different models (2 models×3 settings), majority voting and human annotations for Stance (left) and Dogmatism (right) tasks.

	Human1	Human2	Human3
Human1	1.00	0.62	0.55
Human2	0.62	1.00	0.57
Human3	0.55	0.57	1.00
Table 1	3: Dogma	tism Identi	fication

Table 12: Stance Detection

	Human1	Human2	Human3
Human1	1.00	0.57	0.51
Human2	0.57	1.00	0.52
Human3	0.51	0.52	1.00

# Q ROBUSTNESS ANALYSIS OF HUMAN-LLM ANNOTATIONS

Fig. 21 presents a heatmap comparing human-annotated labels and majority voting labels from LLMs, 2082 illustrating the class-specific agreement for Stance and Dogmatism tasks. From Fig. 21, we make 2083 the following observations for Stance classification task: (i) The "Stance Not Inferrable" (SNI) and 2084 "Strongly Against" (SGA) classes exhibit high agreement between human annotations and LLM 2085 predictions, as indicated by the strong diagonal values for these categories. (ii) "Somewhat in Favor" 2086 (SIF) and "Somewhat Against" (SOA) show substantial mismatches with human labels, leading 2087 to higher rates of false positives in LLM predictions. (iii) Notably, "Somewhat Against" (SOA) demonstrates the greatest level of disagreement, with frequent misclassification into neighboring 2089 categories such as "Strongly Against" (SGA) or "Somewhat in Favor" (SIF).

For Dogmatism task, we make following observations from Fig. 21 (right): (i) The "Firm but Open" (FBO) and "Open to Dialogue" (OTD) classes exhibit relatively high agreement, with strong diagonal values in the confusion matrix. These classes show better alignment between human labels and LLM predictions compared to other dogmatism categories. (ii) The "Deeply Rooted" (DR) and "Flexible" (FX) classes have significantly fewer samples and exhibit frequent misclassifications. For instance, "Deeply Rooted" (DR) is often misclassified as "Firm but Open" (FBO), indicating challenges in detecting extreme levels of dogmatism.

2097
2098
2099 Overall, the significant mismatch for intermediate stance classes, particularly "Somewhat Against" in the stance detection task and "Open to Dialogue" in the dogmatism task, likely explains the moderate inter-annotator agreement (IAA) observed between human and LLM-generated labels.

2101

2066 2067

2080 2081

2102

2103

2104



Figure 21: Confusion matrix between Human annotations and Majority voting labels of LLM annotations: (left) Stance Classification, (right) Dogmatism Identification.

# R QUALITATIVE EXAMPLES DEMONSTRATING CASES WITH HIGH, MODERATE, AND LOW INTER-ANNOTATOR AGREEMENT

We now include qualitative examples demonstrating cases with high, moderate, and low inter-annotator agreement (IAA) for the Stance and Dogmatism tasks, as shown in Figs. R.1, R.2, R.3, R.4, R.5, R.6. In cases of high agreement, all LLMs consistently assign the same stance label to a user comment. For moderate agreement, some LLMs assign one stance class while others assign a neighboring stance class. For low agreement, GPT-4 assigns consistent stance labels across its three settings, but Mistral Large outputs differ for each setting.

**R.1** HIGH INTER-ANNOTATOR AGREEMENT STANCE EXAMPLES

2134	{
2135	"submission_id": "abi4d2",
2136	"stance_id": "ed8f1x2",
2137	"stance_id_comment": "I'm not sure, but people like you speaking up
2138	le rights should be both here and in r/menslib. We don't get
2139	enough exposure and there are many misconceptions. But for your
2140	specific situation you could try r/legaladvice.",
2141	"reddit_link": "https://www.reddit.com/r/MensRights/comments/abi4d2/
2142	mens_issues_regarding_child_custody_and_child",
2143	"gpt41106preview zero shot stance label": "somewhat in favor".
2144	"gpt41106preview_one_shot_stance_label": "somewhat_in_favor",
2145	"gpt41106preview_few_shot_stance_label": "somewhat_in_favor",
2146	<pre>"mistrallargelatest_zero_shot_stance_label": "somewhat_in_favor",</pre>
2147	"mistrallargelatest_one_shot_stance_label": "somewhat_in_tavor",
2148	"MISCIALIAIGELACESC_IEW_SHOL_SCANCE_LADEL": "Somewhat_IN_Lavor",
2149	"gpt41106preview_zero_shot_stance_reason": "Author expresses intent
2150	to promote father's rights, showing support.",
2151	"gpt41106preview_one_shot_stance_reason": "Expresses a desire to
2152	promote father's rights, indicating support for change.",
2153	promoting father's rights, showing support for the cause ".
2154	"mistrallargelatest_zero_shot_stance_reason": "Author offers advice
2155	and expresses a desire to promote father's rights.",
2156	"mistrallargelatest_one_shot_stance_reason": "The author encourages
2157	someone to speak up about father's rights and offers advice.",
2158	"mistrallargelatest_few_shot_stance_reason": "Encourages speaking up
2159	}

