# Improving and Assessing the Fidelity of Large Language Models Alignment to Online Communities

**Anonymous ACL submission**


## Abstract

Large language models (LLMs) have shown promise in representing individuals and communities, offering new ways to study complex social dynamics. However, effectively aligning LLMs with specific human groups and systematically assessing the fidelity of the alignment remains a challenge. This paper presents a robust framework for aligning LLMs with online communities via instruction-tuning and comprehensively evaluating alignment across various aspects of language, including authenticity, emotional tone, toxicity, and harm. We demonstrate the utility of our approach by applying it to online communities centered on dieting and body image. We administer an eating disorder psychometric test to the aligned LLMs to reveal unhealthy beliefs and successfully differentiate communities with varying levels of eating disorder risk. Our results highlight the potential of LLMs in automated moderation and broader applications in public health and social science research[1].


## 1 Introduction

**[Warning: This paper discusses eating disorders, which some may find distressing.]**

Large language models (LLMs) have demonstrated an unparalleled ability to generate detailed, nuanced responses to natural language prompts, suggesting potential for their use in creating high-fidelity proxies of people (Simmons and Hare, 2023). Leveraging LLMs to create digital representations of individuals and human groups could provide powerful tools for studying psycho-social dynamics of human behavior, enhancing and personalizing human-computer interactions, and moderating online spaces to promote prosociality, enhance safety, and reduce harm.

To create digital representations of human subgroups, researchers have aligned LLMs to subgroups via steering—i.e., instructing the LLM to mimic the target subgroup by specifying its core characteristics in the prompt (Santurkar et al., 2023; Durmus et al., 2023). However, this approach does not solve LLMs' misalignment with the target subgroup. Other methods for aligning LLMs to human subgroups include finetuning the base LLMs[2] like GPT-2 on data generated by specific subgroups (Jiang et al., 2022b; He et al., 2024c). Although this method can produce models that reflect the linguistic patterns of the target population, these finetuned models often lack the flexibility to follow diverse instructions, limiting their utility.

Another key challenge in developing digital representations of human subgroups is evaluating the alignment between the LLM and the target group. Traditional methods compare the LLM's responses to surveys with those of the target group (Santurkar et al., 2023; Durmus et al., 2023), but this approach misses critical aspects of human expression like emotional reactions (He et al., 2024b). Additionally, surveys are not scalable due to their cost and time requirements, particularly for marginalized or hard-to-reach groups. Besides, mapping organically-formed online communities to clear demographic identities greatly complicates alignment evaluation.

To address these challenges, we propose a framework for aligning LLMs with online communities through instruction-tuning in a fully unsupervised manner. Additionally, we introduce a comprehensive evaluation framework to assess alignment. This enables the creation of high-fidelity digital representations of online communities, paving the way for new research into human behavior, content moderation, public mental health, and social science. As one example, we can administer psychometric instruments to these digital proxies to identify at-risk communities prone to psychopathologies.

[1] Our data and code will be available upon publication.

[2] By "base LLMs" we refer to LLMs that are not finetuned to follow instructions

Our alignment method takes a corpus of social media posts (e.g., tweets) from an online community and creates a set of demonstrations (instruction-response pairs). In each demonstration (Figure 1), the instruction specifies the task (e.g., generate a tweet) with the response being the exact tweet. We then finetune an LLM on these demonstrations to align it with the community. To assess alignment, we generate a synthetic text corpus using the finetuned LLM and compare it to the original posts along four key aspects: 1) authenticity, 2) emotional tone, 3) toxicity, and 4) harm. These dimensions capture the essential features of online social communication, ensuring the aligned LLM accurately reflects the semantics, affect, and style of the target group's discourse.

**Instruction**: What would you tweet?
**Response**: most of the time the only thing i want in the whole world is to be skinny and lose weight

Figure 1: An example of a demonstration from a pro-eating disorder community, where the response is a tweet from the community.

To demonstrate our framework's utility, we analyze Twitter discussions in diet and fitness communities, where harmful attitudes about body image exist. While these communities can offer support and encouragement, they often promote unhealthy behaviors and beliefs that put people at risk for developing eating disorders (EDs). Applying traditional psychometric instruments to screen individuals for EDs is impractical and potentially unethical; instead, we use our framework to align LLMs with these communities through automatically generated demonstrations and evaluate alignment to show that the finetuned LLMs outperform baseline LLMs in creating high-fidelity proxies of online communities. We then apply an ED screening questionnaire to community-aligned LLMs, revealing significant differences between communities: pro-anorexia communities show a high risk of unhealthy behaviors, while those critical of the diet culture exhibit the lowest risk. These findings highlight our framework's potential for automated moderation by distinguishing communities with varying levels of ED risk.

Our framework offers a scalable approach to modeling and analyzing online communities, with broad implications for understanding and mitigating harmful behaviors. By applying this method to ED communities, we demonstrate its potential to contribute to public health and social science research, highlighting the value of LLMs in studying complex social dynamics.

## 2 Related Work

**Aligning LLMs to Subgroups** There is growing literature (Simmons and Hare, 2023) on aligning LLMs to diverse human subgroups to mimic their language and mindsets. Researchers have aligned LLMs by steering them towards particular demographic groups (Santurkar et al., 2023; Durmus et al., 2023; He et al., 2024b), e.g., by including the target subgroup in the prompt. However, their findings reveal that steering does not solve the model's misalignment with the target subgroup. Moreover, it is non-trivial to summarize an organically-formed community (e.g., communities in retweet networks) into a concise description that can be used in steering.

Others have aligned LLMs with different subgroups by finetuning the model on the text generated by the subgroups. Jiang et al. (2022b) propose COMMUNITYLM by finetuning two GPT-2 models (Radford et al., 2020) using causal language modeling on tweets from liberals and conservatives, and probing their worldviews from their corresponding finetuned models. He et al. (2024c) extend COMMUNITYLM to probe the views of organically-formed online communities and make use of the interactions between different communities. However, GPT-2 is not instruction-tuned and is not able to answer questions in various formats, like the psychometric instruments we discuss in §6. He et al. (2024a) use an advanced LLM (e.g., Claude-3) to distill knowledge from the community's raw data and generate high-quality instruction-response pairs, where the instructions aim to query the community' mindset, and the corresponding responses are abstracted from the ideas conveyed in the raw data. The generated instruction-response pairs are used to finetune a foundational LLM (e.g., Llama-3) for alignment. However, the API costs of querying the advanced LLM are non-negligible.

**Evaluating LLMs' Alignment to Subgroups** Existing works (Santurkar et al., 2023; Durmus et al., 2023) measure an LLM's alignment with a target subgroup using multi-choice surveys. Specifically, they prompt the LLM to respond to a survey question from the perspective of a subgroup and then compare the LLM-generated distribution over the different options of the question to that of the survey respondents belonging to the target group. However, collecting survey responses can be costly and time-consuming. Also, responses on sensitive topics, such as mental health, may be biased due to stigma and social desirability bias (Gordon, 1987). Our framework evaluates LLM alignment by comparing the LLM-generated synthetic text to the original text written by humans is significantly more scalable, unbiased, and cost-effective.

**LLMs and Psychometric Tests** LLMs can respond to psychometric instruments that were originally designed to assess individual human psychological and emotional states. Researchers have administered these instruments to LLMs to probe their decision-making processes, reasoning abilities, cognitive biases, and other psychological traits—Pellert et al. (2024) call this practice "AI Psychometrics". Coda-Forno et al. (2023) show that GPT-3.5 generated consistently high scores on responses to a widely used anxiety questionnaire. Tanmay et al. (2023) measure GPT-4's moral reasoning abilities by applying an ethical measurement instrument for individuals. Researchers also administer personality tests to LLMs to identify their personality traits (Jiang et al., 2022a; Lu et al., 2023; Serapio-García et al., 2023). In contrast, we apply psychometric questionnaires to a specific online community—via a finetuned LLM—to learn more about the mindset of the community members. We show that this helps reveal unhealthy beliefs within these communities and even identify pathologies, like harmful cognitions associated with EDs.

**Online Eating Disorders Communities** Pro-ED (pro-anorexia) communities are online spaces that frame EDs as a lifestyle rather than an illness. While they provide social support, a sense of belonging, and empathy for stigmatized individuals (Juarascio et al., 2010; Oksanen et al., 2016; Yeshua-Katz and Martins, 2013; McCormack, 2010), they also promote harmful behaviors, such as weight loss tips and "thinspiration" imagery, exacerbating EDs and psychological distress (Ging and Garvey, 2018; Mento et al., 2021).

