Evaluating the Effectiveness of Large Language Models in Generating Prosocial Interventions in Reddit Communities

Ipshita Joshi
imjoshi2@illinois.edu
University of Illinois Urbana-Champaign
USA

Rishabh Vallecha
rv14@illinois.edu
University of Illinois Urbana-Champaign
USA

Sanket Nagesh Babu Donty
sdonty2@illinois.edu
University of Illinois Urbana-Champaign
USA

Utkarsh Sharma
usharma4@illinois.edu
University of Illinois Urbana-Champaign
USA

ABSTRACT

Online communities provide platforms for individuals to connect and share experiences, but the quality of interactions can vary greatly. Prosocial behavior is crucial for fostering positive online interactions, and Large Language Models (LLMs) show promise in generating prosocial interventions. This study explores the potential of LLMs in mimicking user style, content, and prosocial behavior within online communities, focusing on Reddit. Through analysis of user-generated content and LLM-generated responses, we demonstrate that LLMs can effectively capture and reproduce the linguistic characteristics and sentiments of users, as supported by high cosine similarity scores for Word2Vec and Doc2Vec embeddings and user archetype clustering results. The findings highlight the potential of LLMs to guide discussions positively, maintain a healthy atmosphere, and support vulnerable individuals by generating targeted prosocial interventions. LLM-based interventions can alleviate the burden on human moderators and foster a supportive and inclusive environment. However, ethical considerations, such as ensuring alignment with community values and maintaining transparency and user consent, must be prioritized. This study underscores the potential of LLMs in promoting prosocial behavior and fostering positive interactions within online communities, with insights applicable to various online platforms and offline contexts. Collaborations between researchers, platform administrators, and community members are essential to develop robust, ethical, and effective strategies for leveraging LLMs to build online communities that promote well-being and encourage constructive dialogue.

KEYWORDS

Large Language Models, Prosocial Behavior, Online Communities, Reddit, User Mimicry, Content Generation, Linguistic Analysis, Sentiment Analysis, Auto-Moderation, Community Guidelines

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
May 2024, University of Illinois Urbana-Champaign, IL, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Ipshita Joshi, Rishabh Vallecha, Sanket Nagesh Babu Donty, and Utkarsh Sharma. 2024. Evaluating the Effectiveness of Large Language Models in Generating Prosocial Interventions in Reddit Communities. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND MOTIVATION

In recent years, online communities have become integral platforms for individuals to connect, share experiences, and seek support. These communities, such as those found on Reddit, serve as virtual spaces where people from diverse backgrounds can engage in discussions, exchange ideas, and form meaningful connections. However, the nature and quality of interactions within these communities can vary greatly, ranging from supportive and empathetic to toxic and harmful. Eisenberg et al. [7] define prosocial behavior as actions that benefit others or society as a whole, and it plays a crucial role in fostering positive online communities. Prosocial interactions, such as offering emotional support, providing helpful advice, and promoting inclusivity, contribute to a sense of belonging and well-being among community members, as noted by Batson et al. [3].

Kang et al. [8] have demonstrated that encouraging prosocial behavior in online communities can lead to numerous benefits, such as increased user engagement, satisfaction, and loyalty. When users experience a supportive and positive environment, they are more likely to actively participate, contribute valuable content, and form long-lasting connections within the community. Moreover, Cheng et al. [5] highlight that prosocial interactions serve as a buffer against negative behaviors, such as trolling, harassment, and the spread of misinformation. By promoting a culture of kindness, empathy, and respect, online communities can mitigate the impact of malicious actors and maintain a healthy and constructive atmosphere.

One promising approach to fostering prosocial behavior in online communities is the use of Large Language Models (LLMs) to generate prosocial interventions. Brown et al. [4] showcase the remarkable capabilities of LLMs, such as OpenAI's GPT-3, in understanding and generating human-like text based on vast amounts of training data. By leveraging the power of LLMs, it becomes possible to automatically generate prosocial responses to posts and comments, guiding discussions in a positive direction. This is particularly valuable in large communities with a high volume of user interactions, where manual moderation becomes challenging.