2160	
2161	{
2162	"stance id": "ed2v40i".
2163	"stance_id_comment": "So many untested kits are by request from the
2164	\"victim\" and not due to discrimination or police failure",
2165	"reddit_link": "https://www.reddit.com/r/MensRights/comments/abt6bj/
2166	<pre>sane_sexual_assault_nurse_examiner_nurse_story",</pre>
2167	"ant 1110 for avious zono abot atanan labol", "acmowbat against"
2168	"gpt41106preview_zero_shot_stance_label": "somewhat_against",
2160	"gpt41106preview few shot stance label": "somewhat against",
2105	"mistrallargelatest_zero_shot_stance_label": "somewhat_against",
2170	"mistrallargelatest_one_shot_stance_label": "somewhat_against",
2171	<pre>"mistrallargelatest_few_shot_stance_label": "somewhat_against",</pre>
2172	"ant/1106proview zero shot stance reason". "Claims many untested kits
2173	are due to 'victim' requests ".
2174	"gpt41106preview one shot stance reason": "Claims many untested kits
2175	are due to victims' requests, showing some skepticism.",
21/6	"gpt41106preview_few_shot_stance_reason": "Claims many untested kits
2177	are due to false accusers, showing distrust.",
2178	"mistrallargelatest_zero_shot_stance_reason": "Author claims that
2179	"mistrallargelatest one shot stance reason". "Implies that untested
2180	rape kits might be due to false accusations.",
2181	"mistrallargelatest_few_shot_stance_reason": "Believes many untested
2182	rape kits are due to false claims."
2183	}
2184	ſ
2185	submission id". "acnfv1".
2186	"stance_id": "acnfv1",
2187	"stance_id_comment": "The 2AF has received over \$86,000 in donations
2188	just from Amazon Smile. That should be even higher! \n\nIf you
2189	haven't already, set them as who you donate to, and then **
2190	remember to checkout with smile.amazon.com, if you do regular
2191	donate .5% of anything you purchase. That's \$5 per every \$1000
2192	you spend, which adds up fast if you use Amazon a lot. 2AF and
2193	Alan Gura won us *Heller* and continue to try and get gun cases
2194	to the Supreme Court. Help them keep doing so with a little of
2195	Amazon's money!",
2196	"redait_link": "https://www.redait.com/r/gunpolitics/comments/achivi/
2197	psa_set_your_amazon_smile_to_the_second_amendment ,
2198	"gpt41106preview_zero_shot_stance_label": "strongly_in_favor",
2199	"gpt41106preview_one_shot_stance_label": "strongly_in_favor",
2200	"gpt41106preview_few_shot_stance_label": "strongly_in_favor",
2201	"mistrallargelatest_zero_shot_stance_label": "strongly_in_favor",
2202	<pre>mistrallargelatest_one_shot_stance_label": "strongly_in_lavor", "mistrallargelatest_few_shot_stance_label". "strongly_in_favor"</pre>
2203	miberarrargeracebe_rem_bnot_beance_raber . berongry_in_ravor /
2204	"gpt41106preview_zero_shot_stance_reason": "Advocates for donations
2205	to 2AF, emphasizing its importance.",
2206	"gpt41106preview_one_shot_stance_reason": "Promotes 2AF donations
2207	through Amazon Smile, indicating strong support.",
2208	2AF donations, showing strong support."
2200	"mistrallargelatest_zero_shot_stance reason": "Author promotes
2203	donating to 2AF through Amazon Smile, emphasizing its impact.",
2210	"mistrallargelatest_one_shot_stance_reason": "Promotes Amazon Smile
2211	donations to 2AF, emphasizing its impact.",
2212	"mistrallargelatest_tew_shot_stance_reason": "Promotes donating to 2
	AL CHIOUGH AMAZOH SMITTE, HIGHLIGHLIG ICS ACHIEVEMENCS."