Previous research has focused on identifying harmful content and at-risk users within these communities. For example, Chancellor et al. (2016a) developed a lexical classifier to predict posts moderated by Instagram for self-harm content, comparing pro-recovery and pro-ED communities (Chancellor et al., 2016b,c). In contrast, our study examines the collective mindset of these communities as expressed through their discussions, using advanced language models to assess attitudes toward mental health and body image issues.

## 3 Communities in Online Discussions

We collect online conversations related to EDs and identify organically-formed communities within the broader context of weight loss, dieting, and fitness discussions.

### 3.1 Data Collection

We collected 2.6M tweets from 557K users from October 2022 to March 2023 using ED-related keywords to query Twitter. For keywords, we start with a set of terms that promote ED (Chancellor et al., 2016a; Pater et al., 2016), such as *thinspo* (thin inspiration), *proana* (pro-anorexia), and *promia* (pro-bulimia), among others. We remove spam terms yielding unrelated content, such as *skinny*. We expanded the query set to include closely related topics such as diet and weight loss through terms such as (*ketodiet*, *weightloss*, . . .), and anti-diet culture (*bodypositivity*, *dietculture*, . . .). See Appendix A.1 for the full set of search terms.

### 3.2 Identifying Communities

We construct a retweet network where nodes are users, and (undirected) edges link users who retweet each other. Visualization of the retweet network is shown in Figure 7 in Appendix A.2. We use Louvain modularity maximization (Blondel et al., 2008) to identify dense clusters of users who frequently retweet one another. These clusters are organically formed based on shared interests, consisting of users who pay attention to each other. Detailed statistics and content of the clusters are shown in Table 3 and Figure 6 in Appendix A.2. Based on the thematic profiling of discussions (Table 4 in Appendix A.2), we categorize the clusters into six communities: *Pro-ED*, *Keto & Diet*, *Weight Loss Drugs*, *Body Image*, *Healthy Lifestyle & Weight Loss*, and *Anti-ED*. This categorization is intended to label the communities for easy reference in subsequent analyses, and the labels do not cover the full spectrum of discussions in the communities.

After identifying communities in the retweet network, we clean the tweets by removing URLs, mentions, hashtags, and emojis, and we filter out retweets and comments, only keeping the original tweets. To ensure high-quality data, we compute the perplexities of the tweets using BERTweet (Nguyen et al., 2020) that is pretrained on tweets, and select a maximum of 10K highest quality (i.e., lowest perplexity) tweets from each community. If there are fewer than 10K tweets from the community, we keep all of them. The numbers of tweets from the community *Pro-ED*, *Keto & Diet*, *Body Image*, *Anti-ED*, *Healthy Lifestyle & Weight Loss*, and *Weight Loss Drugs* are 10K, 10K, 3.3K, 2.9K, 10K, and 10K respectively.

# 4 Aligning LLMs to Communities

There are $n$ online communities $\{C_1, C_2, ...., C_n\}$ on a topic (e.g., EDs), each characterized by their own beliefs and perspectives. Members of a community $C_i$ produce a body of text $D_i$ (e.g., tweets) that reflects their collective opinions and behaviors. Our objective is to align an LLM $f$ to each specific community $C_i$ by training it on the corresponding text corpus $D_i$. The resulting model, $f_i'$, should capture the community's unique collective mindset, enabling it to generate responses that authentically represent the community's voice.

## 4.1 Constructing Instruction-Response Pairs

To align an LLM $f$ to a particular community $C$, we employ a finetuning process using a set of demonstrations (instruction-response pairs). We propose creating demonstrations based on the community's raw text corpus $D$, which is cost-efficient, and yet curated demonstrations can be used to finetune a foundational LLM (e.g., Llama-3) effectively.

For each community $C_i$, we use tweets in $D_i$ as the responses verbatim in the demonstrations. To create instructions that can be answered by the tweets, we focus on the tweet generation task. We curate an instruction pool of 20 different instruction templates (Table 5 in Appendix B.1). For a community, a tweet is paired with an instruction randomly sampled from the instruction pool. As a result, the community has a maximum of 10K demonstrations $Z_i = \{(x_j, y_j)\}_{j=1}^{m}$ for tweet generation, where $m$ is the size of the community's text corpus $D$.

For each community, we augment the demonstrations of tweet generation with the 52K Alpaca (Taori et al., 2023) demonstrations that cover a wide range of tasks to retain the instruction-following capabilities of the LLM and not restrict it to only generating tweets. Ultimately, there are a maximum of 62K demonstrations in the demonstration corpus for a community.

## 4.2 Instruction Tuning LLMs

For each community $C_i$, we align a Llama-3 model $f_i'$ (AI@Meta, 2024) to the community using its demonstration corpus $Z_i$. The LLM is finetuned on 4 Tesla H100-80GB GPUs with batch size 8 for 3 epochs, which takes about 3 hours.

# 5 Assessing Alignment

To assess how effectively a finetuned LLM $f_i'$ aligns with its target community $C_i$, we measure the model's ability to mimic the responses of community members. We first generate a synthetic corpus $D_i^{ft}$ using $f_i'$ and compare it to the original text corpus $D_i$ from the community. The more closely $D_i^{ft}$ resembles $D_i$, the better aligned the LLM is with the community. We evaluate the similarity between $D_i^{ft}$ and $D_i$ across 1) authenticity, 2) emotional tone, 3) toxicity, and 4) harm. While not exhaustive, these aspects capture the essential features of online social communication. Authenticity ensures that the aligned LLM accurately reflects the meaning, content, and linguistic patterns of the target population's language and generates contextually appropriate responses. Emotional tone captures the affective aspects of communication, helping to convey nuances that may not be evident from semantics alone. Toxicity measures the model's ability to reflect hostility and aggression in the population's discourse. Finally, recognizing that certain online conversations can negatively impact users, we compare the types and levels of harm in language across groups. Although in this paper we focus on the domain of EDs, we argue that our LLM alignment framework is naturally generalizable to online communities in other domains.[3]

## 5.1 Synthetic Corpus Generation

Given a community $C_i$, we create a synthetic corpus $D_i^{ft}$ by prompting an LLM $f_i'$ aligned to the community to generate tweets. To diversify the LLM generations, we compile a set of 27 topics relevant to ED discussions, such as *thinspo*, *fitspo*, and *bonespo* (Appendix C.1), and prompt LLMs to generate tweets on these topics. When generating

[3]We acknowledge that evaluating harm is more tailored to the ED domain, but other evaluation aspects should be widely applicable.

