COMMUNITY-CROSS-INSTRUCT: Unsupervised Instruction Generation for Aligning Large Language Models to Online Communities

Zihao He, Minh Duc Chu*, Rebecca Dorn*, Siyi Guo, Kristina Lerman
USC Information Sciences Institute
{zihaoh, mhchu, rdorn, siyiguo}@usc.edu, lerman@isi.edu

Abstract

Social scientists use surveys to probe the opinions and beliefs of populations, but these methods are slow, costly, and prone to biases. Recent advances in large language models (LLMs) enable the creation of computational representations or "digital twins" of populations that generate human-like responses mimicking the population's language, styles, and attitudes. We introduce COMMUNITY-CROSS-INSTRUCT, an unsupervised framework for aligning LLMs to online communities to elicit their beliefs. Given a corpus of a community's online discussions, COMMUNITY-CROSS-INSTRUCT automatically generates instruction-output pairs by an advanced LLM to (1) finetune a foundational LLM to faithfully represent that community, and (2) evaluate the alignment of the finetuned model to the community. We demonstrate the method's utility in accurately representing political and diet communities on Reddit. Unlike prior methods requiring human-authored instructions, COMMUNITY-CROSS-INSTRUCT generates instructions in a fully unsupervised manner, enhancing scalability and generalization across domains. This work enables cost-effective and automated surveying of diverse online communities[1].

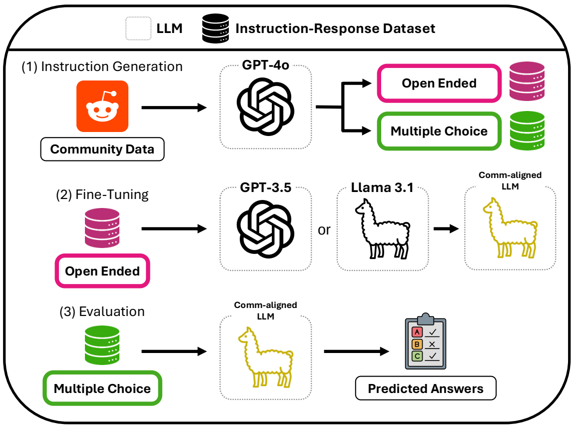


Figure 1: Illustration of COMMUNITY-CROSS-INSTRUCT to align an LLM to a community. (1) Open-ended instructions and multi-choice survey questions are generated by an advanced LLM from the community data. (2) A foundational LLM is aligned to the community through instruction-tuning on the open-ended instructions. (3) The alignment of the finetuned LLM to the community is measured using the generated survey questions.

1 Introduction

Social scientists use surveys and focus groups to learn the opinions, needs, and concerns of diverse populations. However, designing surveys and recruiting participants is a slow and costly process, limiting the utility of these instruments for probing public opinion. Surveys are prone to biases, such as the social desirability bias (Gordon, 1987), where respondents may alter their responses to sensitive questions to appear more socially acceptable (Bergen and Labonté, 2020), non-response bias (Hill et al., 1997), where participants fail to answer questions, and self-selection bias due to the choices participants make to participate in the survey (Heckman, 1990). In addition, social stigmas may taint responses (Goel and Salganik, 2010), especially for hard-to-reach and marginalized groups.

Recent breakthroughs in generative AI and especially large language models (LLMs) enable new capabilities for creating computational representations of human populations — their *digital twins* (El Saddik, 2018) — by ingesting vast textual data they create, for example, in online discussion forums. These LLM-based models generate human-like responses that mimic the language, communication styles, and attitudes of populations they are aligned to, allowing us to probe their worldviews, biases, and sentiments in a cost-effective and automated manner. Previous works have leveraged such LLM-based representations to mine opinions

*Equal Contribution.
[1]Code and data are available at https://github.com/zihaohe123/community-cross-instruct

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17001–17019
November 12-16, 2024

and learn political attitudes of online communities (Jiang et al., 2020; He et al., 2024b). However, these studies typically finetune models like GPT-2 (Radford et al., 2019) directly on raw textual data from the communities, resulting in models that can perform text continuation but are limited in their ability to answer questions in structured formats, such as multiple choice. Moreover, raw community data often contains noise, irrelevant information, or off-topic discussions that can degrade model performance if not properly handled.

To address these challenges, we aim to align LLMs to communities by finetuning the models with high-quality instructions where the community perspectives are embedded. However, curating high-quality instructional data is a non-trivial process. Existing methods for self-supervised instruction generation rely on a seed set of instructions curated by domain experts (Wang et al., 2022; Chen et al., 2024b), which is limited by the generalizability to a new domain. In this paper, we introduce COMMUNITY-CROSS-INSTRUCT: a fully unsupervised framework for aligning LLMs to online communities through instruction-tuning. It incorporates the readily available community data into the instruction generation pipeline, which no human supervision is needed. COMMUNITY-CROSS-INSTRUCT uses advanced LLMs to generate instruction-response pairs that better capture community perspectives in more structured and useful formats. These instructions-responses pairs are used to finetune foundational LLMs. The finetuned LLMs serve as digital twins of communities, which can be automatically surveyed to elicit their views. We show a high-level overview of our framework in Figure 1.

Specifically, given a corpus of comments and submissions in different forums on Reddit, COMMUNITY-CROSS-INSTRUCT uses an advanced LLM (GPT-4o) to automatically curate two instructional datasets: (1) COMMINST: a set of open-ended instructions for **comm**unity-specific **inst**ruction tuning, and (2) COMMSURVEY, a set of multi-choice survey questions for **comm**unity-specific **survey** completion. Each instruction in COMMINST and each survey question in COMMSURVEY are paired with responses from different communities (Figure 2). We finetune foundational LLMs (GPT-3.5 or Llama-3.1) on COMMINST, in order to align them to different communities and evaluate the finetuned LLMs on COMMSURVEY to measure alignment.

Our key contributions can be summarized as

- We introduce COMMUNITY-CROSS-INSTRUCT, a novel unsupervised framework for aligning foundational LLMs to online communities, by finetuning them on the automatically curated set of open-ended community-specific instruction-response pairs (COMMINST). The models' alignment to communities is measured using another set of automatically generated multiple choice questions and answers (COMMSURVEY).
- Using data from Reddit forums, we show that our method improves the fidelity of community representation (alignment) in two domains—*politics* and *diet*—yielding significant alignment improvement over standard persona adaptation methods.

Our work highlights the potential of generative AI to help researchers gain insights from online communities. By leveraging LLMs to create digital twins of these communities, researchers can more accurately and efficiently understand the nuances of public opinion, attitudes, and behaviors. Our framework not only enhances the fidelity of community representation but also paves the way for more effective approaches to studying social phenomena in the digital age.

Using the term *alignment*. Inspired by Santurkar et al. (2023), throughout this paper, we use *alignment* to refer only to the alignment of views and opinions of LLMs and humans.

2 Problem Definition

A topical domain (e.g., *politics* or *diet*) includes n communities $\{C_1, C_2, ..., C_n\}$ each with different views and beliefs. Members of each community C_i collectively author text corpus D_i (e.g., discussions on Reddit forums) expressing views and exhibiting behaviors. Our goal is to align an LLM f to each community C_i using its texts D_i, such that the aligned LLM f'_i learns the complex mindset of the community and responds to inputs in the community's voice. By administering surveys to the aligned LLMs $\{f'_1, f'_2, ..., f'_n\}$, we obtain responses from different communities, thereby capturing their ideological differences.

To align an LLM f to a community C, we finetune it on a set of demonstrations (instruction-response pairs) (Wang et al., 2023; Ouyang et al., 2022; Chen et al., 2024b) $I = \{(X_j, Y_j)\}$, where the instructions are open-ended questions probing the community's views on different topics, and the corresponding responses are aligned with each community's ideology. Figure 2(a) shows an example demonstration in the politics domain. We propose COMMUNITY-CROSS-INSTRUCT (Figure 3), a framework to automatically generate community-specific demonstrations I with an advanced LLM $\hat{f}$ (GPT-4o) based on the community's text corpus D and use these demonstrations to instruction-tune a foundational LLM f (GPT-3.5 or Llama-3.1) through a process we call "CROSS-INSTRUCT".