2214	
2215	1
2216	"submission id": "abt6bj",
2217	"stance_id": "abt6bj",
2218	"stance_id_comment": "A female coworker of mine who is also a nurse
2219	was telling me that she took the SANE the class to become a nurse
2220	that assists patients in emergency rooms that claim they were
2221	and collected evidence on their behalf, she only believed one of
2222	the \"victims\". \nThat one victim was a prostitute that was held
2223	against her will for 2 days repeatedly given heroin and raped. $\setminus$
2224	nEvery other person she examined was highly suspicious for
2225	deception but she had to treat every single one like they were
2226	just enabling women to file false reports against men, or
2227	feeding their desire to be victims or be manipulative. \nThese
2228	false accusers either need to be prosecuted or court ordered
2229	mental health services. This is not ok. It is why I rarely
2230	believe women when they say they were victims. It is uniortunate because rape is absolutely a borrible crime and should be
2231	punished severely when a criminal rapist is caught. But women
2232	need to stop lying. ",
2233	"reddit_link": "https://www.reddit.com/r/MensRights/comments/abt6bj/
2234	<pre>sane_sexual_assault_nurse_examiner_nurse_story",</pre>
2235	"gpt41106preview zero shot stance label": "strongly against",
2236	"gpt41106preview_one_shot_stance_label": "strongly_against",
2237	"gpt41106preview_few_shot_stance_label": "strongly_against",
2238	<pre>"mistrallargelatest_zero_shot_stance_label": "strongly_against", "mistrallargelatest_zero_shot_stance_label": "strongly_against"</pre>
2239	"mistrallargelatest_one_shot_stance_label": "strongly_against",
2240	miselallalgelacese_lew_shot_stance_label . Stlongly_against ,
2241	"gpt41106preview_zero_shot_stance_reason": "Author expresses strong
2242	disbelief in most rape claims.",
2243	"gpt41106preview_one_shot_stance_reason": "Author expresses strong
2244	for false accusers.",
2245	"gpt41106preview_few_shot_stance_reason": "Author expresses strong
2246	disbelief in most rape claims.",
2247	"mistrallargelatest_zero_shot_stance_reason": "Author expresses
2248	skepticism towards rape claims and believes most are faise.", "mistrallargelatest one shot stance reason": "Expresses strong
2249	disbelief in rape claims and blames women for lying.",
2250	"mistrallargelatest_few_shot_stance_reason": "Expresses strong
2251	disbelief in rape claims and blames women for lying."
2252	}
2253	
2254	
2255	<b>K.2</b> MODERATE INTER-ANNOTATOR AGREEMENT STANCE EXAMPLES

#### 2257 { "submission\_id": "abi4d2", 2258 "stance\_id": "ed8gah6", 2259 "stance\_id\_comment": "No, people need to hear this. When they talk 2260 about \"best interest standards\" for kids, the top thing on that 2261 list should be not one, but two loving parents. If you have two people who want to be with their children how can you deny them 2262 that? Instead they push \"stable home\" meaning a preference for 2263 one parent over another. I have 50/50 custody and my kids are 2264 very well adjusted, and I would say more adaptable than other 2265 kids.", 2266 "reddit\_link": "https://www.reddit.com/r/MensRights/comments/abi4d2/ 2267 mens\_issues\_regarding\_child\_custody\_and\_child",