The potential applications of LLMs for generating prosocial interventions on platforms like Reddit are vast. One key use case is the auto-moderation of large communities. LLMs can assist moderators by automatically generating prosocial responses to posts and comments, helping to maintain a positive and constructive discourse. This can alleviate the burden on human moderators and ensure a more consistent and timely response to potentially problematic content. Additionally, LLMs can be trained to identify posts or comments that indicate distress, loneliness, or mental health concerns. By generating targeted prosocial interventions, such as offering support, providing resources, or directing users to appropriate help, LLMs can play a crucial role in supporting vulnerable individuals within the community.

Furthermore, LLMs can be utilized to promote community guidelines and values. By generating interventions that align with the community's standards and expectations, LLMs can gently nudge users towards respectful and constructive behavior. This can help reinforce positive norms and maintain a welcoming and inclusive environment for all members.

To fully realize the potential of LLMs in generating prosocial interventions, it is essential to investigate their effectiveness and understand how well they can mimic the style and content of users in online communities. This leads us to our first research question **(RQ1): How effectively can Large Language Models generate prosocial interventions that mimic the style and content of users in online communities?** By evaluating the ability of LLMs to generate interventions that closely resemble user-generated content, we can assess their potential to seamlessly integrate into online discussions and influence positive behavior.

Additionally, it is crucial to examine the linguistic aspects of LLM-generated interventions. We examine this through our second research question **(RQ2): How well do LLMs emulate the vocabulary, sentence structure, and formality of users in online communities, and to what extent do they replicate users' topical focus and sentiment?** Understanding the linguistic nuances and the ability of LLMs to capture user-specific characteristics is essential for generating interventions that resonate with the target audience and effectively promote prosocial behavior.

Finally, we aim to explore the potential of LLMs to learn from and mimic users who consistently demonstrate prosocial behavior. We address this in our third research question **(RQ3): Can LLMs accurately mimic users with a track record of prosocial contributions and leverage this mimicry to generate similarly positive comments and discussion points?** By identifying and learning from users who exemplify prosocial behavior, LLMs can generate interventions that are more likely to foster positive interactions and contribute to the overall well-being of the community.

In summary, the increasing importance of online communities and the need for prosocial behavior within these spaces highlight the significance of our research. By investigating the effectiveness of LLMs in generating prosocial interventions, we aim to contribute to the development of tools and strategies that can foster positive interactions, support vulnerable individuals, and promote a healthy and thriving online community. The potential impact of this research extends beyond academic curiosity, as it holds promise for real-world applications in moderating and nurturing online spaces that benefit society as a whole.

2 PRIOR WORK

The exploration of prosocial behavior within online communities, particularly through the use of Large Language Models (LLMs), is an active area of research that intersects with the themes of community moderation, user behavior, and content generation.

Bao et al. [2] explore the correlation between the absence of antisocial behavior and the presence of prosocial behavior in online discussions. Utilizing the Perspective API to evaluate toxicity levels, the research aims to understand the dynamic between community moderation efforts and the promotion of prosocial interactions. This aligns with our work's emphasis on using LLMs to generate positive interventions by mimicking prosocial user behaviors within these communities.

Choi et al. [6] discuss the design of ConvEx, a system leveraging AI to augment moderation practices in online platforms. Their approach includes visualizing conversational metrics to help moderators identify and manage problematic discussions before they escalate. The adaptive and proactive moderation techniques explored are similar to our research's focus on LLM-generated content to preemptively guide conversations towards prosocial outcomes.

Aher et al. [1] present a method using LLMs to simulate human responses in controlled experimental setups, such as the Ultimatum Game. Their study demonstrates the consistency and predictive power of LLMs in replicating complex human behaviors based on structured inputs. This directly complements our study's aim of generating LLM-based responses that not only replicate the linguistic style of users but also adhere to prosocial norms.