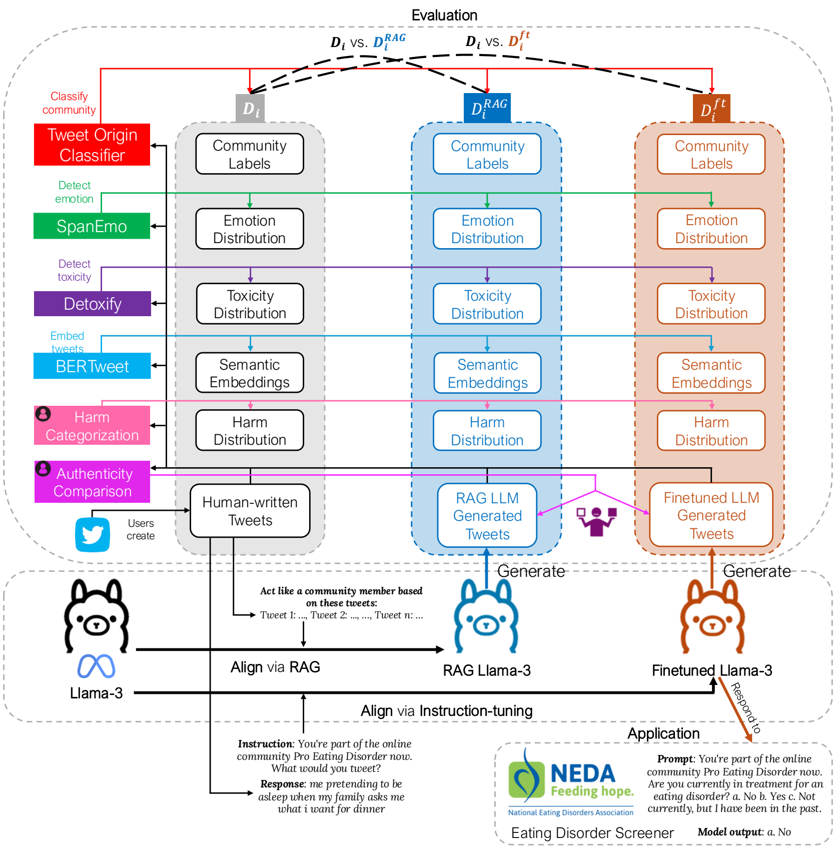


Figure 2: The framework of our method. (1) We align an LLM (Llama-3) to an online community by finetuning the LLM to follow instructions on the task of generating tweets written by users in the community. (2) To prove the effectiveness of alignment, we compare three tweet corpora for each community: human-written tweets $D_i$, RAG LLM-generated tweets $D_i^{RAG}$, and finetuned LLM-generated tweets $D_i^{ft}$. We show that $D_i^{ft}$ is closer to $D_i$ than $D_i^{RAG}$ is, along the following aspects: (a) A classifier trained to classify the tweet origin (what community the tweet belongs to) on $\mathbb{D} = \{D_i\}_{i=1}^n$ performs equally well on $\mathbb{D}^{ft} = \{D_i^{ft}\}_{i=1}^n$, but not on $\mathbb{D}^{RAG} = \{D_i^{RAG}\}_{i=1}^n$; (b) the emotion and toxicity distributions of $D_i^{ft}$ are much closer to that of $D_i$ compared to $D_i^{RAG}$; (c) the semantic embeddings of $D_i^{ft}$ are closer to that of $D_i$ in the embedding space than that of $D_i^{RAG}$ are; (d) a human annotator decides that $D_i^{ft}$ is more aligned to the underlying distribution of $D_i$ than $D_i^{RAG}$ is; (e) two ED experts determine that $D_i^{ft}$ carries harmful narratives that are more similar to $D_i$ than $D_i^{RAG}$. (3) As the LLM is aligned with the community and can speak in the voice of that community, we administer an ED questionnaire to screen the community for EDs.

tweets on a topic, we reuse the diverse instructions from the instruction pool (Table 5 in Appendix B.1). An example instruction is "What would you tweet about **fasting**?" For each topic, the LLM generates 400 tweets, resulting in a synthetic corpus $D_i^{ft}$ with 10,800 tweets for all 27 topics.

We detail the number of tweets in the community's original text corpus $D_i$ that contain the keyword(s) of each of the 27 topics in Table 7 in the Appendix C.1. We observe that $D_i$ contains a very limited number of tweets discussing these topics. This is because we removed the hashtags in tweet processing, and these keywords usually appear in the hashtags. Consequently, when the LLM is finetuned on $D_i$, it is not extensively exposed to tweets that are directly related to these topics. This ensures that **the synthetic corpus $D_i^{ft}$ does not simply replicate** $D_i$. Instead, when the finetuned LLM $f_i'$ generates synthetic tweets on these 27 topics, it extrapolates from existing data in $D_i$ to predict how community members might discuss these previously unseen topics.

**Baseline** We use the LLM with retrieval-augmented generation (RAG) (Lewis et al., 2020) as a baseline. We do not finetune the RAG model. For a community $C_i$, when prompting the model to generate synthetic tweets on topic $t$, we retrieve 250 tweets, consisting of (1) the tweets containing the topic keyword(s), if available, and (2) randomly sampled tweets from $D_i$. Each retrieved tweet is truncated at 20 tokens. We include the retrieved tweets in the prompt, instruct the model to learn the community's mindset from the tweets, and generate synthetic tweets. See Appendix C.2 for the complete prompting template. The synthetic corpus on all topics from the RAG model is denoted as $D_i^{RAG}$.

## 5.2 Alignment Dimensions

### 5.2.1 Automatic Evaluation

**Tweet Origin Classification** We train a classifier to determine the community from which a tweet originated. We achieve this by finetuning Llama-3 using demonstrations with the following template "Instruction: *From these communities: Pro Eating Disorder, Keto & Diet, Body Image, Anti Eating Disorder, Healthy lifestyle & Weight Loss, and Weight Loss Drugs; which community does this Tweet belong to? {Tweet}* Response: *{Community name}*". We randomly sample 3,000 original tweets from each community's corpus $D_i$ and construct a total of 18,000 demonstrations for finetuning. We train the classifier using 95% demonstrations and use the remaining 5% to test, leading to a test accuracy of 0.74. Using this model, we classify the finetuned LLM-generated tweets in $\mathbb{D}^{ft} = \{D_i^{ft}\}_{i=1}^{n}$ and RAG LLM-generated tweets $\mathbb{D}^{RAG} = \{D_i^{RAG}\}_{i=1}^{n}$, leading to an accuracy of 0.75 and 0.59, respectively. These results indicate that the classifier trained on original tweets accurately recognizes the tweets generated by the finetuned LLM. However, it performs poorly on the tweets generated by the RAG LLM, demonstrating that the finetuned LLMs better capture community-specific linguistic characteristics.

**Semantic Comparison** We compute the semantic embeddings of $D_i$, $D_i^{ft}$, and $D_i^{RAG}$ using BERTweet (Nguyen et al., 2020). We then measure the distance between these embeddings using the Fréchet Inception Distance (FID) (Heusel et al., 2017). This metric provides a quantitative measure of the semantic distance between two text corpora. We implement it using the IBM comparing-corpora package (Kour et al., 2022). $FID(D_i, D_i^{ft})$ and $FID(D_i, D_i^{RAG})$ for different communities are shown in Table 1. We see that $FID(D_i, D_i^{ft})$ is much smaller than $FID(D_i, D_i^{RAG})$, implying that the finetuned LLM outputs are more semantically similar responses to the original posts compared to the RAG LLM.

| Community | $FID(D_i, D_i^{RAG})$ | $FID(D_i, D_i^{ft})$ |
|---|---|---|
| Pro-ED | 1.18 | 0.48 |
| Body Image | 1.42 | 0.37 |
| Keto & Diet | 1.05 | 0.51 |
| Anti-ED | 1.00 | 0.52 |
| Healthy Lifestyle & Weight Loss | 1.19 | 0.54 |
| Weight Loss Drugs | 1.04 | 0.40 |

Table 1: Fréchet Inception Distances (FID) (1) between human-written tweets $D_i$ and RAG LLM generated tweets $D_i^{RAG}$, and (2) between human-written tweets $D_i$ and finetuned LLM-generated tweets $D_i^{ft}$. A smaller distance indicates more similarity.

**Emotion and Toxicity Analysis** Emotions and toxicity are vital aspects of online social interactions (Prescott et al., 2019). They can reveal the underlying tone, intent, and style of communication of online users. Within ED communities, these elements heavily impact self-perception of body image (Brytek-Matera and Schiltz, 2011) and can exacerbate body dissatisfaction (Kast, 2018).

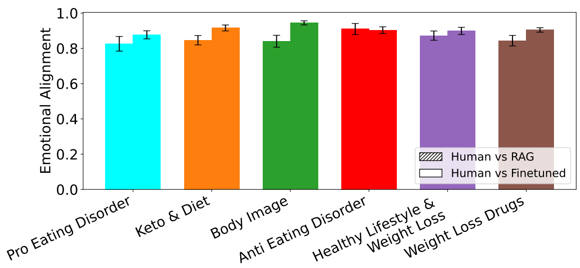


Figure 3: Emotional agreement (a) between human-written tweets and RAG LLM-generated tweets, and (b) between human-written tweets and finetuned LLM-generated tweets. The differences in affective alignment between pairs within each community are statistically significant at a 95% confidence level.

We analyze the emotions of tweets using SpanEmo (Alhuzali and Ananiadou, 2021). For each tweet, SpanEmo returns a vector of confidence scores of eleven emotions: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust. For a community $C_i$, we sum the emotion confidence vectors of all tweets (i.e., $D_i$, $D_i^{ft}$, or $D_i^{RAG}$) and normalize them, resulting in an emotion distribution vector $\mathbf{e}_i$. We then compute the agreement between $\mathbf{e}_i^{ft}$ and $\mathbf{e}_i$, and between $\mathbf{e}_i^{RAG}$ and $\mathbf{e}_i$. The emotional agreement is measured as one minus the Jensen-Shannon distance between the two distribution vectors. As illustrated in Figure 3, for most communities, $D_i^{ft}$ more closely resembles the emotional tone of $D_i$

compared to $D_i^{RAG}$. This demonstrates that finetuning LLMs can effectively capture the authentic emotional tone of posts from communities.