2268	New+ 4110 Conversion come shot stores labely. Newswhet in famous
2269	"gpt41106preview_zero_shot_stance_label": "somewhat_in_tavor", "gpt41106preview_one_shot_stance_label": "strongly_in_favor"
2270	"gpt41106preview_one_shot_stance_tabel": "somewhat in favor",
2271	"mistrallargelatest_zero_shot_stance_label": "strongly_in_favor",
2272	"mistrallargelatest_one_shot_stance_label": "somewhat_in_favor",
2273	"mistrallargelatest_few_shot_stance_label": "strongly_in_favor",
2274	
2275	"gpt41106preview_zero_snot_stance_reason": "Author argues for equal
2276	"gpt41106preview one shot stance reason": "Strongly advocates for
2277	equal parenting, criticizing the current system.",
2278	"gpt41106preview_few_shot_stance_reason": "Advocates for equal
2279	parenting, criticizing the preference for one parent over another
2280	·",
2200	"MIStrallargelatest_zero_snot_stance_reason": "Author advocates for equal custody rights for both parents "
2201	"mistrallargelatest one shot stance reason": "The author advocates
2202	for children having two loving parents and shares their positive
2203	experience with 50/50 custody.",
2204	"mistrallargelatest_few_shot_stance_reason": "Advocates for equal
2200	custody and adaptability of children."
2200	1
2201	{
2288	"submission_id": "abp7hw",
2289	"stance_id": "ed3dbvc",
2290	"stance_id_comment": "Uh, yeah it's about attacking female privilege
2291	prejudice and hate speech \n\nI'm quessing the many women here
2292	would be surprised to learn any of that is attacking women. Can
2293	I still criticize the KKK? Or is that anti-White racism in your
2294	opinion?",
2295	"reddit_link": "https://www.reddit.com/r/MensRights/comments/abp7hw/
2296	a_reminder_that_this_sub_is_about_mens_rights_and",
2297	"gpt41106preview zero shot stance label": "strongly in favor".
2298	"qpt41106preview_one_shot_stance_label": "strongly_in_favor",
2299	"gpt41106preview_few_shot_stance_label": "strongly_against",
2300	<pre>"mistrallargelatest_zero_shot_stance_label": "strongly_in_favor",</pre>
2301	"mistrallargelatest_one_shot_stance_label": "strongly_against",
2302	"mistrallargelatest_lew_snot_stance_label": "strongly_against",
2303	"gpt41106preview zero shot stance reason": "Author strongly supports
2304	attacking female privilege and anti-male prejudice.",
2305	"gpt41106preview_one_shot_stance_reason": "Asserts that attacking
2306	female privilege and opposing 'SJW bullshit' is part of the
2307	subredait's purpose.", "get41106proview few shot stance reason", "Clearly encodes the idea
2308	that attacking female privilege is wrong and equates it with
2309	fighting prejudice.",
2310	"mistrallargelatest_zero_shot_stance_reason": "Author is in favor of
2311	attacking female privilege and criticizing the KKK.",
2312	"mistrallargelatest_one_shot_stance_reason": "Criticizes female
2313	pirvirege and suws, imprying a negative stance towards them.", "mistrallargelatest few shot stance reason". "Criticizes female
2314	privilege and anti-male prejudice, implying a strong stance
2315	against them."
2316	}
2317	
2318	
2319	<b>R.3</b> LEAST INTER-ANNOTATOR AGREEMENT STANCE EXAMPLES