Park et al. [9] introduced 'Social Simulacra,' a technique leveraging large language models like GPT-3 to simulate a wide variety of social interactions within online communities. This method allows designers to prototype and refine community interactions by exploring various "what if?" scenarios and adjusting designs accordingly. This work aligns with our research, highlighting the potential of LLMs to generate prosocial interventions and manage online communities effectively, thereby emphasizing the importance of aligning generated content with community values.

Santurkar et al. [10] provide a foundational analysis of how LLMs align with the diverse set of human opinions across various demographic groups. They developed the OpinionQA dataset to measure the representativeness and steerability of LLMs against public opinion data. Their findings indicate a significant misalignment between LLM-generated responses and the opinions of specific U.S. demographic groups, such as individuals over the age of 65 and those identifying as widowed. Additionally, even with targeted steering attempts, LLMs often fail to move away from their inherent biases, mainly reflecting the viewpoints of younger, more liberal, and technologically adept demographics. This misalignment raises critical considerations for our study as we explore the use of LLMs to emulate user vocabulary and sentence structures in prosocial interventions on platforms like Reddit. Ensuring that such models sufficiently represent the full diversity of community opinions is crucial for promoting genuinely supportive and inclusive environments.

These studies collectively emphasize the potential of LLMs to influence online community dynamics positively. They highlight

how advanced technologies in Artificial Intelligence like Large Language Models can be harnessed not only for understanding user interactions but also for actively shaping them in a manner that promotes supportive, engaging, and healthy online environments. The integration of LLMs offers a promising opportunity for enhancing the quality of interactions in digital spaces.

3 DATA COLLECTION

We chose Reddit as our data source for conducting this research because of its diverse user base, active discussions, and availability of good-quality textual data. Reddit is home to many thriving online communities, also called subreddits, that can be found for a wide range of topics. These communities contain extensive user interactions and discussions.

Reddit is an ideal platform for this study due to the many prosocial communities that are dedicated specifically to promoting positivity, empathy, kindness, and support among users. These communities provide an abundance of textual data in the form of posts, comments, and user interactions. This textual data is crucial as it captures the writing styles, users' sentiments, and also the topical focus of users within the communities. The user base found on Reddit is also extremely diverse - spanning across different demographics, backgrounds, and varying perspectives. This diversity helps enhance the generalizability of the findings to a great extent as it covers many language patterns and styles.

Data on Reddit is made available to the public through a publicly accessible API which makes it easy to use. For this research, we made use of PRAW (Python Reddit API Wrapper), which is a Python package that allows researchers and developers to access and retrieve data from Reddit using the API in an extremely simple and efficient manner.

For this research study, we took into consideration only communities with prosocial interactions. Based on community popularity, positive and constructive interactions amongst users, and their supportive environment, we carefully curated a list of seven subreddits for the research - CasualConversation, LifeProTips, GetMotivated, DecidingToBeBetter, CongratsLikeImFive, offmychest, and AskReddit.

The CasualConversation subreddit is dedicated to light-hearted and casual conversations without a topical focus that promotes a friendly and welcoming environment for user engagement. LifeProTips, on the other hand, is more of a hub that is used to share practice advice in life that is aimed at improving various aspects of life. This community is focused on personal growth and mutual support. Subreddits like GetMotivated and DecidingToBeBetter cater to Reddit users that are seeking motivation and encouragement to overcome challenges and bring about positivity in life - also without a topic focus. These communities have super uplifting content that aligns closely with the goal of this study. The CongratsLikeImFive and offmychest subreddits are dedicated to celebrating user achievements, no matter how small, and also provide a safe space for individuals to share personal stories and struggles in a supportive community. Finally, we also included the AskReddit community to explore a more diverse range of perspectives and discussions on more open-ended questions and further broaden the spectrum of language patterns and interaction styles.