We use Detoxify (Hanu and Unitary team, 2020) to measure toxicity in tweets (Rajadesingan et al., 2020; Sheth et al., 2022). For a tweet, Detoxify returns a value between 0 and 1 indicating the level of toxicity. We only include tweets with toxicity levels equal to or greater than 0.05 for clarity and to reduce noise. Figure 4 shows the distributions of toxicity scores of human-written tweets $D_i$, RAG LLM-generated tweets $D_i^{RAG}$ and finetuned LLM-generated tweets $D_i^{ft}$. We observe that the toxicity distribution of $D_i^{ft}$ matches more closely to that of $D_i$ compared to $D_i^{RAG}$ for most communities, and tweets from the anti-ED community have the highest toxicity.

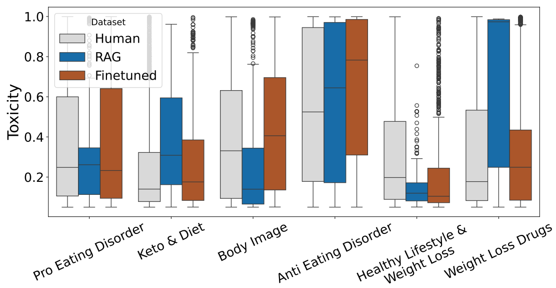


Figure 4: Toxicity distributions across different communities of human-written posts, RAG LLM generated posts, and finetuned LLM generated posts.

### 5.2.2 Human Evaluation

**Authenticity Comparison** An annotator with expertise in EDs on social media was presented with 300 triplets, 50 from each community, where a triplet consists of a community name, a RAG LLM-generated tweet $d_{i,j}^{RAG} \in D_i^{RAG}$, and a finetuned LLM-generated tweet $d_{i,k}^{ft} \in D_i^{ft}$. Both tweets in a triplet are on the same topic and from the same community. For each triplet, the annotator was asked to decide which tweet was more aligned with the given community, by referring to the following characteristics: mis/use of ingroup language, references to themes in underlying distribution (e.g. the Body Image community often references nudity), use of capitalization, and coherence of message. In 225 out of 300 triplets, the annotator chose $d_{i,j}^{ft}$ as a better match, indicating the finetuned LLM is more aligned with the community.

**Harm Categorization** Online ED communities pose significant risks by promoting and normalizing harmful behaviors (Lerman et al., 2023). Harm and toxicity are distinct in online discourse where toxicity detection algorithms may mistakenly flag explicit yet harmless language as toxic (Sánchez et al., 2024). We come up with a dimension tailored to this ED domain where we assess harm by focusing on the underlying semantic content, as opposed to surface-level style. Our goal is for the finetuned LLM, $f_i'$, to accurately capture the level of harm within these communities.

There are no existing classifiers for automatic harm detection in the context of EDs. In collaboration with ED experts, we developed a comprehensive taxonomy of harm specific to ED online content, covering dimensions such as body image, relationships with food and exercise, and self-disclosure. Harm is defined as the promotion or glorification of unhealthy dieting and body objectification.

We sampled 360 tweets across six communities, with 20 each from $D_i$, $D_i^{RAG}$, and $D_i^{ft}$. Two annotators with ED expertise labeled these tweets based on two tasks: (1) determine whether a tweet is harmful, and (2) classify harmful tweets into one of three fine-grained categories—*body image objectification*, *relationship to food and exercise*, or *self-disclosure*. Annotators achieved a Cohen's Kappa score of 0.384 for identifying if harm was present and 0.519 for classifying fine-grained harm categories, indicating fair to moderate agreement.

A tweet was assigned to a harm category if both annotators agreed. Out of 360 tweets, 34 were classified into harm categories across $\mathbb{D}$, $\mathbb{D}^{RAG}$, and $\mathbb{D}^{ft}$. Figure 5 shows the distribution of these categories, demonstrating that finetuned LLMs better replicate the distribution of harm found in the community's conversations.

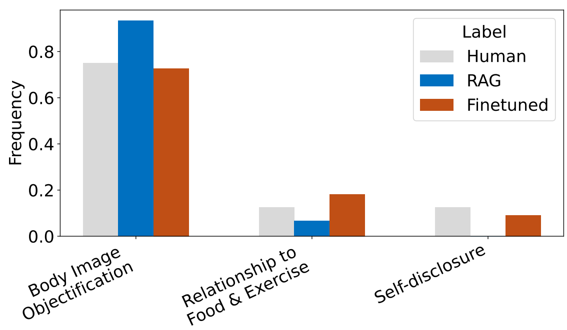


Figure 5: Distribution of the three fine-grained harm categories

## 6 Case Study: Screening Online Communities for Eating Disorders

In §5, we demonstrated that the finetuned LLM learns a more accurate representation of the community than the baseline method. This motivates us to apply psychometric instruments designed to evaluate an individual's risk of EDs to online communities to help uncover unhealthy body and eating concerns within them.

**Eating Disorder Screener** The Stanford-Washington University Eating Disorder Screener (SWED) (Graham et al., 2019) is a concise screening tool for ED behaviors. The screener has been widely used in both men and women (Fitzsimmons-Craft et al., 2018) and incorporated into an online tool (NEDA, 2019) by the National Eating Disorders Association (NEDA, n.d). SWED consists of 11 questions (see Appendix D.1), both multiple-choice and open-ended, covering demographics, height, and weight, ED behaviors, weight and shape concerns, and impairment.

We focus on a subset of SWED questions and evaluate responses using four key criteria (Fitzsimmons-Craft et al., 2018): C1, C2, C3, and C4. These items indicate a higher risk of EDs when the score is elevated (C1) or when being true (C2, C3, C4). For details, see Appendix D.2 and D.3.

**Screening Online Communities via Finetuned LLMs** We prompt finetuned LLMs to respond to questions on the SWED screener. To account for randomness, for each item on the SWED questionnaire, the finetuned LLM generates 50 responses. The responses (Table 8 in Appendix D.5) are aggregated using a majority vote for each question. The results, Table 2, indicate that the *Pro Eating Disorder* community exhibits the highest levels of body image concerns, followed by the *Keto & Diet* community. Furthermore, both communities meet all three criteria signaling a high risk of ED pathology, whereas responses of the *Anti Eating Disorder* community are consistent with a low risk of ED.

These findings align with our empirical observations. The content shared by the *Pro Eating Disorders* community glorifies thinness and includes tips to promote disordered behaviors and body dysmorphia. Conversely, the *Anti Eating Disorder* community is critical of the diet culture and people who glorify EDs. The relatively high-risk score of the *Keto & Diet* community is a concerning indicator that this community may serve as a gateway to EDs. In contrast, the Body Image community, which mostly posts about body positivity, has a low risk of EDs, as does the Healthy Lifestyle & Weight Loss community. Although the latter focuses on weight loss, it appears to achieve this goal through healthy behaviors.

| Community | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| Pro Eating Disorder | **45.0** | T | T | T |
| Keto & Diet | 33.3 | T | T | T |
| Weight Loss Drugs | 16.7 | F | F | T |
| Body Image | 15.0 | F | F | F |
| Healthy Lifestyle & Weight Loss | 13.3 | T | F | F |
| Anti Eating Disorder | 13.3 | F | F | F |

Table 2: Eating disorder risk assessment on the finetuned LLMs for different communities, using four criteria–C1 through C4. For C1, a higher score indicates a higher risk of an ED. For C2, C3, and C4, being positive implies higher risk.

## 7 Conclusion

We demonstrate that aligning LLMs to online communities helps create high-fidelity digital proxies, which can be queried to reveal the implicit mindsets of these communities. When applied to online diet and fitness communities, the method uncovers communities with unhealthy body image and dieting beliefs that put their members at risk of EDs. This is important, as harmful communities that indoctrinate users into extreme ideologies (Schmitz et al., 2022) or glorify EDs and self-harm (Goldenberg et al., 2022) often evade moderation by using coded language that is opaque to outsiders or obfuscate harmful content via coded language and misspellings (Chancellor et al., 2016d; Cobb, 2017; Bickham et al., 2024). Our method can help online platforms bridge these obstacles to create safe and supportive online communities.