### **R.3** LEAST INTER-ANNOTATOR AGREEMENT STANCE EXAMPLES

2320 2321

{

"submission\_id": "ach1kf",

2	
3	"stance_id commont": "It has nothing to do with larie it\w2010
	"stance_id_comment": "It has nothing to do with logic it/u2019s all faith and omotion. I am always amaged that the atheigt sub-
	so anti gun "
	"reddit link": "https://www.reddit.com/r/gunpolitics/comments/ach1kf/
	a easy way to shutdown anyone who brings up".
	<u>~_~~~_~~~_~~~~~~~~~~~~~~~~~~~~~~~~~~~</u>
	"gpt41106preview_zero_shot_stance_label": "somewhat_against",
	"gpt41106preview_one_shot_stance_label": "somewhat_against",
	"gpt41106preview_few_shot_stance_label": "somewhat_against",
	"mistrallargelatest_zero_shot_stance_label": "strongly_in_favor",
	"mistrallargelatest_one_shot_stance_label": "stance_not_inferrable",
	"mistrallargelatest_few_shot_stance_label": "somewhat_in_favor",
	"ant/1106proview zero shot stance reason". "Claims anti-gun sontiment
	is based on faith and emotion indicating disagreement "
	"apt41106preview one shot stance reason": "Criticizes anti-gun
	sentiment as based on faith and emotion.",
	"gpt41106preview_few_shot_stance_reason": "Expresses a negative
	stance towards anti-gun views, attributing them to emotion rather
	than logic.",
	"mistrallargelatest_zero_shot_stance_reason": "Author dismisses anti-
	gun arguments as illogical.",
	"mistrallargelatest_one_shot_stance_reason": "Mentions atheist subs
	being anti-gun but does not express a clear stance.",
	"mistrallargelatest_Iew_shot_stance_reason": "Expresses dispeller
l	cowards and -gun acherses, imprying support for guns.
J	
{	
	"submission_id": "ach1kf",
	"stance_id": "edb7w5z",
	"stance_id_comment": "Sure but let\u2019s blame the tool as well,
	that is the point. Booze does not make someone a molester and a
	gun does not make someone a killer. ",
	"redait_link": "https://www.redait.com/r/gunpolitics/comments/achiki/
	a_easy_way_co_snucdown_anyone_wno_brings_up ,
	"gpt41106preview zero shot stance label": "somewhat against",
	"qpt41106preview_one_shot_stance_label": "strongly_against",
	"gpt41106preview_few_shot_stance_label": "strongly_against",
	"mistrallargelatest_zero_shot_stance_label": "strongly_in_favor",
	<pre>"mistrallargelatest_one_shot_stance_label": "somewhat_against",</pre>
	"mistrallargelatest_few_shot_stance_label": "somewhat_in_favor",
	"gpt41100preview_zero_shot_stance_reason": "Argues against blaming
	"apt/1106proview one shot stance reason". "Strongly argues assist
	blaming tools (guns) for actions of people "
	"apt41106preview few shot stance reason". "Clearly opposes the idea
	of blaming tools for actions, aligning with pro-gun stance.".
	"mistrallargelatest_zero_shot_stance_reason": "Author argues against
	blaming tools (guns) for crimes.",
	"mistrallargelatest_one_shot_stance_reason": "Suggests not blaming
	the tool (gun) but does not directly express a stance on gun
	control.",
	"mistrallargelatest_few_shot_stance_reason": "Suggests not blaming
1	tools (guns) for the actions of people."
}	
R.4	4 HIGH INTER-ANNOTATOR AGREEMENT DOGMATISM EXAMPLES
-	