We aimed to capture a comprehensive representation of user interactions and community dynamics in the chosen prosocial subreddits for this user study. To achieve this, we took into consideration three primary data points - posts, comments, and karma points associated with posts and comments.

The posts represent content submissions that are made by users within a given subreddit community. Posts can vary from a text-based discussion, question, or personal story with additional media content. These posts were then used to analyze the language patterns and writing styles of users within subreddit communities. In addition to posts, we also collected comments made by the users in response to the posts. Comments help provide more context in understanding the community's reactions and opinions surrounding the original post. Finally, Reddit's Karma system is a key feature of the application that helps quantify user engagement and community appreciation for specific posts and comments through Karma points. We used these Karma points to shortlist top posts for different subreddit communities.

Through our extensive data collection pipeline developed using Python, we assembled a significant-sized dataset consisting of 603 posts from the different prosocial communities. These posts were contributed by a total of 201 unique authors. This ensured a diverse representation of writing styles and perspectives. We further ensured that each of the unique authors also had a total minimum of three posts in our dataset. This threshold was to ensure that there was a sufficient sample size for each user that allowed for accurate modeling and representation.

Even though the dataset included postings from several subreddits, we realized that in order to enable effective analysis, both the quantity and the complexity of the data needed to be managed. As a result, we set a 75-comment limit on the total number of comments that can be gathered for each post. This way, we were able to preserve a manageable dataset size for our computational resources while still capturing a representative sample of community interactions.

Finally, We obtained a broad and diverse corpus of user-generated content from the dataset which comprised of 603 posts from 201 distinct authors, with a minimum of three posts per user and a maximum of 75 comments per post. Our study about the efficacy of large language models in mimicking prosocial interventions was made possible by this dataset.

4 METHODOLOGY

4.1 Data Preparation

We begin our methodology by separating the user data into training and evaluation sets. For each user, we allocate the first two posts to the training data and the last post to the evaluation set. This separation allows us to train our models on a portion of the user's content while reserving a post for evaluating the effectiveness of our approach.

4.2 Data Preprocessing

To ensure data quality and consistency, we perform a series of preprocessing steps on both the posts and comments of the user data. The preprocessing pipeline includes the following steps:

(1) **Removing URLs:** We eliminate any URLs present in the text to focus on the actual content.
(2) **Removing punctuation:** We remove all punctuation marks from the text to simplify the data.
(3) **Converting to lowercase:** We convert the entire text to lowercase to ensure uniformity and avoid case-sensitive variations.
(4) **Tokenizing the text:** We tokenize the text, breaking it down into individual words or tokens.
(5) **Removing English stopwords:** We remove common English stopwords that do not contribute significantly to the semantic meaning of the text.
(6) **Joining tokens:** Finally, we join the remaining tokens back into a string to form the preprocessed text.

4.3 Word2Vec and Doc2Vec Training

To capture the semantic meaning of words and perform an analysis of the user posts and comments, we train Word2Vec and Doc2Vec models on the preprocessed data. Word2Vec is a neural network-based model that learns word embeddings, representing words as dense vectors in a high-dimensional space. Doc2Vec, an extension of Word2Vec, learns document-level embeddings, representing entire documents as dense vectors.

4.4 K-Means Clustering and User Archetypes

After obtaining the word and document embeddings, we perform k-means clustering on the vectors generated. We choose the number of clusters to be seven, corresponding to the user archetypes we have defined based on common Reddit user interactions. The user archetypes considered in this experiment are:

(1) **Expert:** Users who demonstrate expertise and provide informative responses in their domain of knowledge.
(2) **Critic:** Users who offer critical opinions and constructive feedback on various topics.
(3) **Karma Seeker:** Users who actively seek attention and upvotes through their posts and comments.
(4) **Story Teller:** Users who engage the community by sharing personal stories and experiences.
(5) **Moderator:** Users who actively moderate and guide discussions to maintain a positive community atmosphere.
(6) **Humorist:** Users who entertain and engage others through humorous and lighthearted content.
(7) **Lurker:** Users who primarily consume content without actively participating in discussions.