## Limitations

**Dataset Bias.** The anonymized version of our dataset may contain implicit biases reflecting societal prejudices or unequal representation of demographic sub-groups. More specifically, ED symptoms have a history of being under-diagnosed in African American and Hispanic adolescents, in part due to stereotypical representation of ED being Caucasian adolescent girls (Gordon et al., 2002). This historical bias could be inadvertently learned by our model, resulting in discriminatory behavior. In our future work, we hope to evaluate the model's fairness across different user groups, allowing us to properly mitigate dataset biases.

**Evolving Nature of Online Communities.** Capturing the evolving nature of online communities is potentially difficult. Online discourse is dynamic, with language, topics, and sentiments shifting over time. Our finetuning process may not fully account for these temporal changes, which could result in misalignment when the model is applied to current discussions within the community.

**Synthetic Corpus Artifacts.** The synthetic corpus generated by the LLM might also introduce artifacts that do not fully represent the authentic discourse of the community. Although we strive for diversity in the generated content, the model's predictions on previously unseen topics may not always accurately reflect how community members would engage with those topics in real-life scenarios.

**Evaluation Metrics.** While the aspects of authenticity, emotional tone, toxicity, and harm capture important aspects of online communication, they may not encompass all the subtle and complex features of human discourse. As a result, some aspects of community interaction may be underrepresented or overlooked in our evaluation process.

**Complete Coverage of Eating Disorders.** This paper looks at the discussions of ED in online communities. We focus on a conglomeration of ED, including bulimia nervosa, anorexia nervosa, and binge eating disorder. Besides ED, our dataset captures other discussions related to weight concerns, such as weight loss, diet, body positivity, etc. Unfortunately, our data does not comprehensively represent all existing ED. However, our methods ensure that if a large ED community has some overlap with our keyword list, the community will be identified.

## Ethics Statement

**Risk of Finetuning Models Towards Harm** In our study, we expect the finetuned LLMs to replicate harmful narratives from online communities. This is done solely to demonstrate that, through our alignment framework, LLMs can accurately capture the nuanced language of these communities, including harmful content. Our objective is not to create models with malicious intent. We strongly advise that any future replication of this work be conducted with extreme caution.

**Community-Level Diagnosis.** Diagnosing psychiatric illness at the community level comes with the risk of falsely diagnosing some community members. This could lead to unjust actions against users, such as unwarranted bans or removal of content. Additionally, approximating community behavior inherently excludes minority group members. Simultaneously, anorexia is one of the deadliest mental health disorders[4] and participation in online pro-ED spaces heightens one's disease risk (Mento et al., 2021). By evaluating psychiatric illness on the community level, we can identify toxic communities, helping content moderation experts deploy proper interventions to promote healthy and safe online environments. We encourage the use of human moderators to review and validate the decisions made by our model, particularly in cases with low confidence scores.

**Topic Sensitivity and Privacy** The sensitive nature of our topic means that our outputs could be misused, such as targeted advertising. Additionally, our dataset includes some tweets that disclose deeply personal information such as medical diagnoses, weight information, and personal struggles. Many of these tweets are posted under the assumption of anonymous identity. By collecting these tweets, user-specific information may be pieced together thus de-anonymizing some users. For these reasons, we take precautions to anonymize the social media posts before feeding them to the language models. Additionally, researchers can be granted access to generated tweets upon detailed inquiry.

**Hallucination Risk.** Our finetuned models can exhibit hallucinations, generating incorrect or nonsensical information. Hallucination in the context of community alignment can lead to community misrepresentation. In future work, we hope to utilize some factual-based evaluation datasets to measure model hallucination.

## References

AI@Meta. 2024. Llama 3 model card.

Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification

[4]https://www.state.sc.us/dmh/anorexia/statistics.htm

as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584.

Charles Bickham, Kia Kazemi-Nia, Luca Luceri, Kristina Lerman, and Emilio Ferrara. 2024. Hidden in plain sight: Exploring the intersections of mental health, eating disorders, and content moderation on tiktok. *Preprint*, arXiv:2404.15457.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Anna Brytek-Matera and Lony Schiltz. 2011. Association between attitudes towards body image, negative emotions about one's own body and self-state representations in a clinical sample of eating disordered women. *Archives of Psychiatry and Psychotherapy*, 2:37–43.

Stevie Chancellor, Zhiyuan Lin, and Munmun De Choudhury. 2016a. "this post will just get taken down"' characterizing removed pro-eating disorder social media content. In *CHI*, pages 1157–1162.

Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016b. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *CSCW*, pages 1171–1184.

Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. 2016c. Recovery amid pro-anorexia: Analysis of recovery in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2111–2123.

Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016d. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 1201–1213.

Gemma Cobb. 2017. "this is not pro-ana": Denial and disguise in pro-anorexia online spaces. *Fat Studies*, 6(2):189–205.

Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models increases exploration and bias. *arXiv preprint arXiv:2304.11111*.

Esin Durmus, Karina Nguyen, 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*.

Ellen Fitzsimmons-Craft, Marie-Laure Firebaugh, Andrea Graham, Dawn Eichen, Grace Monterubio, Katherine Balantekin, Anna Karam, Annie Seal, Burkhardt Funk, C. Taylor, and Denise Wilfley. 2018. State-wide university implementation of an online platform for eating disorders screening and intervention. *Psychological Services*, 16.

Debbie Ging and Sarah Garvey. 2018. 'written in these scars are the stories i can't explain': A content analysis of pro-ana and thinspiration image sharing on instagram. *New Media & Society*, 20:1181–1200.

Alex Goldenberg, John Farmer, Lee Jussim, Loree Sutton, Danit Finkelstein, Cristian Ramos, Pamela Paresky, and Joel Finkelstein. 2022. Online communities of adolescents and young adults celebrating, glorifying, and encouraging self-harm and suicide are growing rapidly on twitter. Technical report, NCRI.

Kathryn H Gordon, Marisol Perez, and Thomas E Joiner Jr. 2002. The impact of racial stereotypes on eating disorder recognition. *International Journal of Eating Disorders*, 32(2):219–224.

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

AK Graham, M Trockel, H Weisman, EE Fitzsimmons-Craft, KN Balantekin, DE Wilfley, and CB Taylor. 2019. A screening tool for detecting eating disorder risk and diagnostic symptoms among college-age women. *Journal of American College Health*, 67(4):357–366. Epub 2018 Oct 9.

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

Zihao He, Rebecca Dorn, Siyi Guo, Minh Duc Chu, and Kristina Lerman. 2024a. Community-cross-instruct: Unsupervised instruction generation for aligning large language models to online communities. *arXiv preprint arXiv:2406.12074*.

Zihao He, Siyi Guo, Ashwin Rao, and Kristina Lerman. 2024b. Whose emotions and moral sentiments do language models reflect? *arXiv preprint arXiv:2402.11114*.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Guangyuan Jiang, Manjie Xu amd Song-Chun Zhu, Wenjuan Han, and Yixin Zhu Chi Zhang. 2022a. Evaluating and inducing personality in pre-trained language models. *arXiv preprint arXiv:2206.07550*.

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

Adrienne S Juarascio, Amber Shoaib, and C Alix Timko. 2010. Pro-eating disorder communities on social networking sites: a content analysis. *Eating disorders*, 18(5):393–407.

Hinde Kast. 2018. The unspoken power of toxic words on body image. Master's thesis, University of Southern California.

JD Killen, C Hayward, DM Wilson, CB Taylor, LD Hammer, I Litt, B Simmonds, and F Haydel. 1994. Factors associated with eating disorder symptoms in a community sample of 6th and 7th grade girls. *International Journal of Eating Disorders*, 15(4):357–367.

Joel Killen, C. Taylor, Chris Hayward, Katherine Haydel, Lawrence Hammer, Helena Kraemer, Anne Blair-Greiner, and Diane Strachowski. 1996. Weight concerns influence the development of eating disorders: A 4-year prospective study. *Journal of consulting and clinical psychology*, 64:936–40.

Joel D. Killen, C. Barr Taylor, Lawrence D. Hammer, Iris Litt, Darrell M. Wilson, Tia Rich, Chris Hayward, Beverly Simmonds, Helena Kraemer, and Ann Varady. 1993. An attempt to modify unhealthful eating attitudes and weight regulation practices of young adolescent girls. *International Journal of Eating Disorders*, 13(4):369–384.

George Kour, Samuel Ackerman, Eitan Farchi, Orna Raz, Boaz Carmeli, and Ateret Anaby-Tavor. 2022. Measuring the measuring tools: An automatic evaluation of semantic metrics for text corpora. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2022)*. Association for Computational Linguistics.