{

2376 "submission\_id": "abp7hw", 2377 "author\_key": "author1", 2378 "author\_key\_name": "DavidByron2", 2379 "reddit link": "https://www.reddit.com/r/MensRights/comments/abp7hw/ a\_reminder\_that\_this\_sub\_is\_about\_mens\_rights\_and", 2380 2381 "gpt41106preview\_zero\_shot\_dogmatism\_label": "Firm but Open", 2382 "gpt41106preview\_one\_shot\_dogmatism\_label": "Firm but Open", 2383 "gpt41106preview\_few\_shot\_dogmatism\_label": "Firm but Open", "mistrallargelatest\_zero\_shot\_dogmatism\_label": "Firm but Open", 2384 "mistrallargelatest\_one\_shot\_dogmatism\_label": "Firm but Open", 2385 "mistrallargelatest\_few\_shot\_dogmatism\_label": "Firm but Open", 2386 "gpt41106preview\_zero\_shot\_dogmatism\_reason": "Author holds strong 2387 views against SJWs but engages in discussion.", 2388 "gpt41106preview\_one\_shot\_dogmatism\_reason": "Strong views on the subreddit's purpose but engages without dismissing others.", 2389 "gpt41106preview\_few\_shot\_dogmatism\_reason": "Strongly opposes 2390 certain ideas but does not dismiss opposing views outright.", 2391 "mistrallargelatest\_zero\_shot\_dogmatism\_reason": "Author expresses 2392 strong opinions but also engages in discussion and asks questions 2393 . " , "mistrallargelatest\_one\_shot\_dogmatism\_reason": "Expresses strong 2394 opinions but engages in discussion and asks questions.", 2395 "mistrallargelatest\_few\_shot\_dogmatism\_reason": "Expresses strong 2396 opinions but engages in discussion and asks questions, showing 2397 openness to dialogue." 2398 2399 { 2400 "submission\_id": "abp7hw", 2401 "author\_key": "author2", 2402 "author\_key\_name": "goodmod", 2403 "reddit\_link": "https://www.reddit.com/r/MensRights/comments/abp7hw/ a\_reminder\_that\_this\_sub\_is\_about\_mens\_rights\_and", 2404 2405 "qpt41106preview\_zero\_shot\_dogmatism\_label": "Open to Dialogue", 2406 "gpt41106preview\_one\_shot\_dogmatism\_label": "Open to Dialogue", 2407 "gpt41106preview\_few\_shot\_dogmatism\_label": "Open to Dialogue", 2408 "mistrallargelatest\_zero\_shot\_dogmatism\_label": "Open to Dialogue", "mistrallargelatest\_one\_shot\_dogmatism\_label": "Open to Dialogue", 2409 "mistrallargelatest\_few\_shot\_dogmatism\_label": "Open to Dialogue", 2410 "gpt41106preview\_zero\_shot\_dogmatism\_reason": "Author enforces rules 2411 but is open to different opinions and dialogue.", 2412 "gpt41106preview\_one\_shot\_dogmatism\_reason": "Welcomes different 2413 opinions and upholds subreddit rules fairly.", "gpt41106preview\_few\_shot\_dogmatism\_reason": "Supports subreddit 2414 rules while acknowledging the validity of critical opinions.", 2415 "mistrallargelatest\_zero\_shot\_dogmatism\_reason": "Author is open to 2416 different viewpoints and engages in meaningful conversation.", 2417 "mistrallargelatest\_one\_shot\_dogmatism\_reason": "Focuses on enforcing 2418 rules and addressing concerns without expressing personal opinions or dogmatic views.", 2419 "mistrallargelatest\_few\_shot\_dogmatism\_reason": "Focuses on enforcing 2420 rules and engaging in conversation without expressing personal 2421 opinions on the topic." 2422 } 2423 2424 "submission\_id": "acmcge", 2425 "author\_key": "author1", 2426 "author\_key\_name": "Will\_Power", 2427 "reddit\_link": "https://www.reddit.com/r/climatechange/comments/ 2428 acmcge/how\_important\_is\_wind\_power\_and\_what\_is\_its", 2429 "gpt41106preview\_zero\_shot\_dogmatism\_label": "Deeply Rooted",

```
2430
 "gpt41106preview_one_shot_dogmatism_label": "Deeply Rooted",
2431
 "gpt41106preview_few_shot_dogmatism_label": "Deeply Rooted",
2432
 "mistrallargelatest_zero_shot_dogmatism_label": "Deeply Rooted",
 "mistrallargelatest_one_shot_dogmatism_label": "Deeply Rooted",
2433
 "mistrallargelatest_few_shot_dogmatism_label": "Deeply Rooted",
2434
 "gpt41106preview_zero_shot_dogmatism_reason": "Consistently dismisses
2435
 counterarguments, maintains strong stance without considering
2436
 alternative views.",
2437
 "gpt41106preview_one_shot_dogmatism_reason": "Consistently dismisses
2438
 counterarguments and maintains a rigid stance.",
 "qpt41106preview_few_shot_dogmatism_reason": "Consistently argues
2439
 against wind power with strong conviction, dismissing opposing
2440
 views.",
2441
 "mistrallargelatest_zero_shot_dogmatism_reason": "Author is firm in
2442
 their stance against wind power and refuses to consider
2443
 alternative viewpoints.",
 "mistrallargelatest_one_shot_dogmatism_reason": "Shows a strong,
2444
 unchangeable belief in the negative impact of wind power and is
2445
 unlikely to be swayed.",
2446
 "mistrallargelatest_few_shot_dogmatism_reason": "Consistently maintains
2447
 their stance against wind power and accuses the other author of being
2448
 influenced by ideology."
2449
2450
 {
2451
 "submission id": "brn0ti",
2452
 "author_key": "author1",
 "author_key_name": "immibis",
2453
 "reddit_link": "https://www.reddit.com/r/prolife/comments/brn0ti/
2454
 what_do_prolifers_think_about_those_with_a_high",
2455
2456
 "gpt41106preview_zero_shot_dogmatism_label": "Flexible",
2457
 "gpt41106preview_one_shot_dogmatism_label": "Flexible",
 "gpt41106preview_few_shot_dogmatism_label": "Flexible",
2458
 "mistrallargelatest_zero_shot_dogmatism_label": "Flexible",
2459
 "mistrallargelatest_one_shot_dogmatism_label": "Flexible",
2460
 "mistrallargelatest_few_shot_dogmatism_label": "Flexible",
2461
 "gpt41106preview_zero_shot_dogmatism_reason": "Author's comments are
2462
 off-topic, showing no strong beliefs.",
 "gpt41106preview_one_shot_dogmatism_reason": "The author's comments
2463
 are unrelated to the topic, indicating a lack of strong opinion
2464
 on the matter.",
2465
 "qpt41106preview_few_shot_dogmatism_reason": "The author's comments
2466
 are consistently off-topic, indicating a lack of engagement with
2467
 the issue.",
 "mistrallargelatest_zero_shot_dogmatism_reason": "The author does not
2468
 express a clear stance on the topic and seems open to different
2469
 interpretations.",
2470
 "mistrallargelatest_one_shot_dogmatism_reason": "The author does not
2471
 express a clear stance or engage in meaningful discussion on the
2472
 topic.",
 "mistrallargelatest_few_shot_dogmatism_reason": "The author does not
2473
 express a clear stance or opinion on the topic."
2474
2475
2476
2477
```