4.5 Clustering Evaluation and Visualization

To assess the quality of the clustering results, we calculate the silhouette score, which measures the cohesion within clusters and the separation between clusters. A higher silhouette score indicates better-defined and well-separated clusters.

To visualize the clustering results, we employ t-Distributed Stochastic Neighbor Embedding (t-SNE), a dimensionality reduction technique that projects the high-dimensional vectors onto a 2D plot. The t-SNE visualization provides insights into the distribution and similarity of user archetypes based on their proximity in the plot.

Additionally, we analyze the most similar words to the cluster centers using the trained Word2Vec model. This analysis helps in understanding the key characteristics and topics associated with each user archetype.

4.6 Prompt Generation and LLM Response

To generate prosocial interventions, we prepare a prompt to be fed to the Large Language Model (LLM). In this experiment, we utilize the OpenAI GPT-3.5 language model. The prompt is constructed based on the training posts of each user and includes the following instructions:

> Instructions: Based on the training posts provided, generate a new post that the user might create in the same subreddit. Consider the following:
> (1) Mimic the writing style, vocabulary, sentence structure, and formality of the user's existing posts.
> (2) Focus on topics and sentiments similar to the user's previous posts.
> (3) Generate a thoughtful and coherent post that aligns with the user's interests and contributes positively to the subreddit.

The LLM generates a response based on the prompt, aiming to produce a post that resembles the user's writing style, focuses on similar topics and sentiments, and contributes positively to the subreddit.

4.7 Similarity Evaluation

To evaluate the effectiveness of the LLM in mimicking user style and content, we calculate cosine similarity scores between each author's texts and the generated LLM response for that author. The similarity scores are computed based on the Word2Vec and Doc2Vec embeddings obtained earlier.

We calculate the average cosine similarity scores across all authors/users for both the Word2Vec and Doc2Vec models. These average scores provide an overall measure of how well the LLM responses resemble the writing style and content of the users based on the respective embedding representations.

4.8 Topic Modeling and KL Divergence

To further analyze the thematic similarity between user posts and LLM-generated responses, we perform Latent Dirichlet Allocation (LDA) topic modeling. LDA is a probabilistic model that discovers latent topics within a collection of documents.

We apply LDA separately to the user embeddings and LLM embeddings, obtaining topic distributions for each. To quantify the similarity between the user and LLM topic distributions, we calculate the Kullback-Leibler (KL) divergence. KL divergence measures the difference between two probability distributions, with a lower value indicating higher similarity.

We compute the KL divergence between each pair of user and LLM topic distributions and calculate the mean KL divergence across all pairs. This average KL divergence provides an overall measure of the thematic alignment between user posts and LLM-generated responses.

4.9 Visualization and Interpretation

Finally, we visualize the results of the topic modeling for both user and LLM embeddings. For each topic, we display the top words and their corresponding weights, providing insights into the dominant themes and concepts within the user posts and LLM-generated responses.

The combination of clustering, similarity evaluation, topic modeling, and visualization techniques allows us to comprehensively assess the effectiveness of LLMs in generating prosocial interventions that mimic user style, content, and thematic focus. The results obtained from these analyses provide valuable insights into the potential of LLMs in fostering positive interactions and supporting the well-being of online communities.

5 FINDINGS AND IMPLICATIONS

5.1 RQ1: Effectiveness of LLMs in Mimicking User Style and Content

5.1.1 Findings. The high average cosine similarity scores for both Word2Vec (0.80834573) and Doc2Vec (0.68903158) embeddings indicate that Large Language Models (LLMs) can effectively generate prosocial interventions that closely mimic the style and content of users in online communities. These scores suggest that LLMs can capture and replicate the vocabulary, word-level patterns, and overall document-level structure of user-generated content with a high degree of similarity.