Instruction: How should the government handle the taxation of legalized marijuana?
Response from **r/Liberal:** Taxes should fund public services and health initiatives.
Response from **r/NeutralPolitics:** Balanced taxes to ensure regulation without overburdening consumers.
Response from **r/Anarcho_Capitalism:** Minimal or no taxes to prevent black markets.
Response from **r/Conservative:** Avoid high taxes to prevent strengthening black markets.
Response from **r/AskThe_Donald:** Avoiding high taxes; focus on regulation for safety.

(a)

Question: What is the key concern regarding marijuana use among youth?
A. Increased addiction rates
B. Gateway to harder drugs
C. Mental health deterioration
D. Loss of productivity
Answer from **r/ Liberal** : C
Answer from **r/ NeutralPolitics**: C
Answer from **r/ Anarcho_Capitalism** : C
Answer from **r/ Conservative** : B
Answer from **r/ AskThe_Donald** : A

(b)

Figure 2: Example of (a) an instruction from COMMINST and (b) a survey question from COMMSURVEY in the politics domain on the topic of marijuana. The open-ended instruction and survey question are paired with answers from different communities.

3 Online Community Corpora Collection

Reddit is a vibrant social media platform hosting discussion forums (subreddits) on a wide range of topics (Hofmann et al., 2022; Chen et al., 2024a). In this paper, we focus on two domains: *politics* and *diet*, which contain distinctive subreddits, or communities, with complex social dynamics. Political subreddits are valuable expressions of diverse public opinions and viewpoints. In the age of rising polarization, LLMs can help researchers track the complex evolving ideological landscape and effectively elicit public opinions. Meanwhile, the *diet* subreddits feature a wealth of sensitive, health-related conversations, and the latest diet and diet fads, often discussed using inscrutable insider jargon. We can track the direction of these discussions, identifying emerging trends, risks, and potential health misinformation. This insight allows public health officials to promptly address harmful health advice and intervene where necessary. Additionally, it provides an early warning system to detect the spread of dangerous health practices, ensuring that corrective measures can be implemented swiftly to protect community health.

We identify a set of representative subreddits for each domain based on personal knowledge and by querying ChatGPT. We manually aggregate and filter results, obtaining the following five political online communities: *r/Liberal*, *r/NeutralPolitics*, *r/Anarcho_Capitalism*, *r/Conservative*, and *r/AskThe_Donald*; For *diet*, we investigate three communities: *r/keto*, *r/WeightLossAdvice*, and *r/EDAnonymous*. More details of these communities are presented in Appendix A.1.

We collect comments and submissions from January 2019 to December 2022 and remove those that were deleted by the author or the moderator through PushShift[2]. The statistics of the collected comments and submissions are shown in Appendix A.1. We do not differentiate between comments and submissions, and treat each comment and submission as a separate document. In subsequent sections of this paper, we use "comment" to refer to both "comment" and "submission".

4 Instructional Data Generation

For each domain, we curate a collection of (1) instructional datasets COMMINST for **comm**unity-specific **inst**ruction tuning, which is used to finetune LLMs to views of different communities, and (2) survey datasets COMMSURVEY for evaluating the alignment of the finetuned LLMs' views to the **comm**unities using **survey** questions. The pipeline is depicted in Figure 3.

4.1 Topic Modeling

For a domain, denote the combined corpus from all n communities by $\mathbb{D} = D_1 \cup D_2 \cup .... \cup D_n$. We use BERTopic (Grootendorst, 2022) on $\mathbb{D}$ to identify topics T. More details about BERTopic can be found in Appendix B.1. After topic modeling, each text d is assigned a topic $t(d)$. For each com-

[2]https://pushshift.io

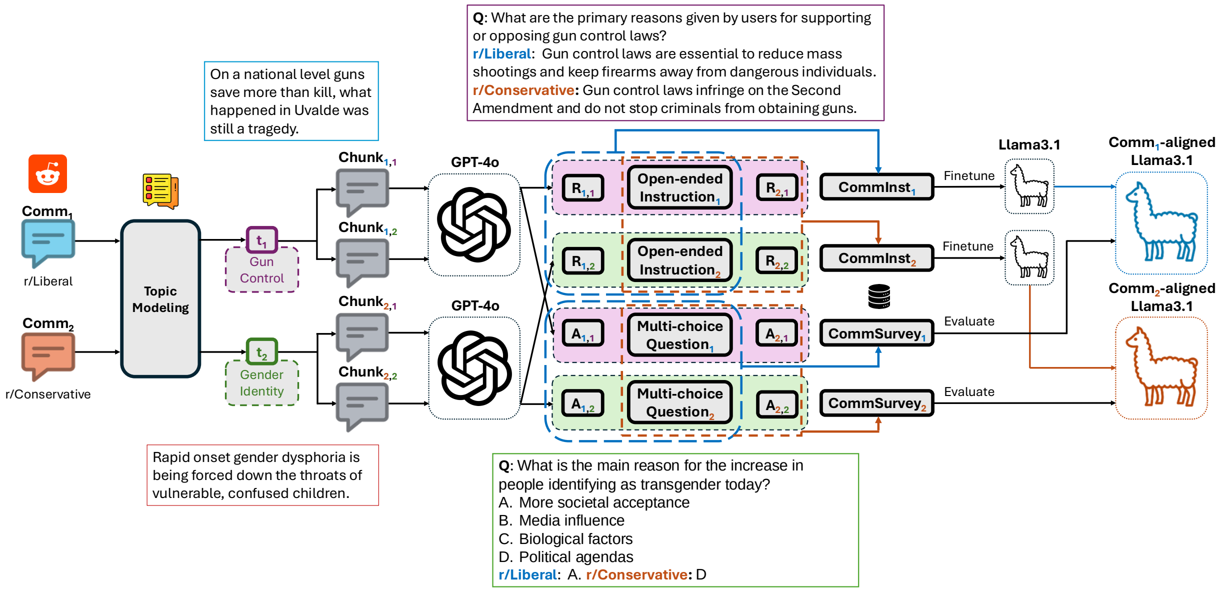


Figure 3: Overview of COMMUNITY-CROSS-INSTRUCT, with an illustrative example of the politics domain. (1) Data is collected for each community within the desired domain. (2) BERTopic clusters the data and identifies prominent topics. A chunk is a set of documents from a community on the same topic. Chunk$_{i,j}$ represents the chunk from community i on topic j. (3) For each topic, the advanced LLM is prompted with (i) on-topic chunks from each community and (ii) task definition of the instructional data generation (see Appendix B.2), which leads the LLM to generate (a) open-ended instruction-response pairs and (b) multi-choice question-answer pairs. R$_{i,k}$ represents the response of community i to instruction k; A$_{i,k}$ represents the answer of community i to question k. (4) The open-ended instructions across all topics, along with the corresponding responses of community i, are added to COMMINST$_i$, which is used to finetune a foundational LLM, to align the LLM to the community. (5) The multi-choice questions across all topics, along with the corresponding answers from community i, are added to COMMSURVEY$_i$, which is used to evaluate the finetuned LLM.

munity C_i, we split the text corpus D_i into smaller chunks $D_i = \{\hat{D}^i_1, \hat{D}^i_2, ...\}$, where each chunk $\hat{D}^i$ consists of 50 randomly sampled texts belonging to a single topic $t(\hat{D}^i)$. A chunk $\hat{D}^i$ on topic $t(\hat{D}^i)$ is considered an occurrence of the topic. To investigate different views across the communities, we only keep the topics that are discussed by at least $n-1$ communities. To balance the occurrences of different topics, for each topic in a community, we keep a maximum of 5 chunks. As a result, for *politics* and *diet*, there are 41 and 148 topics respectively. We present the keywords of 10 topics that appear 5 times in all communities in Appendix B.1. *Politics* topics include COVID vaccines, abortion, climate, guns. *diet* topics include body measures, keto diet, and eating disorder recovery.