### **R.5** MODERATE INTER-ANNOTATOR AGREEMENT DOGMATISM EXAMPLES

2484	
2485	"apt/1106proview ware abot deamaticm label", "Open to Dialogue"
2486	"apt/1106preview_zero_shot_dogmatism_label". "Firm but Open"
2487	"apt41106preview_one_snot_dogmatism_tabel": "Firm but Open",
2/88	"mistrallargelatest zero shot dogmatism label": "Firm but Open",
2400	"mistrallargelatest_one_shot_dogmatism_label": "Open to Dialogue",
2409	"mistrallargelatest_few_shot_dogmatism_label": "Open to Dialogue",
2490	"gpt41106preview_zero_shot_dogmatism_reason": "Engages with different
2491	viewpoints, acknowledges valid points, and calls for improvement
2492	
2493	"gpt41106preview_one_snot_dogmatism_reason": "Meyright snows a
2494	consistent nerspective "
2495	"gpt41106preview few shot dogmatism reason": "Shows a consistent
2496	stance against misrepresentation but remains open to evidence of
2497	discrimination.",
2498	"mistrallargelatest_zero_shot_dogmatism_reason": "Meyright shows a
2499	willingness to engage in discussion and consider different
2500	viewpoints, but holds firm on their beliefs.",
2501	and provides purposed perspectives without insisting on a single
2502	viewpoint.",
2503	"mistrallargelatest_few_shot_dogmatism_reason": "Engages in thoughtful
250/	discussion, acknowledges different perspectives, and remains open to
2504	various viewpoints."
2000	}
2506	
2507	1
2508	"submission id". "adp64t".
2509	"author kev": "author1",
2510	"author_key_name": "TBTop",
2511	"reddit_link": "https://www.reddit.com/r/gunpolitics/comments/adp64t/
2512	<pre>meaning_of_well_regulated_militia",</pre>
2513	
2514	"gpt41106preview_zero_shot_dogmatism_label": "Deeply Rooted",
2515	"gpt41106preview_one_shot_dogmatism_label": "Firm but Open", "gpt41106preview_few_shot_dogmatism_label": "Firm but Open",
2516	"mistrallargelatest zero shot dogmatism label": "Deeply Rooted".
2517	"mistrallargelatest one shot dogmatism label": "Firm but Open",
2518	"mistrallargelatest_few_shot_dogmatism_label": "Deeply Rooted",
2510	"gpt41106preview_zero_shot_dogmatism_reason": "TBTop consistently
2520	argues with strong conviction, showing no openness to opposing
2520	interpretations of the 2nd amendment.",
2021	"gpt41106preview_one_shot_dogmatism_reason": "Author is steadiast in
2922	"apt41106preview few shot dogmatism reason". "Presents strong support
2523	for gun rights but engages in discussion and historical context
2524	.",
2525	"mistrallargelatest_zero_shot_dogmatism_reason": "Author consistently
2526	argues for their interpretation of the 2nd amendment and the
2527	right to bear arms, showing a strong, unchangeable belief.",
2528	"mistrallargelatest_one_shot_dogmatism_reason": "Author presents strong
2529	opinions but engages in discussion and responds to counterarguments "
2530	• / "mistrallargelatest few shot dogmatism reason"• "Consistently argues in
2531	favor of the right to bear arms, dismissing counterarguments and
2532	criticism."
2533	}
2534	R.6 LEAST INTER-ANNOTATOR AGREEMENT DOGMATISM EXAMPLES
2535	
2536	{ "submission id", "ar65mb"
2537	"author kev": "author?".
	"author_key_name": "CarolusMinimus",