5.1.2 Implications. The effectiveness of LLMs in mimicking user style and content has significant implications for auto-moderating large communities on Reddit. By generating prosocial responses that closely resemble user-generated content, LLMs can help guide discussions in a positive direction and maintain a healthy community atmosphere. This can be particularly beneficial in communities with a high volume of interactions, where manual moderation by human moderators becomes challenging.

The high similarity scores suggest that LLM-generated interventions can seamlessly blend into the conversation, making them more likely to be accepted and influential among users. This implies that implementing LLM-based auto-moderation can alleviate the burden on human moderators, ensuring a more scalable and efficient moderation process while maintaining the quality and tone of the discussions.

5.2 RQ2: LLMs' Ability to Mimic Prosocial User Archetypes

5.2.1 Findings. The user archetype clustering results in Figure 1 show that LLMs can accurately identify and mimic users with a track record of prosocial contributions, such as the "Expert," "Story Teller," and "Humorist" archetypes. This finding suggests that LLMs can effectively recognize and emulate the language patterns and sentiments of users who consistently engage in positive and constructive interactions within the community.

5.2.2 Implications. The ability of LLMs to accurately mimic prosocial user archetypes has important implications for identifying and responding to users in need within Reddit communities. By replicating the language patterns and sentiments of supportive and empathetic users, LLMs can generate targeted interventions for individuals expressing distress, loneliness, or mental health concerns. This can help ensure that vulnerable individuals receive timely support and appropriate assistance.

LLM-generated responses that mimic prosocial user archetypes can provide timely support, offer helpful resources, and direct users to appropriate assistance. This implies that LLMs can play a crucial role in creating a more supportive and inclusive environment on Reddit, where users feel heard and valued, ultimately fostering a sense of belonging and well-being within the community.

5.2.3 User Archetype Visualization.

5.2.4 Findings. While some user archetypes form relatively compact clusters (e.g., Lurker, Humorist), others are more dispersed and overlapping (e.g., Expert, Critic, Karma Seeker). The overlapping nature of certain clusters indicates that users may exhibit characteristics of multiple archetypes simultaneously, highlighting the complexity and diversity of user interactions in online communities.

Users within the same archetype cluster tend to be closer to each other in the visualization, indicating similarity in their text data and behavioral patterns. The proximity of users within a cluster suggests that they share common language patterns, topics of discussion, and engagement styles.

The visualization may reveal outliers or users who do not clearly belong to any specific archetype cluster. These outliers represent users with unique or exceptional behavior patterns that deviate from the identified archetypes.

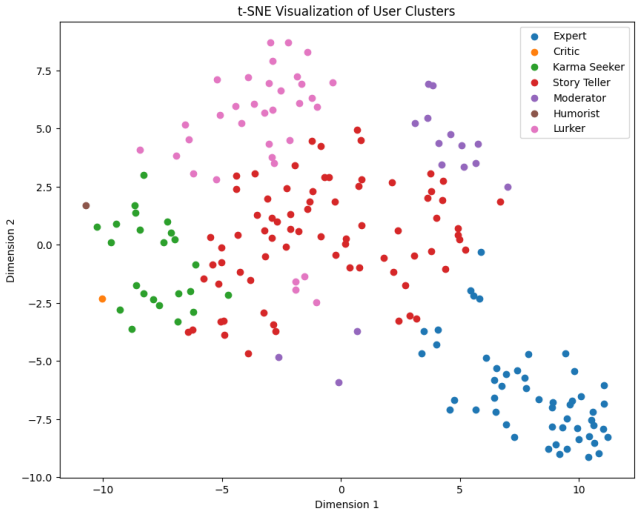

Figure 1: t-SNE Visualisation of User Archetype Clusters

5.2.5 Implications. Understanding the distinct user archetypes can help in tailoring prosocial interventions and moderation strategies to specific user groups. By leveraging the characteristics and preferences of each archetype, moderators can craft interventions that align with the language, tone, and interests of users within those clusters, making the interventions more effective and resonant.