4.2 Open-ended Instructions Generation

For a specific domain with n communities, we initialize n empty demonstration pools $\{\text{COMMINST}_i\}_{i=1}^n$. For a topic $t_k \in T$ that is discussed by m communities[3], from each community C_i, we randomly sample without replacement a text chunk $\hat{D}^i$ on t_k. We include the m sampled text chunks $\{\hat{D}^1, \hat{D}^2, ..., \hat{D}^m\}$ on topic t_k into a prompt and then use the prompt to query the advanced LLM (GPT-4o) to generate open-ended instructions and responses on the topic. The prompting template is shown in Appendix B.2. In the prompt, we first specify the topic by the topic keywords, and then the comments in the sampled chunks from different communities. Next, we instruct the model to generate an instruction that can be answered based on the texts, and that the instruction should elicit different responses from the communities. In addition, we instruct the model to **not rely on its pre-knowledge about the community, but solely focus on the given texts**. The generated instruction X is paired with different responses $\{Y_i\}_{i=1}^m$ from different communities, as shown in Figure 2(a). Each demonstration $\{X, Y_i\}$ is then added to the corresponding pool for the i-th community. We iteratively repeat this process until there are fewer than $n-1$ communities that have chunks left on topic t_k, then we shift to the next topic t_{k+1}. After

[3]As we require a topic to appear at least $n-1$ communities, $m = n$ or $m = n-1$.

the iterative generation process, each demonstration pool COMMINST_i is used to finetune a foundational LLM for community C_i. Please refer to Appendix B.3 for more details about COMMINST.

4.3 Multi-choice Surveys Generation

To measure the alignment of the finetuned foundational LLM to the community, we administer a survey of multi-choice questions to the model and evaluate the agreement between the model's responses and the community members' responses. However, manually designing surveys and collecting human responses from online communities is a non-trivial process, which is costly and time-consuming. Instead, we curate another collection of datasets called COMMSURVEY for evaluating the alignment of the finetuned LLMs' views to the communities.

We initialize n empty multi-choice question pools $\{\text{COMMSURVEY}_i\}_{i=1}^{n}$. Then, following the open-ended instruction-response pairs generation process in §4.2, we iteratively generate multi-choice questions and answers (Figure 2(b)) using the same advanced LLM (GPT-4o) and add them to the pools[4]. The prompting template is shown in Appendix B.2. Each question pool COMMSURVEY_i is used to evaluate a finetuned foundational LLM on community C_i. Please refer to Appendix B.3 for more details about COMMSURVEY.

We assume that the answers generated by the advanced LLM are the "semi-ground truths", and that they faithfully represent the views of the corresponding communities. As an empirical evaluation, we compute the pairwise agreement between different communities using their responses to questions in COMMSURVEY. The results are presented in Figure 5 in Appendix B.4. We observe significant polarization between the left-leaning and right-leaning communities, and strong agreement between the two right-leaning communities (*r/Conservative* and *r/AskThe_Donald*). However, in the *diet* domain where the communities are more practical, there is less polarization. For more rigorous evaluation, we verify the "semi-ground truths" by human annotation, as detailed in Appendix C.

[4]In practice we use the advanced LLM to generate open-ended instructions and multi-choice questions at the same time in a single query. For clarity of presentation, we articulate them separately.

5 Experiments

We align foundational LLMs to Reddit communities by finetuning them on relevant demonstrations from COMMINST. After finetuning, we evaluate the models using surveys from COMMSURVEY, comparing their responses to "semi-ground truths" to assess alignment.

5.1 Experimental Setup

Data Generation. We use OpenAI's GPT-4o Batch API to generate the datasets. For each query, three open-ended instructions and two multi-choice questions are created, as shown in Appendix B.2. The API costs approximately \$2 for the *politics* domain and \$5 for the *diet* domain.

Data Splitting. Among all the queries used to prompt the advanced LLM to generate data, we randomly select 85% of them and use the instructions generated from these queries as the training instructions denoted as COMMINST-TRAIN. For the rest 15% queries, we use the corresponding multi-choice questions as test questions, denoted as COMMSURVEY-TEST. In addition to randomly splitting the queries, we perform a topic-wise split, and make sure that the 85% training instructions, COMMINST-TRAIN-TOPIC, and 15% test questions, COMMSURVEY-TEST-TOPIC, do not cover the same topics.

Finetuning. We focus on two strong foundational LLMs – Llama-3.1-8B-Instruct (Dubey et al., 2024) and GPT-3.5-Turbo (Ouyang et al., 2022). For each community C_i, we finetune the LLM on COMMINST-TRAIN_i (§5.3) and $\text{COMMINST-TRAIN-TOPIC}_i$ (§5.4). The input and the output are the instruction and response verbatim. Llama-3.1 is finetuned with LLAMAFACTORY (Zheng et al., 2024), using both full and LoRA finetuning (Hu et al., 2022), with batch size 16 for 3 epochs. We report the results of the model (full or LoRA) that achieves better loss on the validation set (5% of the training data). Full finetuning takes around 3 minutes on 4 NVIDIA H100 GPUs, and LoRA finetuning takes around 30 seconds on 1 GPU. For GPT-3.5, we use the OpenAI API, which completes finetuning in around 10 minutes for 1 epoch, where the batch size is automatically determined.

Measuring Alignment. We administer survey questions from COMMSURVEY-TEST (§5.3) and COMMSURVEY-TEST-TOPIC (§5.4) to the finetuned foundational LLMs. Each prompt includes the question and options verbatim, followed by the instruction, "Select only one answer by stating either A, B, C, or D. Do not provide any additional explanation or rationale for your choice," to facilitate easier matching of the model's responses to the options. We set the temperature to 0.8, generating 20 responses per question, and take the majority vote. Accuracy is calculated by comparing the model's answers to the "semi-ground truths".

5.2 Baselines

LLM+Context. For each community C_i, we provide the vanilla (unfinetuned) LLM with the context about the community by appending 300 most relevant comments from it. For example, when prompting the LLM to answer a question in COMMSURVEY-TEST$_i$, we retrieve the most relevant 300 comments to the question, from the text chunks that are used to generate COMMSURVEY-TRAIN$_i$, by calculating the embedding similarity with *sentence-transformers*. The prompt is augmented with the instruction "`According to the following statements, learn the mindset and select only one most relevant answer by stating either A, B, C, or D. Do not provide any additional explanation or rationale for your choice. [comments]`". This baseline is inspired by the idea of in-context learning, where the LLM learns the community's mindset within the context provided in the prompt.

Providing the LLM with context becomes less efficient when the model is deployed to answer a large number of survey questions, as each question requires processing an extremely long input sequence due to the added comments. This significantly increases computational costs, memory usage, and processing time, making it inefficient for large-scale or real-time applications. In contrast, our framework, which finetunes the model on community-specific data, is a one-time effort. Once the model is finetuned, it can efficiently answer survey questions without requiring additional long contextual inputs, making it far more scalable and resource-friendly in practical use cases.

LLM+Steering. When prompting the vanilla LLM to answer survey questions, we steer it to mimic the community, by specifying in the prompt that "`Select only one answer` **`that best aligns with the opinions of members from subreddit r/[subreddit]`**". Steering the vanilla LLMs can nudge them to respond to the community. For fair comparison to this baseline, we also steer the finetuned LLM via COMMUNITY-CROSS-INSTRUCT, where the LLM is aligned to the community both in the finetuning and steering process.