Kristina Lerman, Aryan Karnati, Shuchan Zhou, Siyi Chen, Sudesh Kumar, Zihao He, Joanna Yau, and Abigail Horn. 2023. Radicalized by thinness: Using a model of radicalization to understand pro-anorexia communities on twitter. *arXiv preprint arXiv:2305.11316*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yang Lu, Jordan Yu, and Shou-Hsuan Stephen Huang. 2023. Illuminating the black box: A psychometric investigation into the multifaceted nature of large language models. *Preprint*, arXiv:2312.14202.

Abby McCormack. 2010. Individuals with eating disorders and the use of online support groups as a form of social support. *CIN: Computers, Informatics, Nursing*, 28:12–19.

Carmela Mento, Maria Catena Silvestri, Maria Rosaria Anna Muscatello, Amelia Rizzo, Laura Celebre, Martina Praticò, Rocco Antonio Zoccali, and Antonio Bruno. 2021. Psychological impact of pro-anorexia and pro-eating disorder websites on adolescent females: A systematic review. *International Journal of Environmental Research and Public Health*, 18:2186.

NEDA. 2019. Eating disorder online screening tool. https://www.nationaleatingdisorders.org/screening-tool/. [Online; accessed 14-May-2024].

NEDA. n.d. Eating disorder statistics & research.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Atte Oksanen, David Garcia, and Pekka Räsänen. 2016. Proanorexia communities on social media. *Pediatrics*, 137(1).

Jessica A Pater, Oliver L Haimson, Nazanin Andalibi, and Elizabeth D Mynatt. 2016. "hunger hurts but starving works" characterizing the presentation of eating disorders online. In *CSCW*, pages 1185–1200.

Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 0(0):17456916231214460. PMID: 38165766.

Julie Prescott, Terry Hanley, and Katalin Ujhelyi Gomez. 2019. Why do young people use online forums for mental health and emotional support? benefits and challenges. *British Journal of Guidance & Counselling*, 47(3):317–327.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2020. Language models are unsupervised multitask learners.

Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 557–568.

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.

Matheus Schmitz, Goran Muric, and Keith Burghardt. 2022. Quantifying how hateful communities radicalize online users. In *ASONAM*, pages 139–146. IEEE.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

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

Cinthia Sánchez, Minh Duc Chu, Zihao He, Rebecca Dorn, Stuart Murray, and Kristina Lerman. 2024. Feelings about bodies: Emotions on diet and fitness forums reveal gendered stereotypes and body image concerns. *Preprint*, arXiv:2407.03551.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Probing the moral development of large language models through defining issues test. *Preprint*, arXiv:2309.13356.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

C Taylor, Susan Bryson, Kristine Luce, Darby Cunning, Angela Doyle, Liana Abascal, Roxanne Rockwell, Parvati Dev, Andrew Winzelberg, and Denise Wilfley. 2006. Prevention of eating disorders in at-risk college-age women. *Archives of general psychiatry*, 63:881–8.

Daphna Yeshua-Katz and Nicole Martins. 2013. Communicating stigma: The pro-ana paradox. *Health Communication*, 28(5):499–508.

# A Online Communities in ED Discussions

## A.1 Search Terms

The terms used for tweet collection are: *anatips, bodygoals, bodyimage, bodypositivity, chloetingchallange, cleaneating, cleanvegan, eatingdisorder, edrecovery, edtwt, edvent, fatspo, fearfood, foodistheenemy, healthyliving, intermittentfasting, iwillbeskinny, juicecleanse, ketodiet, losingweight, lowcalrestriction, meanspo, midriff, ozempic, proana, proanatips, redbracetpro, semaglutide, skinnycheck, slimmingworld, sweetspo, thighgapworkout, thinspo, thinspoa, watercleanse, wegovy, weightlossjourney, weightlossmotivation, whatieatinaday, bonespo, fatacceptance, keto, promia, skinnydiet, dietculture, m34nspo, weightloss, weightlosstips*.

Below are the explanations of these keywords used in the context of the online ED community:

- chloetingchallange: a popular fitness trend created by YouTuber Chloe Ting.
- edtwt: refers to the general ED community on Twitter/X.
- fatspo: promotes body positivity and acceptance of larger body sizes.
- fearfood: a term for foods that cause anxiety or avoidance in those with ED.
- redbracetpro: refers to the bracelet patients wear at a treatment facility when they are medically unstable or fragile.
- meanspo, m34nspo: be deliberately mean or insulting to motivate someone to do something.
- midriff: refers to the area of the body between the chest and the waist. It often shows one's ribcage and is closely associated with being skinny.
- ozempic, wegovy, semaglitude: refers to a medication primarily used to treat type 2 diabetes but has gained attention for its use as a weightloss drug
- thighgapworkout: refers to exercises aimed at achieving a gap between the thighs, a controversial and unrealistic body goal often associated with unhealthy body image standards.
- thinspo: short for "thinspiration," referring to content or imagery that promotes extreme thinness.
- bonespo: refers to content that glorifies extreme thinness by focusing on images of prominent bones.
- promia: the promotion of bulimia-related behaviors, often found in harmful online communities.

## A.2 Profiling Communities

The statistics of the top 20 largest user clusters detected by Louvain modularity maximization are shown in Table 3. The word clouds of tweets in these 20 clusters are shown in Figure 6. The retweet network, with users from different clusters showing different colors, is shown in Figure 7.

To profile discussions, we provide a random sample of 200 posts from each user cluster to GPT-4 with the prompt: "*Given this list of posts, summarize the main ideas in 1 sentence*". We observe that using different random samples of posts leads to substantially similar summaries. After reviewing generated summaries, we note significant thematic and content overlaps and group the clusters based on their common topics of discussion into clusters: *Pro-ED*, *Keto & Diet*, *Body Image*, *Anti-ED*, *Healthy Lifestyle & Weight Loss*, *Weight Loss Drugs*, and *spam* (not included).

Members of clusters 0, 7, 8, and 9 use "edtwt" to self-identify as part of the ED community, and their posts promote disordered behaviors. Interestingly, members of clusters 8 and 9 post in Spanish and Portuguese, respectively. They are also placed close to pro-ED clusters 0, 7 in Figure 7. Cluster 2, although also uses "edtwt" label, is well separated from the rest. This cluster takes a critical—anti-ED—stance on ED, as seen from the summary in Table 4.

The remaining clusters are loosely connected in the retweet network and less insular than the pro-ED cluster. Clusters 1, 15, 16, 18 discuss the risks and benefits of the keto diet; clusters 3, 6 and 19 focus on issues surrounding the use of weight loss drugs like Ozempic and Wegovy; Clusters 4, 13, 17 examine issues of healthy lifestyle and weight loss, while clusters 5, 10 cover body image topics, like body positivity and self-acceptance. Clusters 11, 12, and 14 are on other random issues not relevant to ED, as can be observed from word clouds and thus we exclude them in our subsequent analysis.

# B Aligning LLMs

## B.1 Demonstration Template for LLM finetuning

The instructions for finetuning LLMs are shown in Table 5. For tweet generation demonstrations, each tweet is paired with a randomly sampled instruction from the table. An example prompt template is shown below. More demonstrations for different communities are shown in Table 6.

```
Instruction: What would you tweet?
Response: {Tweet}
```

# C Assessing Alignment

## C.1 Topics for Creating Synthetic Tweets

The 27 topics used for creating the synthetic tweets are: *thinspo, fitspo, bonespo, deathspo, caloric restriction, meanspo, ozempic, wegovy, fatspo, fatphobia, thighgap, caloric counting, purging, food rules, extreme diet, food fear, hiding food, fasting, starving, steroid, excessive exercising, body dysmorphia, working out, anorexia, bulimia, orthorexia, binge eating*.

The number of tweets mentioning the topics for each community is shown in Table 7.

## C.2 Prompt Template for Tweet Generation by RAG LLM

An example prompt template is shown below.

```
You're part of an online community now. To
help you describe this online community,
here are the tweets made by members in
this community about the topic of {topic}.
Tweet 1: {tweet_1}
Tweet 2: {tweet_1}
  ...
Tweet 250: {tweet_250}
What would you tweet about {topic}? Learn
the ideas and mindset of the community
from these tweets and speak like a member
from this community.  Only generate one
tweet.
```

## C.3 Demonstration Template for Tweet Origin Classification

```
Instruction: From these communities: Pro
Eating Disorder, Keto & Diet, Body Image,
Anti Eating Disorder, Healthy lifestyle &
Weight Loss, and Weight Loss Drugs, which
community does this Tweet belong to?
{Tweet}
Response: {community_name}
```

# D Screening Online Communities

## D.1 Stanford-Washington University Eating Disorder (SWED) 3.0 Screener

The 11 questions in the questionnaire are shown below.