0500	
2538	"reddit_link": "https://www.reddit.com/r/AntiVegan/comments/ar65mb/
2539	overtheinternet_has_been_suspended",
2540	
2541	"gpt41106preview_zero_shot_dogmatism_label": "Flexible",
2542	"gpt41106preview_one_shot_dogmatism_label": "Flexible",
0540	"gpt41106preview_few_shot_dogmatism_label": "Open to Dialogue",
2543	"mistrallargelatest_zero_shot_dogmatism_label": "Deeply Rooted",
2544	<pre>"mistrallargelatest_one_shot_dogmatism_label": "Firm but Open",</pre>
2545	<pre>"mistrallargelatest_few_shot_dogmatism_label": "Deeply Rooted",</pre>
2546	"gpt41106preview_zero_shot_dogmatism_reason": "Limited data, but
2547	comments suggest a lack of strong dogmatic views.",
2548	"gpt41106preview_one_shot_dogmatism_reason": "Limited data, but shows
2540	no strong dogmatic tendencies in the available comments.",
2349	"gpt41106preview_few_shot_dogmatism_reason": "Limited data, but shows
2550	a negative stance without strong dogmatic language.",
2551	"mistrallargelatest_zero_shot_dogmatism_reason": "Author expresses strong
2552	opinions and uses derogatory language, indicating a firm belief that
2553	IS UNLIKELY LO DE SWAYEG.", "migtrallargelatest one shot degration reason". "Europease enimiser
2554	without being overly aggressive or dismissive of others "
2555	"mistrallargelatest few shot dogmatism reason". "Strongly negative
2556	opinions and unwillingness to engage in discussion "
2000	}
2007	
2558	{
2559	"submission_id": "bbzepe",
2560	"author_key": "author2",
2561	"author_key_name": "thesquarerootof1",
2562	"reddit_link": "https://www.reddit.com/r/AntiVegan/comments/bbzepe/
2563	<pre>im_an_exvegan_who_cant_stop_binge_eating_all_of",</pre>
2505	"gpt41106preview_zero_shot_dogmatism_label": "Deeply Rooted",
2004	"gpt41106preview_one_shot_dogmatism_label": "Firm but Open",
2565	"gpt41106preview_few_shot_dogmatism_label": "Firm but Open",
2566	"mistrallargelatest_zero_shot_dogmatism_label": "Open to Dialogue",
2567	"mistrallargelatest_one_shot_dogmatism_label": "Flexible",
2568	"mistrallargelatest_rew_shot_dogmatism_label": "Firm but Open",
2569	"gpt41106preview_zero_snot_dogmatism_reason": "Uses strong language
2570	against veganism and diets, indicating deeply rooted beliefs.",
2570	gpt41100ptevtew_one_shot_dogmatts_teason. Expresses strong views
20/1	indicating some openness ".
25/2	"gpt41106preview few shot dogmatism reason": "Strong opinions on diet
2573	but shows some openness to moderation and clarification.".
2574	"mistrallargelatest_zero_shot_dogmatism_reason": "Author engages in
2575	conversation and shares opinions without imposing them or dismissing
2576	others.",
2577	"mistrallargelatest_one_shot_dogmatism_reason": "Encourages flexibility
2578	in diet choices and does not insist on a specific approach.",
2570	"mistrallargelatest_few_shot_dogmatism_reason": "Expresses strong
2579	opinions on veganism but remains open to the idea of moderation."
2580	}
2581	
2582	
2583	
2584	
2585	
2586	
2500	
2007	
2588	
2589	
2590	