It is worth noting that **steering only applies to predefined communities that are developed via a manually specified tag**, such as subreddits, where we can easily reference community names in the prompt. While this paper focuses on forum-based communities, COMMUNITY-CROSS-INSTRUCT generalizes to other online communities, including organically formed communities in the retweet network (Chu et al., 2024) or the news co-sharing network (Mosleh et al., 2021; He et al., 2024b), as long as their relevant text data is readily available. Although it is not always feasible to concisely summarize organically formed communities in text, COMMUNITY-CROSS-INSTRUCT allows for LLM alignment to such communities without requiring explicit textual descriptions.

5.3 Main Results

We compare the finetuned LLMs' generated answers to multi-choice questions to the "semi-ground truths" and report the accuracy in Figure 1 as a measure of alignment with the corresponding community. The LLMs are finetuned on COMMINST-TRAIN and evaluated on COMMINST-TEST.

Politics. In the *politics* domain, CROSSINST consistently outperforms CONTEXT across both LLMs, demonstrating the strength of finetuning on community-specific instructional data. By aligning LLMs to explicit community instructions, CROSSINST enables the models to better capture the nuances of political discourse, where ideological divides and subtle differences in language are critical. This makes the model more adept at accurately reflecting community values without being overwhelmed by noisy or redundant information, which is a common issue with CONTEXT. We observe that the a large portion of generated answers from Llama-3.1 using CONTEXT are texts irrelevant to the survey questions, indicating that it struggles in dealing with a lengthy prompt with 300 examples.

An important aspect to consider is that LLMs come with pre-existing knowledge about various communities, which is learned during pretraining on large-scale internet data. During STEERING, the model attempts to retrieve and apply this pre-

Subreddit	Llama-3.1-8B				GPT-3.5-Turbo			
	Without Steering		With Steering		Without Steering		With Steering	
	Context	CrossInst	Steering	Steering+CrossInst	Context	CrossInst	Steering	Steering+CrossInst
r/Liberal	8.3	**54.2**	55.8	**58.3**	41.7	**62.5**	45.8	**62.5**
r/NeutralPol	3.8	**50.0**	55.0	**63.3**	**55.0**	**55.0**	40.0	**50.0**
r/Anar_Cap	54.1	**76.7**	**70.0**	66.7	50.0	**66.7**	66.7	**73.3**
r/Conservative	18.4	**63.3**	**60.5**	56.7	50.0	**53.3**	**53.3**	**53.3**
r/AT_Donald	23.7	**56.7**	66.7	**70.0**	30.0	**56.6**	50.0	**63.3**
avg. politics	21.7	**60.2**	61.6	**63.0**	45.3	**58.8**	51.2	**60.5**
r/keto	26.0	**75.0**	67.0	**72.0**	56.0	**68.0**	**66.0**	**66.0**
r/WLAdvice	27.7	**60.6**	**65.2**	61.7	**59.6**	58.5	**66.1**	64.8
r/EDAnony	29.1	**72.1**	65.1	**69.8**	61.6	**72.1**	62.7	**66.3**
avg. diet	27.6	**69.2**	65.8	**67.8**	59.1	**66.2**	64.9	**65.7**

Table 1: Evaluation results of Llama-3.1-8B and GPT-3.5-Turbo on COMMSURVEY-TEST. The community-aligned LLMs (CROSSINST and STEERING+CROSSINST) are finetuned on COMMINST-TRAIN, so there is potential topic overlap between training and evaluation. For each model family, the results are divided into two groups, one without and one with steering. The best results in each group are highlighted in bold.

Subreddit	Llama-3.1-8B				GPT-3.5-Turbo			
	Without Steering		With Steering		Without Steering		With Steering	
	Context	CrossInst	Steering	Steering+CrossInst	Context	CrossInst	Steering	Steering+CrossInst
r/Liberal	11.1	**50.0**	58.3	**61.1**	41.7	**61.1**	63.9	**65.1**
r/NeutralPol	20.8	**60.0**	54.2	**65.0**	52.5	**65.0**	**67.5**	**67.5**
r/Anar_Cap	23.9	**56.5**	52.2	**56.5**	54.3	**58.7**	58.7	**71.7**
r/Conservative	20.0	**54.3**	52.5	**65.2**	50.0	**67.4**	**69.6**	65.2
r/AT_Donald	25.0	**56.5**	**57.5**	56.5	52.2	**60.9**	61.2	**63.0**
avg. politics	20.2	**55.5**	54.9	**60.9**	50.1	**62.6**	64.2	**66.5**
r/keto	29.7	**64.4**	**65.2**	63.6	59.3	**67.8**	61.0	**63.5**
r/WLAdvice	26.9	**63.5**	61.5	**62.5**	58.7	**62.5**	**63.5**	62.5
r/EDAnony	25.4	**55.9**	50.0	**54.9**	57.8	**60.8**	58.8	**61.8**
avg. diet	27.3	**61.3**	58.9	**60.3**	58.6	**63.7**	61.1	**62.6**

Table 2: Evaluation results of Llama-3.1-8B and GPT-3.5-Turbo on COMMSURVEY-TEST-TOPIC. The community-aligned LLMs (CROSSINST and STEERING+CROSSINST) are finetuned on COMMINST-TRAIN-TOPIC, so there is **no topic overlap** between training and evaluation. For each model family, the results are divided into two groups, one without steering and one with steering. The best results in each group are highlighted in bold.

existing knowledge to align its responses with the given community. However, this knowledge may not always be fully accurate or consistent with the actual values of the community being evaluated. For example, STEERING might prompt the model to draw on generalizations or stereotypes that are present in its pretraining data, which might not reflect the current or specific views of the subreddit in question. This inconsistency can explain why STEERING alone sometimes leads to suboptimal performance.

When combining STEERING with CROSSINST, a potential conflict arises between the knowledge gained from finetuning and the pre-existing community knowledge the model tries to retrieve through steering. The model is essentially being pulled in two directions: one based on the explicit, finetuned instructions from CROSSINST and the other based on its internalized, sometimes outdated, pretraining knowledge. This conflict can result in STEERING+CROSSINST underperforming compared to STEERING alone, as observed in certain subreddits, such as *r/Anarcho_Capitalism* on Llama-3.1. In these cases, the inconsistency between the steering prompts and the finetuned knowledge creates confusion for the model, leading to lower performance.

Diet. In the *diet* domain, we observe a similar pattern where CROSSINST significantly outperforms the CONTEXT baseline. Communities like *r/keto* and *r/EDAnonymous*, which are focused on specific health and lifestyle goals, benefit greatly from instructional finetuning. These communities are characterized by practical and focused discussions, and CROSSINST allows the model to adapt to the specific language and norms within these subreddits, ensuring a more accurate reflection of their values and objectives.

Interestingly, in *r/WeightLossAdvice*, STEERING alone performs the best across the four methods within each LLM. One possible reason for this is that in certain communities, particularly those that are well-represented in pretraining data, the model's pre-existing knowledge might be more aligned with the actual community values. In such communities, where strong ideological markers may already be embedded in the model's pretraining data, steering allows the model to retrieve this relevant information effectively, leading to stronger performance. In such cases, STEERING helps amplify the model's pre-learned alignment with the community, which can be sufficient for capturing the community's voice without the need for additional finetuning.

5.4 Out-of-Topic Generalizability

Our final goal is to create LLMs aligned to different communities, which can be used to answer any survey question from the perspective of the community. In the real world, the surveys may contain questions covering topics that do not appear in the finetuning data COMMINST. In this study, we finetune the LLMs on COMMINST-TRAIN-TOPIC, and evaluate them on COMMSURVEY-TEST-TOPIC, to make sure that there is no overlap in topic coverage.

The results are shown in Table 2. CROSSINST models continue to outperform the CONTEXT baselines, demonstrating their strong generalization capability to new topics, suggesting that CROSSINST maintains its robustness even when exposed to entirely new topics.