1. Are you currently in treatment for an eating disorder?

| Comm | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # of users | 61,954 | 24,400 | 21,887 | 20,631 | 9,901 | 9,031 | 9,000 | 8,084 | 7,702 | 7,020 | |
| # of tweets | 805,249 | 112,674 | 32,883 | 37,788 | 193,348 | 24,395 | 21,369 | 82,702 | 70,764 | 71,970 | |
| **Comm** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **total** |
| # of users | 6,477 | 6,158 | 5,181 | 4,528 | 3,682 | 3,672 | 3,360 | 3,163 | 3,086 | 2,865 | 221,887 |
| # of tweets | 15,796 | 9,254 | 7,019 | 103,177 | 260,971 | 5,338 | 4,881 | 5,065 | 4,612 | 7,021 | 1,876,276 |

Table 3: Number of users (community size) and tweets in the top 20 largest user clusters respectively and in total.

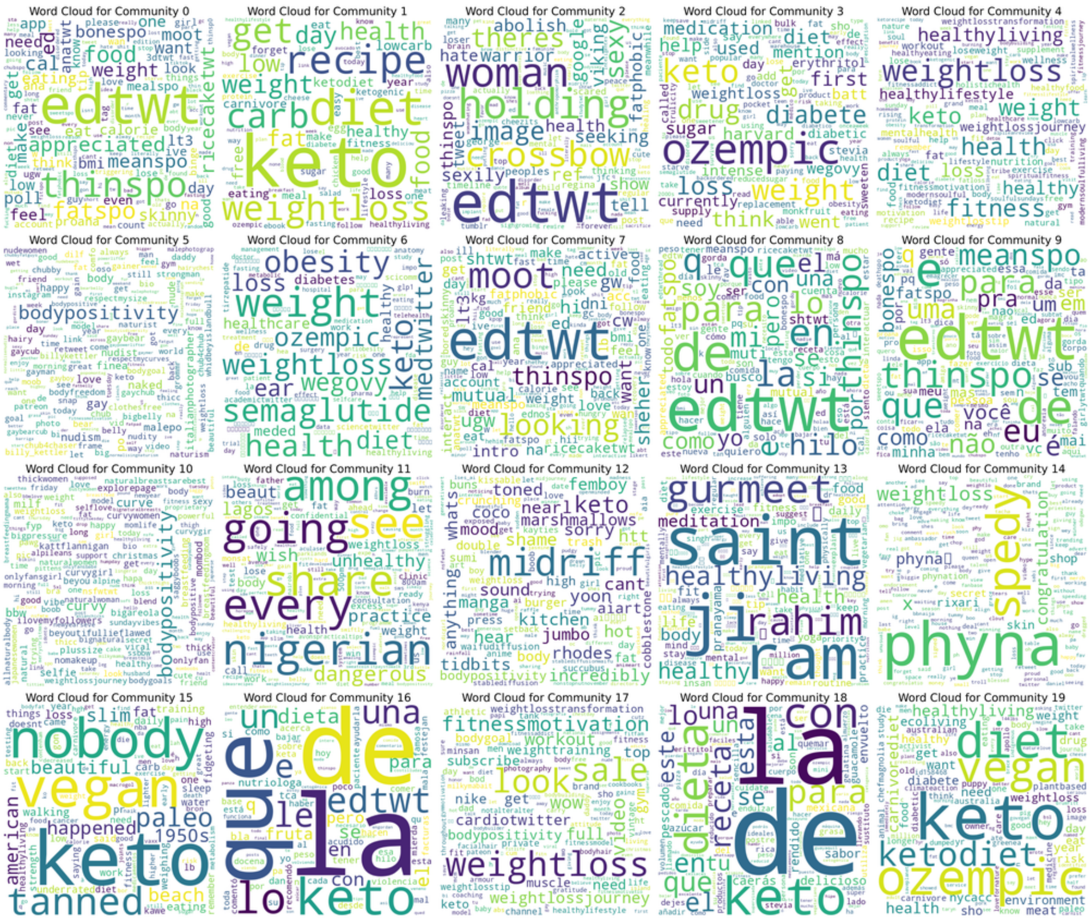


Figure 6: Wordclouds of popular terms appearing in the original tweets posted within each user cluster.

(a) No
(b) Yes
(c) Not currently, but I have been in the past

2. What was your lowest weight in the past year, including today, in pounds?

3. What is your current weight in pounds?

4. What is your current height in inches?

5. How much more or less do you feel you worry about your weight and body shape than other people your age?

(a) I worry a lot less than other people
(b) I worry a little less than other people
(c) I worry about the same as other people
(d) I worry a little more than other people
(e) I worry a lot more than other people

6. How afraid are you of gaining 3 pounds?

(a) Not afraid of gaining
(b) Slightly afraid of gaining
(c) Moderately afraid of gaining
(d) Very afraid of gaining
(e) Terrified of gaining

7. When was the last time you went on a diet?

(a) I have never been on a diet

| Community Tag | Summary of Community Discussions | User Cluster ID |
|---|---|---|
| Pro Eating Disorder | This community revolves around the online eating disorder community (edtwt), sharing tips, thinspo (thin inspiration), meanspo (mean inspiration), fasting strategies, and discussing body image and weight loss goals, often in a way that promotes disordered eating behaviors. | 0,7,8,9 |
| Keto & Diet | This community focuses on a range of topics related to ketogenic diets, weight loss, metabolic health, and low-carb recipes, with discussions on the effectiveness of keto for various health conditions, debates on prescribing obesity drugs to children, and personal testimonials about the benefits of a keto. | 1,15,16,18 |
| Body Image | This community dives into a variety of personal updates, including fitness activities, body positivity, nudism, modeling, and social interactions, with some tweets promoting content or expressing motivational thoughts. | 5, 10 |
| Anti Eating Disorder | This community expresses strong negative sentiments towards "edtwt" (presumably "eating disorder Twitter"), criticizing it for being toxic, fatphobic, and harmful, with calls to abolish it and stop interacting with its content. | 2 |
| Healthy Lifestyle & Weight Loss | This community covers a variety of health and wellness topics, including weight loss methods, dietary plans, fitness advice, healthy eating, keto diet, fasting, moxibustion, and motivational messages for maintaining a healthy lifestyle. | 4,13,17 |
| Weight Loss Drugs | This community discusses the controversial use of the diabetes drug Ozempic for weight loss, the impact of its shortage on diabetic patients, the cost of the medication, and related topics such as body positivity, keto diets, and the role of influencers and celebrities in promoting certain health trends and products. | 3,6,19 |

Table 4: Summary of posts in the communities with GPT-4. Similar communities are merged.

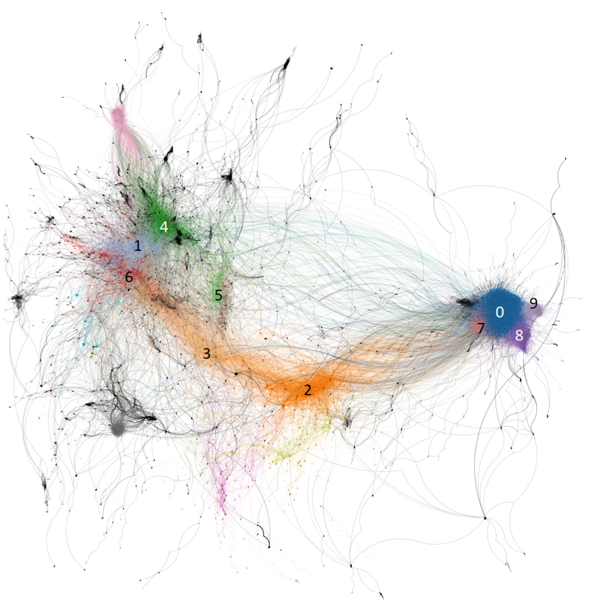


Figure 7: Retweet network, where nodes are individual users and edges indicate the retweeting activities. Node colors represent different user clusters identified by the Louvain modularity method.