The overlapping nature of some clusters highlights the need for flexible and adaptive intervention strategies that can cater to

users with multiple archetype characteristics. This implies that a one-size-fits-all approach may not be sufficient, and interventions should be designed to accommodate the diversity of user behaviors and needs.

The proximity and similarity of users within archetype clusters can be leveraged to identify similar users and generate targeted prosocial interventions that resonate with specific user groups. This implies that interventions can be personalized based on the common language patterns, topics of discussion, and engagement styles of users within each cluster, enhancing their effectiveness.

Analyzing and understanding the outliers in the visualization can provide insights into niche user behaviors and potential areas for further investigation or specialized intervention strategies. This implies that exceptional cases may require additional attention and tailored approaches to effectively address their unique needs and behaviors within the community.

The visualization of user archetypes can inform the development of targeted prosocial interventions and help identify influential users or opinion leaders within each archetype cluster. By focusing on user archetypes with a higher propensity for prosocial behavior (e.g., Expert, Story Teller) and leveraging their influence, moderators can foster positive interactions and encourage constructive engagement within the community.

Monitoring the evolution of user clusters over time can provide insights into shifts in community dynamics and emerging behavioral patterns. This implies that regular analysis of user archetypes can enable proactive moderation and intervention strategies, allowing moderators to adapt to changing community needs and maintain a healthy and positive environment.

5.3 RQ3: LLMs' Emulation of User Vocabulary, Sentence Structure, and Formality

5.3.1 Findings. The average KL divergence between user and LLM topic distributions (0.606358118973754) indicates a moderate level of similarity in the topical focus of user posts and LLM-generated responses. While LLMs can capture the overall sentiment and tone of user posts, as evidenced by the cosine similarity scores, there is room for improvement in replicating the precise topical focus of users.

5.3.2 Implications. The ability of LLMs to emulate user vocabulary, sentence structure, and formality has implications for promoting community guidelines and values on Reddit. By generating responses that align with the language norms and conventions of specific subreddits, LLMs can help reinforce positive behavior and discourage rule violations. LLM-generated interventions can gently remind users to adhere to community standards, engage in respectful discussions, and maintain a constructive and inclusive environment.

The moderate similarity in topical focus between user posts and LLM responses suggests that LLMs can generate content that is relevant to the ongoing discussions, helping to keep conversations on track and aligned with the subreddit's purpose. By replicating the sentiment and tone of user posts, LLMs can help foster a positive and supportive atmosphere, encouraging users to engage in prosocial behavior and contribute to the overall well-being of the community.

6 CONCLUSION

In this study, we investigated the potential of Large Language Models (LLMs) in generating prosocial interventions within online communities, focusing on Reddit. Our research addressed the effectiveness of LLMs in mimicking user style and content, their ability to emulate user vocabulary and sentence structure, and their potential to accurately mimic prosocial users and generate positive discussions.

Through analysis of user-generated content and LLM-generated responses, we demonstrated that LLMs can effectively capture and reproduce the linguistic characteristics and sentiments of users. High cosine similarity scores for Word2Vec and Doc2Vec embeddings and user archetype clustering results support these findings. The implications of our study are significant for online community moderation and well-being. LLM-generated prosocial interventions can guide discussions positively, maintain a healthy atmosphere, and support vulnerable individuals. LLMs can alleviate the burden on human moderators and foster a supportive and inclusive environment by accurately mimicking prosocial user archetypes.

Ethical considerations, such as ensuring generated content aligns with community values and maintaining transparency and user consent, must be prioritized when using LLMs for prosocial interventions.

Our research highlights the immense potential of LLMs in promoting prosocial behavior and fostering positive interactions within online communities. The insights gained can be applied to various online platforms and offline contexts.