We observe that STEERING alone occasionally performs the best (e.g., *r/AskThe_Donald* on both LLMs, and *r/WeightLossAdvice* on GPT-3.5). We argue that this might be because ideologically-polarized or practical communities are consistent in their mindsets over time, so the LLM's pre-knowledge about them would well predict their ideologies in the future.

6 Related Works

6.1 Self-Improved Instruction Tuning.

A survey of instruction-tuning approaches (Zhang et al., 2023) outlines several methods for models to autonomously self-improve their instruction set. These include generating instructions for pre-existing texts (Li et al., 2023), eliciting interaction between different model iterations (Chen et al., 2024c), and bootstrapping from an existing instruction set to generate new ones (Wang et al., 2022; Chen et al., 2024b). Additionally, a human-in-the-loop framework builds upon self-generalization by iterating between human and machine-generated instructions (Guo et al., 2024). Despite these advances, these strategies require manual input. In this work, we build on *self-instruct*, a method to enhance instruction tuning by eliciting synthetic model instruction generations. We alter the original framework by removing the required set of manually written seed instructions (Wang et al., 2022). This work significantly contributes to the field by offering a scalable and fully autonomous solution to instruction tuning, paving the way for more adaptive and intelligent models.

6.2 Aligning LLMs to Subgroups

Existing work has aligned LLMs to different human subgroups to discover their mindset and opinions (Dorn et al., 2023, 2024). Subpopulation representative models (SRMs) (Simmons and Hare, 2023) can be used to emulate some characteristics of a particular subpopulation, particularly as LLMs can provide fine-grained, demographically-correlated outputs (Argyle et al., 2023a). Argyle et al. (2023b) find that exposing GPT-3 to thousands of socio-demographic backstories leads the model to obtain a complex understandings of sociological concepts.

To learn about partisan communities, Jiang et al. (2022) propose COMMUNITYLM by finetuning GPT-2 on tweets authored by prominent community members, and prompt the finetuned model to generate opinions about various political entities. He et al. (2024b) extends COMMUNITYLM to organically-formed online communities with more fine-grained ideologies. However, these finetuned GPT-2 models can only be used for text continuation tasks and cannot answer survey questions. In our paper, we finetune the foundational instruction-tuned LLMs to serve as their digital twins. The resulting LLMs retain the instruction-following capabilities and are able to complete various tasks as specified in the instructions, including survey completion.

6.3 Evaluating LLMs' Subgroup Alignment

Social scientists use surveys to systematically collect data from populations to characterize their beliefs, attitudes, opinions, and behaviors (Hill et al., 1997; Choi and Pak, 2005; Gordon, 1987; Goel and Salganik, 2010). LLM developers and practitioners focus on measuring LLM's alignment with different human subgroups (Santurkar et al., 2023; Durmus et al., 2023; He et al., 2024a) using use real-world survey responses as ground truth for comparing values. However, they often ignore under-represented groups, as it is difficult to administer surveys to people from those groups. To administer surveys, researchers sample individuals from a target population and ask them to manually respond to survey questions. This can be costly and slow, which limits the utility and timeliness of surveys. Our method addresses this challenge by using social media data that is easily accessible, even for those hard-to-reach subgroups. By automatically creating multi-choice survey questions and answers COMMSURVEY for these groups, our method enables a more comprehensive and cost-effective evaluation of LLMs' alignment with diverse groups.

7 Discussions

We aim to align LLMs to online communities through instruction-tuning so that the aligned LLMs can be flexibly surveyed to probe the communities' views on different topics. To prepare data for instruction tuning, we propose COMMUNITY-CROSS-INSTRUCT, and the automatic creation of the open-ended instruction-response pairs (COMMINST) is the core contribution. However, in order to demonstrate that the finetuned LLMs are indeed aligned to communities and can be treated as their digital twins, we propose to automatically generate survey questions and answers (COMMSURVEY) to efficiently evaluate LLM alignment. This is a secondary contribution of this paper, and we do not claim that COMMSURVEY is the best way to measure alignment. Therefore, our main goal is to propose a method to prepare LLMs as a proxy for surveying different online communities, rather than automatically creating such surveys.

8 Conclusion

In this paper, we introduced and explored COMMUNITY-CROSS-INSTRUCT, a fully automated framework to represent online communities and elicit their views. By automating the process of surveying diverse online communities, COMMUNITY-CROSS-INSTRUCT offers a cost-effective and efficient alternative to traditional survey methods used by social scientists. Through finetuning LLMs based on community-specific instructions and answers, COMMUNITY-CROSS-INSTRUCT enables the generation of accurate responses that align with the beliefs and perspectives of different online communities. This innovative framework has demonstrated its effectiveness in representing various communities, including political and diet forums on Reddit. COMMUNITY-CROSS-INSTRUCT opens up new possibilities for researchers to gain insights into online communities' diverse views and opinions, paving the way for deeper understanding and analysis of digital societies.

Moving forward, we will apply the framework to model organic online communities, such as those in the retweet network. In addition, we are interested in aligning LLMs to communities through reinforcement learning from human feedback (RLHF).

Limitations

Ignoring Thread Structure in Reddit Data. We treat each Reddit comment and submission as independent data points, without taking into account the thread structure in which they are embedded. Reddit discussions are inherently hierarchical, with comments often building upon or responding to previous comments and submissions. By ignoring this structure, we may lose important contextual information that could help capture the nuances of community interactions and opinions. Future work could explore leveraging the thread structure to better model the flow of conversations, capturing how community members engage with each other's ideas and how opinions evolve within discussions.

Identical LLM for Generating both Datasets. We use GPT-4o for generating both COMMINST for finetuning LLMs and COMMSURVEY for evaluating LLM alignment, as we couldn't find a different advanced LLM that was able to generate high-quality data as GPT-4o. This reliance on a single LLM introduces a limitation, as it may lead to a bias where the evaluation dataset (CommSurvey) is inherently aligned with the model used for instruction generation (CommInst), potentially overestimating the alignment accuracy of the finetuned models. To address this limitation, we plan to incorporate a diverse set of advanced LLMs as they become available in future iterations of our framework. Additionally, we will include human-in-the-loop validation to ensure the generated datasets maintain high quality and representativeness, mitigating any biases introduced by the single LLM dependency. Furthermore, we will explore cross-validation techniques and third-party evaluations to benchmark the performance of our finetuned models, ensuring the robustness and generalizability of our results beyond the influence of a single LLM.

Group Approximation. To approximate group-level behavior, community members represented in the minority might be inherently excluded. Further, measuring group alignment using a single-answer multiple choice questionnaire does not account for a more complex distribution of the various opinions within the community. We hope to build on this work by experimenting with survey designs that account for more diversity in communities.

Hallucination Potential. Adapting models to communities poses a risk of language models hallucinating or providing misinformation in their community representation. We hope to build upon this work with in-depth experiments on model hallucination and misrepresentation in the subgroup representation task.

Prompt Perturbations. LLM responses are sensitive to slight changes in prompts (Salinas and Morstatter, 2024). In this work, we work primarily with one prompt for instruction generation. We would be interested to see how the model's generated instructions shift with different prompting schema.

Ethics Statement

Finetuning LLMs towards Bias. Aligning LLMs with specific communities may result in models that appear more biased, as they are finetuned to the distinctive views and perspectives of those communities. However, this process is done solely for the purpose of accurately reflecting the values and attitudes prevalent within each community, rather than intentionally making the models more biased. Our goal is not to reinforce or amplify harmful biases, but rather to provide a computational tool that can represent the views of a given community for research purposes. To mitigate potential misuse, these models should be used in controlled environments and in contexts where understanding specific community perspectives is essential for social, cultural, or political research.