(b) I was on a diet about one year ago
(c) I was on a diet about 6 months ago
(d) I was on a diet about 3 months ago
(e) I was on a diet about 1 month ago
(f) I was on a diet less than 1 month ago
(g) I'm on a diet now

8. Compared to other things in your life, how important is your weight to you?

    (a) My weight is not important compared to other things in my life
    (b) My weight is a little more important than some other things
    (c) My weight is more important than most, but not all, things in my life
    (d) My weight is the most important thing in my life

9. Do you ever feel fat?

    (a) Never
    (b) Rarely
    (c) Sometimes
    (d) Often
    (e) Always

10. In the past 3 months, how many times have you had a sense of loss of control AND you also ate what most people would regard as an unusually large amount of food at one time, defined as definitely more than most people would eat under similar circumstances?

11. In the past 3 months, how many times have you done any of the following as a means to control your weight and shape:

    (a) Made yourself throw up?
    (b) Used diuretics or laxatives?
    (c) Exercised excessively? i.e. pushed yourself very hard; had to stick to a specific exercise schedule no matter what

| Index | Instruction |
| --- | --- |
| **1** | What would you tweet? |
| **2** | What tweet would you send out? |
| **3** | What's your tweet today? |
| **4** | What would you want to tweet about? |
| **5** | What's on your mind to tweet? |
| **6** | What tweet would you drop? |
| **7** | What would you say? |
| **8** | What's your tweet? |
| **9** | Tweet something. |
| **10** | What would you tweet? |
| **11** | What kind of tweet would you send out to engage with fellow members? |
| **12** | Draft a tweet that captures the interests and spirit of the community. |
| **13** | Craft a relatable tweet that resonates with members. |
| **14** | Share a tweet that sparks conversation on relevant topics. |
| **15** | Compose a tweet that reflects the shared voice and passions. |
| **16** | Author an insightful tweet that inspires dialogue among members. |
| **17** | Tweet something that provokes intellectual discourse. |
| **18** | Tweet an observation or perspective that contributes meaningfully. |
| **19** | Craft a tweet that elevates the ongoing conversations. |
| **20** | Compose a tweet that encourages enriching engagement. |

Table 5: Instructions used to finetune the LLMs.

– for example even when you were sick-/injured or if it meant missing a class or other important obligation; felt compelled to exercise

(d) Fasted? i.e. intentionally not eating anything at all for at least 24 hours in an attempt to prevent weight gain (e.g., that is feared as a result of binge eating) or to lose weight

12. Have you experienced significant weight loss (or are at a low weight for your age and height) but are not overly concerned with the size and shape of your body?

    (a) Yes

    (b) No

### D.2 Weight Concerns Scale

The Weight Concerns Scale (WCS) (Killen et al., 1993) is a brief, validated psychometric instrument designed to measure concerns about weight and body shape, fear of gaining weight, dieting history, and feelings of fatness. The scale was created to assess risk factors for ED and has demonstrated excellent stability and sensitivity to treatment differences. Longitudinal studies have shown that high WCS scores predict the onset of ED, making it a valuable instrument for identifying at-risk individuals (Killen et al., 1994, 1996; Taylor et al., 2006). Our work deploys this questionnaire to gain insights into body-related concerns of members of online communities.

### D.3 Diagnostic Criteria

C1 is a numerical score, the Weight Concerns Scale (WCS) (Killen et al., 1993) - a metric that measures concerns about weight and shape. This criterion composes answers from questions 5 to 9 in SWED, maps the options to numerical values to a 0-100 scale, and calculates the average score across the 5 questions. The higher the WCS score is, the higher the risk factor for body and shape concerns.

C2, C3, and C4 are boolean values, and being true implies a higher risk of EDs. C2 is True if the model responds “more/most important" (options c or d) to Q8 “Compared to other things in your life, how important is your weight to you?". C3 is True if the model responds “moderately afraid" or more (options c, d, or e) to Q6: “How afraid are you of gaining 3 pounds?”. C4 is True if the model responds “Yes” at least 3 times to Q11: “In the past 3 months, how many times have you done any of the following as a means to control your weight and shape: a. Made yourself throw up? b. Used

| Community | Instruction | Response |
|---|---|---|
| Pro-ED | What would you tweet? | most of the time the only thing i want in the whole world is to be skinny and lose weight |
| Keto & Diet | What tweet would you send out? | ready to jumpstart your weight loss journey? Try these tips to help you lose weight in a month |
| Body Image | What's your tweet today? | everyone has something about their body they 're not completely happy with. Don't focus on that! Love the body you have! |
| Anti-ED | What would you want to tweet about? | do not follow me if you're on edtwt. I don't know how many times I have to say this |
| Healthy Lifestyle | What's on your mind to tweet? | we don't stop exercising because we grow old, we grow old because we stop exercising. |
| Weight Loss Drugs | What tweet would you drop? | are our keto diet pills effective and safe to use? The truth about keto diet pills benefits, risks, and effectiveness |

Table 6: Demonstration examples for LLM finetuning for different communities.

diuretics or laxatives? c. Exercised excessively? d. Fasted?"

### D.4 Prompt Template for SWED Question Answering

```
{Question}
Respond to the following question only
with the letter at the beginning of each
option or with a number.
```

### D.5 LLM Responses to SWED Questions

Responses from the finetuned LLMs to SWED questions are shown in Table 8.

| Topic | Pro-ED | Keto and Diet | Body Image | Anti-ED | Healthy lifestyle and Weight Loss | Weight Loss Drugs |
|---|---|---|---|---|---|---|
| thinspo | 20 | 0 | 0 | 24 | 0 | 2 |
| fitspo | 0 | 0 | 0 | 0 | 0 | 0 |
| bonespo | 4 | 0 | 0 | 0 | 0 | 0 |
| deathspo | 0 | 0 | 0 | 0 | 0 | 0 |
| caloric restriction | 0 | 0 | 0 | 0 | 0 | 0 |
| calorie counting | 0 | 0 | 0 | 1 | 0 | 0 |
| purging | 0 | 0 | 0 | 0 | 0 | 0 |
| food rules | 0 | 0 | 0 | 0 | 0 | 0 |
| extreme diet | 0 | 0 | 0 | 0 | 0 | 0 |
| food fear | 0 | 0 | 0 | 0 | 0 | 0 |
| hiding food | 0 | 0 | 0 | 0 | 0 | 0 |
| fasting | 0 | 1 | 0 | 1 | 0 | 2 |
| starving | 1 | 0 | 1 | 1 | 0 | 1 |
| steroid | 0 | 0 | 0 | 0 | 0 | 0 |
| meanspo | 0 | 0 | 0 | 0 | 0 | 0 |
| ozempic | 0 | 0 | 0 | 0 | 0 | 0 |
| wegovy | 0 | 0 | 0 | 0 | 0 | 0 |
| fatspo | 2 | 0 | 0 | 3 | 0 | 0 |
| fatphobia | 0 | 0 | 0 | 0 | 0 | 0 |
| thigh gap | 4 | 0 | 0 | 0 | 0 | 0 |
| excessive exercising | 0 | 0 | 0 | 0 | 0 | 0 |
| body dysmorphia | 0 | 0 | 1 | 1 | 0 | 0 |
| working out | 1 | 2 | 2 | 0 | 0 | 1 |
| anorexia | 0 | 0 | 0 | 2 | 0 | 0 |
| bulimia | 0 | 0 | 0 | 0 | 0 | 0 |
| orthorexia | 0 | 0 | 0 | 0 | 0 | 0 |
| binge eating | 1 | 0 | 0 | 0 | 0 | 0 |

Table 7: Number of tweets mentioning topic keyword(s) from each community.

| Question | Pro ED | Keto & Diet | Body Image | Anti ED | Healthy Lifestyle & Weight Loss | Weight Loss Drugs |
|---|---|---|---|---|---|---|
| 5 | b | c | b | a | a | b |
| 6 | c | c | a | c | a | b |
| 7 | c | a | b | b | a | a |
| 8 | c | c | b | a | c | b |
| 9 | c | a | a | a | a | a |
| 11a | c | a | a | c | a | a |
| 11b | c | a | a | c | a | a |
| 11c | a | b | b | b | b | a |
| 11d | a | b | b | b | b | a |

Table 8: Responses from the finetuned LLMs to the questions in SWED that are used to compute the diagnosis criteria. The responses displayed are the majority of answers for each question.