Collaborations between researchers, platform administrators, and community members are essential to develop robust, ethical, and effective strategies for leveraging LLMs to foster prosocial behavior. By working together, we can harness the potential of language models to build online communities that promote well-being, encourage constructive dialogue, and contribute to a more positive digital landscape.

7 LIMITATIONS AND FUTURE WORK

While the research study provides valuable insights into the effectiveness of Large Language Models (LLMs) in generating prosocial interventions, it is important for us to understand and address certain limitations that can potentially impact the generalizability of the research findings.

One significant limitation lies in the potential bias present in our data source. Our main source of data was a very particular subset of prosocial communities on Reddit that may have their own distinct dynamics, user base, and viewpoints. Although they were carefully chosen for their prosocial characteristics, it is possible that they may not have fully represented all the different online communities and user interactions found on various platforms in various circumstances. This bias could potential restrict the applicability of our findings to more extensive online community platforms.

Another limitation of this study was that the concept of utilizing LLMs to mimic and potentially manipulate user behavior would always raise ethical questions with regard to authenticity and also have the potential for misuse. Although the goal of the study is to use LLMs to mimic human prosocial interactions, it could also

potentially be exploited for malicious purposes like spreading misinformation or engaging in deceptive practices. It is important to address these ethical concerns and establish robust guidelines to ensure a positive impact on online communities.

Another disadvantage of our analysis is its complete dependence on Reddit platform data. Despite being a vast and varied source of user-generated material, Reddit's interaction dynamics and conventions could not be the same as those on other social networking sites, online forums, or community spaces.

Future work for this study could include developing real-time prosocial interventions as users engage in online discussions. This would enable more dynamic and responsive moderation efforts that would further help promote prosocial behavior. Additionally, user experience studies to evaluate how users perceive and respond to LLM-generated prosocial interventions can also help provide valuable insights into the effectiveness of these interventions. Next, integrating non-textual cues, such as images, videos, or other multimedia content into the analysis could also enhance the LLM's understanding of the context and potentially improve the accuracy of the generated interventions further.

It would also help to collaborate with interdisciplinary experts including ethicists and social scientists to develop comprehensive ethical guidelines for the responsible use of LLMs in online community interventions. Finally, expanding the research to other data sources and other platforms can also increase the diversity of the dataset and make it more generalizable. Investigating platform-specific nuances and then tailoring the intervention generation process accordingly can lead to more effective, platform-specific interventions.

By acknowledging and addressing these limitations and also pursuing the outlined future directions, we can strengthen the validity and generalizability of the findings to effectively deploy LLMs to foster prosocial interactions across diverse online communities.

REFERENCES

[1] Gati Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. *arXiv preprint arXiv:2208.10264* (2023).

[2] Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of the Web Conference 2021*. ACM, 1417–1429.

[3] C Daniel Batson and Adam A Powell. 2003. Altruism and prosocial behavior. In *Handbook of psychology: Personality and social psychology*. John Wiley & Sons, Inc., 463–484.

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[5] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1217–1230.

[6] Frederick Choi, Tanvi Bajpai, Sowmya Pratipati, and Eshwar Chandrasekharan. 2023. ConvEx: A Visual Conversation Exploration System for Discord Moderators. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–30.

[7] Nancy Eisenberg and Paul Henry Mussen. 1989. The roots of prosocial behavior in children. (1989).

[8] Jiwon Kang, Haeyoon Kim, Hyunwoo Chu, and Chulmo Koo. 2013. The effects of prosocial orientation, social capital, and community engagement on user loyalty in online communities. *Journal of the Korean Society for Quality Management* 41, 4 (2013), 591–602.

[9] Joon Sung Park, Lindsay Popowski, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. *arXiv preprint arXiv:2208.04024* (2022).

[10] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? *arXiv preprint arXiv:2303.17548* (2023).