User Privacy and Consent. Users might not be informed of how their data (reflected in their posts) are used to produce digital twins that mirror their voices and the purpose of the survey constructed based on their data. Furthermore, automatically surveying communities can reveal unconsented insights of certain individuals or groups of people online. To address these ethical concerns, we implement several measures. Firstly, we ensure that all data used for creating digital twins is anonymized, stripping any personally identifiable information to protect user privacy. Secondly, we seek to aggregate data in a manner that focuses on community-level insights rather than individual-level analysis, thereby reducing the risk of unconsented personal exposure. Thirdly, we will obtain explicit consent from users where possible, clearly communicating the purpose and scope of our research. Lastly, we will adhere to ethical guidelines and institutional review board (IRB) requirements to ensure that our methods respect the privacy and rights of the individuals and communities involved.

Acknowledgements

We are grateful to Patrick Gerard, Ashwin Rao, and DJ Berry for helping review responses to surveys. This project was funded in part by the Defense Advanced Research Projects Agency (DARPA) under Agreement Nos. HR00112290021 and HR001121C0168.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023a. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023b. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Nicole Bergen and Ronald Labonté. 2020. "everything is perfect, and we have no problems": detecting and limiting social desirability bias in qualitative research. *Qualitative health research*, 30(5):783–792.

Kai Chen, Zihao He, Keith Burghardt, Jingxin Zhang, and Kristina Lerman. 2024a. Isamasred: A public dataset tracking reddit discussions on israel-hamas conflict. *arXiv preprint arXiv:2401.08202*.

Kai Chen, Zihao He, Jun Yan, Taiwei Shi, and Kristina Lerman. 2024b. How susceptible are large language models to ideological manipulation? *arXiv preprint arXiv:2402.11725*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024c. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Bernard CK Choi and Anita WP Pak. 2005. Peer reviewed: a catalog of biases in questionnaires. *Preventing chronic disease*, 2(1).

Minh Duc Chu, Aryan Karnati, Zihao He, and Kristina Lerman. 2024. Characterizing online eating disorder communities with large language models. *arXiv preprint arXiv:2401.09647*.

Rebecca Dorn, Lee Kezar, Fred Morstatter, and Kristina Lerman. 2024. Harmful speech detection by language models exhibits gender-queer dialect bias. *arXiv preprint arXiv:2406.00020*.

Rebecca Dorn, Negar Mokhberian, Julie Jiang, Jeremy Abramson, Fred Morstatter, and Kristina Lerman. 2023. Non-binary gender expression in online interactions. *arXiv preprint arXiv:2303.04837*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Abdulmotaleb El Saddik. 2018. Digital twins: The convergence of multimedia technologies. *IEEE multimedia*, 25(2):87–92.

Sharad Goel and Matthew J Salganik. 2010. Assessing respondent-driven sampling. *Proceedings of the National Academy of Sciences*, 107(15):6743–6747.

Randall A Gordon. 1987. Social desirability bias: A demonstration and technique for its reduction. *Teaching of Psychology*, 14(1):40–42.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hongyi Guo, Yuanshun Yao, Wei Shen, Jiaheng Wei, Xiaoying Zhang, Zhaoran Wang, and Yang Liu. 2024. Human-instruction-free llm self-alignment with limited samples. *arXiv preprint arXiv:2401.06785*.

Zihao He, Siyi Guo, Ashwin Rao, and Kristina Lerman. 2024a. Whose emotions and moral sentiments do language models reflect? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6611–6631, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zihao He, Ashwin Rao, Siyi Guo, Negar Mokhberian, and Kristina Lerman. 2024b. Reading between the tweets: Deciphering ideological stances of interconnected mixed-ideology communities. *arXiv preprint arXiv:2402.01091*.

James J Heckman. 1990. Selection bias and self-selection. In *Econometrics*, pages 201–224. Springer.

Anthony Hill, Julian Roberts, Paul Ewings, and David Gunnell. 1997. Non-response bias in a lifestyle survey. *Journal of Public Health*, 19(2):203–207.

Valentin Hofmann, Hinrich Schütze, and Janet B Pierrehumbert. 2022. The reddit politosphere: A large-scale text and network resource of online political discourse. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1259–1267.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. Communitylm: Probing partisan worldviews from language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6818–6826.

Julie Jiang, Emily Chen, Shen Yan, Kristina Lerman, and Emilio Ferrara. 2020. Political polarization drives online conversations about covid-19 in the united states. *Human Behavior and Emerging Technologies*, 2(3):200–211.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Mohsen Mosleh, Gordon Pennycook, Antonio A Arechar, and David G Rand. 2021. Cognitive reflection correlates with behavior on twitter. *Nature communications*, 12(1):921.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv.org/abs/2203.02155*, 13.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *arXiv preprint arXiv:2401.03729*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Gabriel Simmons and Christopher Hare. 2023. Large language models as subpopulation representative models: A review. *arXiv preprint arXiv:2310.17888*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Reddit Data

A.1 Subreddits

We briefly introduce each subreddit in Table 3. Statistics of data from different subreddits are shown in Table 4.

B Instructional Data Generation

B.1 Topic Modeling with BERTopic

Following the pipeline of BERTopic (Grootendorst, 2022), we first obtain the text embeddings using *all-mpnet-base-v2* (Song et al., 2020). The embedding dimensions are reduced with UMAP, with n_neighbors 15 and n_components 5. Then the embeddings are clustered using HDBSCAN, with min_cluster_size 40. After fitting the model, we remove texts that are not assigned any topics. Table 5 and Table 6 present the top 10 most frequent topics in *politics* and *diet*.

B.2 Prompting Template for CommInst

The prompting template for generating the open-ended instruction-response pairs in CommInst and multi-choice question-answer pairs in CommSurvey is shown in Figure 4.

B.3 Statistics of CommInst and CommSurvey

Table 7 describes the number of open-ended instruction-response pairs in CommInst and multi-choice question-answer pairs in CommSurvey. The variance is due to their different topic coverage.

B.4 Community Agreement Analysis

For a pair of communities C_i and C_j, we identify their common survey questions in CommSurvey_i and CommSurvey_j, and compute their agreement on the questions using Cohen's Kappa. The agreement matrices for *politics* and *diet* are shown in Figure 5.

In the *politics* domain, we observe significant polarization between communities. For example, the agreement between *r/Liberal* and both *r/Conservative* (0.01) and *r/AskThe_Donald* (0.03) is extremely low, indicating the deep ideological divide between these left-leaning and right-leaning communities. This low agreement reflects how vastly different their responses are to the same political survey questions, aligning with the broader patterns of political polarization seen across online platforms. On the other hand, *r/Conservative* and *r/AskThe_Donald* exhibit much higher agreement (0.59), highlighting the ideological overlap between these conservative-leaning communities, likely reflecting their shared values and viewpoints on key political issues.

Meanwhile, *r/NeutralPolitics* shows moderate agreement with both *r/Anarcho_Capitalism* (0.24) and *r/Liberal* (0.28), suggesting that as a centrist or neutral forum, it shares certain perspectives with both libertarian and liberal ideologies. Additionally, *r/Anarcho_Capitalism* shows higher agreement with conservative-leaning subreddits, such as *r/Conservative* (0.36) and *r/AskThe_Donald* (0.44), reflecting shared economic or libertarian principles, despite divergences on other political issues.

In the *diet* domain, the relationships between the communities are more practical and less polarized compared to *politics*. Communities like *r/keto* and *r/WeightLossAdvice* exhibit moderate

Subreddit	Description
r/Liberal	r/Liberal is a subreddit dedicated to discussions and news from a liberal political perspective, focusing on progressive policies, social justice, and left-leaning viewpoints.
r/NeutralPolitics	r/NeutralPolitics is a community dedicated to evenhanded, empirical discussion of political issues. It is a space to discuss policy and the tone of political debate.
r/Anarcho_Capitalism	r/Anarcho_Capitalism is a subreddit dedicated to discussing free-market capitalist anarchism and related topics, advocating for a society where voluntary interactions enhance liberty and opportunity for all.
r/Conservative	r/Conservative offers the largest space on Reddit for fiscal and social conservatives to explore and discuss political and cultural issues from a distinctly conservative perspective.
r/AskThe_Donald	r/AskThe_Donald is a subreddit that serves as a hub for supporters of former President Donald Trump, fostering discussions around conservative politics, pro-Trump views, and right-wing ideologies.
r/keto	r/keto is a community for sharing experiences and advice about the low-carb Ketogenic diet, which supports a range of health conditions from diabetes to epilepsy
r/WeightLossAdvice	r/WeightLossAdvice is a subreddit where users share tips, strategies, and support for healthy and sustainable weight loss, with a focus on practical advice and personal experiences.
r/EDAnonymous	r/EDAnonymous is a subreddit that provides a supportive, anonymous space for individuals struggling with eating disorders to share their experiences, seek advice, and offer encouragement on the path to recovery.

Table 3: Reddit forums used in this study.

	r/Lib	**r/NeutralPol**	**r/Anarcho_Cap**	**r/Conserv**	**r/ATDonald**
# of comments	31,233	35,725	670,686	2,243,842	142,543
# of submissions	5,243	2,072	22,551	153,813	15,645
	r/keto	**r/WLAdvice**	**r/EDAnonymous**		
# of comments	603,466	225,635	347,477		
# of submissions	49,571	42,840	99,925		

Table 4: Number of comments and submissions in each subreddit.

agreement (0.39), likely reflecting shared goals related to weight loss and health, even though they may emphasize different strategies. Similarly, *r/EDAnonymous* and *r/keto* (0.38) show moderate overlap, suggesting that while their focus areas are different—one being centered on eating disorders and the other on ketogenic diets—they share some common ground in their approaches to health and diet-related topics.

The strongest alignment within the *diet* domain is between *r/WeightLossAdvice* and *r/EDAnonymous* (0.41), which likely stems from overlapping concerns about diet and body image, which are central to both communities' discussions.

C Human Annotation

We verify the faithfulness of COMMSURVEY as the "semi-ground truths" for the communities via human annotation. We also tried directly posting the survey questions to the subreddits and having community members answer them. However, our posts were immediately removed by moderators and our accounts were banned from certain subreddits. This further indicates the difficulty of directly surveying online communities.

C.1 Annotator Recruiting

The four annotators for the *politics* domain closely follow political news and are knowledgeable about the platforms of the parties represented in our sample. The three annotators for the *diet* domain are active members of diet and diet and have substantial knowledge of diet-related topics and trends. To ensure the annotators were capable of accurately answering the survey questions for different communities, we implemented a thorough selection and training process. Annotators were chosen based on their demonstrated expertise and familiarity with the relevant domains. Additionally, we conducted preliminary tests where annotators answered a subset of questions, and their responses were evaluated for consistency and accuracy.

C.2 Annotation Process

We randomly sampled 20 questions from COMMSURVEY-POLITICS (Figure 6) and 10 questions from COMMSURVEY-DIET (Figure 7) for the annotators to answer. When answering

idx	Topic Keywords
1	vaccine, covid, vaccinated, vaccines, flu, unvaccinated, vaccination, mrna, pandemic, data
2	climate, prices, solar, change, co2, emissions, coal, cars, fuels, earth
3	ballots, fraud, election, voter, ballot, evidence, voters, machines, audits, rigged
4	abortion, abortions, fetus, roe, murder, unborn, conception, choice, womb, cells
5	ukraine, russia, putin, nato, war, ukrainians, crimea, invade, conflict, eu
6	twitter, facebook, musk, google, social, platforms, companies, users, censorship, ban
7	gun, guns, firearms, shootings, firearm, rifles, laws, amendment, ammo, armed
8	impeachment, mueller, fbi, impeach, documents, collusion, comey, congress, whistleblower, crimes
9	gender, trans, transgender, lgbt, dysphoria, children, sexuality, identity, pronouns, genders
10	capitol, blm, antifa, riots, riot, insurrection, rioters, protesters, terrorism, january

Table 5: Top-10 most frequent topics in *politics*. Each topic is represented by the topic-10 keywords.

idx	Topic Keywords
1	keto, protein, ketosis, started, macros, back, weeks, carbs, diet, calories
2	skinny, thin, girls, like, overweight, body, underweight, weight, hate, myself
3	thank, therapy, relapse, recover, help, life, support, yourself, proud, treatment
4	scale, week, water, lose, weighing, daily, trend, fluctuates, months, plateau
5	potassium, sodium, magnesium, electrolytes, mg, supplement, ketoade, citrate, 5000mg, chloride
6	alcohol, drinking, beer, drink, drinks, liquor, alcoholic, gin, booze, drinker
7	protein, macros, fat, min, kcal, 20g, carbs, bf, lean, need
8	binging, restricting, binges, eating, hunger, control, bingeing, cycle, guilt, feeling
9	ed, eds, recovery, recover, therapist, treatment, me, help, coping, talk
10	coffee, tea, chocolate, brew, creamer, starbucks, latte, unsweetened, flavors, sweetener

Table 6: Top-10 most frequent topics in *diet*. Each topic is represented by the topic-10 keywords.

each question, the annotators were instructed to search the relevant posts from the subreddit and learn their views from the discussions, instead of solely relying on their pre-assumption about the community. Each annotator filled out the question for a different subset of the communities, as shown in Table 8. If multiple annotators annotated for a community, they had another round of discussions to resolve their disagreement. As a result, for community, the annotators delivered one set of annotations of the survey questions.

C.3 Annotation Evaluation

For each subreddit, we compare the LLM-generated “semi-ground truths” to the label from human annotators, and present the accuracy in Table 9. In both domains, “semi-ground truths” achieve strong agreement with human annotations for most subreddits, which gives confidence that the advanced LLM-generated survey answers can be used as “semi-ground truth” to evaluate fine-tuned foundational LLMs.

	r/Lib	r/NeutralPol	r/Anarcho_Cap	r/Conserv	r/ATDonald
# of instructions	234	234	300	303	303
# of survey questions	156	156	200	202	202
	r/keto	**r/WLAdvice**	**r/EDAnonymous**		
# of instructions	1,032	885	870		
# of survey questions	688	590	580		

Table 7: Number of generated open-ended instruction-response pairs in COMMINST$_i$ and multi-choice question-answer pairs in COMMSURVEY$_i$, for each community.

	r/Lib	r/NeutralPol	r/Anarcho_Cap	r/Conserv	r/ATDonald
Annotator 1			x	x	
Annotator 2	x		x	x	x
Annotator 3	x		x		
Annotator 4	x			x	
	r/keto	**r/WLAdvice**	**r/EDAnonymous**		
Annotator 1	x	x	x		
Annotator 2	x		x		
Annotator 3		x	x		

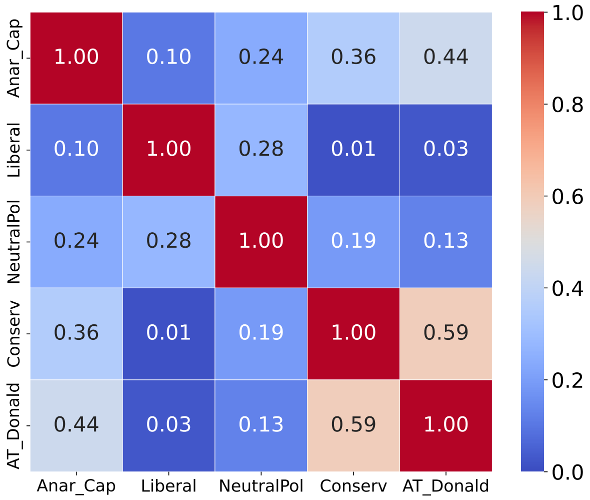
Table 8: Communities that each annotator annotated for.

	r/Lib	r/NeutralPol	r/Anarcho_Cap	r/Conserv	r/ATDonald
Human-LLM Acc	0.75	NA	0.65	0.55	0.55
	r/keto	**r/WLAdvice**	**r/EDAnonymous**		
Human-LLM Acc	0.7	0.6	0.5		

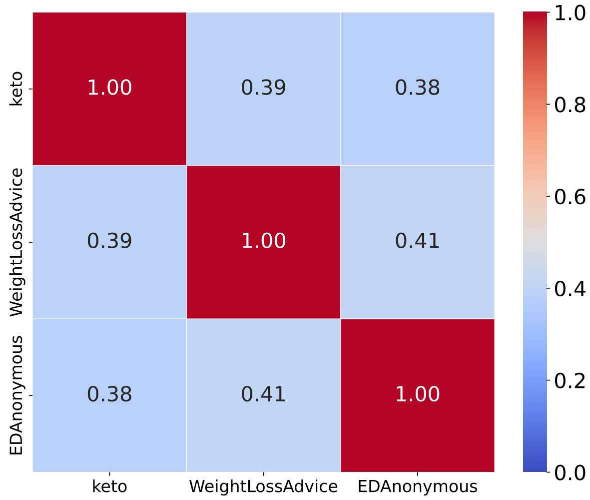
Table 9: Agreement between human annotators' survey responses and the advanced LLM (GPT-4o) generated survey answers, measured by accuracy.

```
Below are comments from 5 different subreddits related to a topic. The topic can be represented using these keywords: ballots, fraud, election, voter, ballot, evidence, voters, machines, audits, rigged.


Comments from r/Liberal
Comment 1: xxx
…
Comment 50: xxx

Comments from r/Anarcho_Capitalism
Comment 1: xxx
…
Comment 50:xxx


Comments from r/NeutralPolitics
Comment 1: xxx
…
Comment 50: xxx


Comments from r/Conservative
Comment 1: xxx
…
Comment 50: xxx

Comments from r/AskThe_Donald
Comment 1: xxx
…
Comment 50: xxx


Write 5 questions (Q1 through Q5) on this topic that can be answered based on these comments. For a subreddit, each question should be answered in a way that the members from the subreddit would do, and the answers should echo the comments shown above. Do NOT rely on your background knowledge about the specific subreddits to answer the questions. The questions should be low-level, detailed, trigger different responses that differentiate between different subreddits. Don't ask too high-level questions. The questions should not be in the style of reading comprehension ones, and they are intended for members in the subreddits to answer. The questions should not contain "comment" in them. Each question should be paired with answers from all {n_nonzero_subreddits} subreddits. For the first 3 questions, they are open-ended. The answers should be concise (fewer than 32 tokens), legible, grammatically correct. For the second 2 questions, they are multi-choice questions and are associated with four options (A through D). Try to come up questions that members from different subreddits would answer differently. Below is the format of generated questions.


Q1: [open-ended question]
Response from r/Liberal: [answer in clean text]
Response from r/Anarcho_Capitalism: [answer in clean text]
Response from r/NeutralPolitics: [answer in clean text]
Response from r/Conservative: [answer in clean text]
Response from r/AskThe_Donald: [answer in clean text]

Q2: …

Q3: …

Q4: [multi-choice question]
A.xxx
B.xxx
C.xxx
D.xxx
Answer from r/Liberal: A/B/C/D
Answer from r/Anarcho_Capitalism: A/B/C/D
Answer from r/NeutralPolitics: A/B/C/D
Answer from r/Conservative: A/B/C/D
Answer from r/AskThe_Donald: A/B/C/D

Q5: …
```

Figure 4: Prompting template to generate COMMINST and COMMSURVEY in the *politics* domain.



(a) politics



(b) diet

Figure 5: Pairwise agreement between different communities, measured by Cohen's Kappa.

1. **Which measure do users from these subreddits most commonly support to address school shootings?**
   A. More gun control laws
   B. Arming teachers and increasing school security
   C. Mental health interventions
   D. Stricter penalties for illegal gun possession
2. **How should society address the issue of transgender athletes in sports?**
   A. Allow them to compete according to their gender identity
   B. Create separate categories for transgender athletes
   C. Require them to compete according to their biological sex
   D. Ban transgender athletes from competitive sports
3. **How do members of your subreddit view the term "echo chamber"?**
   A. As a necessary evil
   B. As something to avoid
   C. As a positive thing
   D. As a misunderstood concept
4. **According to the comments, what is the main reason some users believe the US should not get involved in Ukraine?**
   A. It's Europe's problem
   B. Fear of nuclear escalation
   C. Financial burden
   D. Corruption in Ukraine
5. **What is a commonly suggested alternative to taking on large student loans for college?**
   A. Community college
   B. Trade school
   C. Scholarships and grants
   D. All of the above
6. **How do different groups perceive the involvement of Trump in the January 6th events?**
   A. He incited the violence directly
   B. His rhetoric contributed to the atmosphere but didn't incite violence directly
   C. He had no role; it was spontaneous
   D. It was a false flag operation
7. **Which argument is often used by pro-choice supporters to counter the claim that abortion is murder?**
   A. Fetus viability criteria
   B. Women's right to bodily autonomy
   C. Economic impact on unwanted pregnancies
   D. Historical context of abortion laws
8. **How should the government address tax loopholes used by the wealthy?**
   A. Increase audits and enforcement
   B. Simplify the tax code and eliminate loopholes
   C. Maintain current practices
   D. Increase taxes across all brackets
9. **What is the stance on abortions in cases of rape from your subreddit?**
   A. Always unacceptable
   B. Acceptable only if the mother's life is at risk
   C. Acceptable in the first trimester
   D. Acceptable in all cases
10. **Which policy is believed to most impact gasoline prices?**
   A. Increasing oil drilling
   B. Subsidizing EVs
   C. Imposing carbon taxes
   D. Restricting oil imports

Figure 6: Ten multiple choice survey questions (out of the twenty) in CommSurvey for human evaluation in the *politics* domain.

1. **What is a realistic weight loss rate per week?**
   A. 5 pounds
   B. 0.5 to 2 pounds
   C. 3 to 4 pounds
   D. 4 to 6 pounds
2. **When experiencing a relapse, what should be your next step?**
   A. Give up and wait for better conditions
   B. Reflect, learn, and try again
   C. Punish yourself for failing
   D. Seek immediate professional help
3. **What strategy do you use to avoid eating food you don't want during family meals?**
   A. Politely decline and change the subject
   B. Take the food and don't eat it
   C. Explain your dietary restrictions
   D. Avoid family meals altogether
4. **What is your primary reason for tracking calories?**
   A. Weight loss
   B. Ensuring proper nutrition
   C. Managing eating habits
   D. I don't track calories
5. **What role do genetics play in where you store and lose fat first?**
   A. No role
   B. Minor role
   C. Major role
   D. Complete control
6. **How do community members feel about using protein shakes as meal replacements?**
   A. They are convenient but should not replace whole foods entirely
   B. They are effective and can help with quick weight loss
   C. They cause bowel issues and are not sustainable long-term
   D. They are great for muscle recovery but should be used sparingly
7. **What is one method suggested to help control emotional eating?**
   A. Tracking macros
   B. Going for a walk
   C. Drinking more coffee
   D. Joining a support group
8. **How do you handle unsolicited advice or comments about your weight?**
   A. I ignore them and focus on my goals
   B. I try to educate the person giving advice
   C. I get upset but don't confront them
   D. I change the topic immediately
9. **What do members of your subreddit often replace breakfast with?**
   A. Black coffee
   B. Protein shakes
   C. Small snacks
   D. Nothing, they wait until lunch
10. **When you have a cheat meal, what is the most important aspect to consider?**
    A. The type of food you are eating
    B. The timing of the cheat meal
    C. Balancing it with exercise
    D. How it fits into your weekly calorie intake

Figure 7: Ten multiple choice survey questions COMMSURVEY for human evaluation in the *diet* domain.