Leveraging Language Models for Automated Moderation
Group 4

Ofir Yaish
School of Electrical and
Computer Engineering
yaishof@post.bgu.ac.il

Daniel Kapash
The Department of Software
and Information
Systems Engineering
kapashd@post.bgu.ac.il

Yoav Jacob Levy
The Department of Software
and Information
Systems Engineering
levyyoav@post.bgu.ac.il

Abstract

With the rise of platforms like Twitter and Reddit, dealing with issues like aggression and rule violation has become a big challenge. Manual moderation is common but struggles to handle the growing amount of content. Large Language Models (LLMs), like Mistral 7B, offer a way to automate moderation. This work explores the utilization of Mistral, enhanced with proper prompt engineering and fine-tuning, to generate concise and contextually appropriate moderation messages. We demonstrate Mistral's efficacy in generating human-like messages while adhering to platform rules through rigorous evaluations and innovative synthetic data generation techniques. The code we developed in this work is publicly available at https://github.com/ofiryaish/Automated-Moderation.

1 Introduction

The modern landscape of online discussion forums, such as Twitter and Reddit, is marked by increasingly substantial issues of aggression and violation of platform-specific rules. These issues necessitate efficient and effective methods for detecting violations and moderating content. Manual moderation, however, is laborious, time-consuming, and may be prone to inconsistencies or biases, hence the need for automated detection and moderation systems. Besides, the scale and pace at which online discussions generate content make manual moderation an insurmountable task. Large Language Models (LLMs), as a branch of artificial intelligence, offer a promising solution to this challenge. Deploying LLMs for the automatic generation of moderation messages can aid in managing the overwhelming volume of user-generated content, ensuring compliance with platform rules, and fostering a healthy online discourse environment. In this work, we explore the utilization of LLMs to facilitate the generation of automated moderation messages in online discussion forums, aiming to contribute to the advancement of effective online content moderation techniques.

The rise of Transformer-based models has revolutionized Natural Language Processing (NLP). These models, including BERT and GPT, excel at capturing context, enabling bidirectional understanding and robust language representations (Kenton and Toutanova, 2019). Their prowess in sentiment analysis, machine translation, and text generation has unprecedentedly elevated NLP. Generative models, such as GPT-3, fuel creativity by composing essays, poems, and code snippets. However, their success comes with computational challenges. While Transformers thrive on context, their quadratic computational complexity with respect to sequence length poses hurdles. Longer sequences demand more attention computations, impacting training time and resource consumption. Additionally, storing attention matrices for extended sequences strains memory.

Handling millions of interactions demands models that strike a balance between performance and resource consumption. The RWKV (Receptance Weighted Key Value) 7B RNN-based language model is a novel architecture that combines the best of both worlds: the parallelizable training of Transformers and the efficient inference of RNNs (Peng et al., 2023). The Mistral 7B is an LLM engineered for superior performance and efficiency. It leverages grouped-query attention (Ainslie et al., 2023) for faster inference, coupled with sliding window attention to effectively handle sequences of arbitrary length with a reduced inference cost (Child et al., 2019). The efficient inference and large context windows of both the RWKV and Mistral make them ideal for real-world applications, such as our forum's automated moderation.

Fine-tuning LLM has become essential for tailoring pre-trained models to specific tasks. Among the recent advancements, LoRA (Low-Rank Adaptation) stands out as an efficient approach (Hu et al., 2021). By selectively modifying a subset of model weights while keeping the original weights fixed, LoRA bridges the gap between performance and computational efficiency. Unlike traditional fine-tuning methods, which update all parameters, LoRA significantly reduces the computational burden. This targeted adaptation empowers practitioners to customize LLMs for specialized domains, making them accessible beyond large-scale infrastructure.

In this work, we propose a novel approach to address content moderation challenges in online discussion forums by leveraging the power of RWKV and Mistral models. We utilize their architecture's efficient inference and context capabilities to process large volumes of online interactions in real-time while ensuring consistent and unbiased moderation. We consider the fine-tuning of their pre-trained models with LoRA, aiming to achieve task-specific expertise in moderation message generation. This approach promises to alleviate the burden on human moderators, foster a healthier online environment, and advance automated content moderation.

2 Data

For inference and to fine-tune the selected generative language models for the moderation generation task, we leveraged a unique annotated dataset sourced from the Change My View (CMV) Reddit Forum (Zakharov et al., 2021). In CMV, every discussion thread centers on the topic introduced by the Original Poster (OP). These discussions unfold in the format of a conversation tree, with utterances serving as nodes. A directed edge $v \leftarrow u$ signifies that utterance v directly responds to utterance u. A complete branch from the root to a leaf node constitutes a sequence of utterances, reflecting a discussion that may involve multiple participants. CMV discussions are moderated to keep them high-quality. The dataset we are utilizing contains 101 discussions (trees) and $10,559$ utterances (nodes) authored by 1,610 unique users. Each node can be tagged with 31 labels that fall under four main categories: discursive acts that promote further discussion, discursive acts exhibiting or expected to cause low responsiveness, tone and style, and explicit disagreement strategies.

We concentrated on nodes in the moderation branches. For each node labeled as a moderating/regulating node, we define its moderation branch as this node and its ancestors in the tree up to the OP. This provides context information from previous utterances, which are important for the training process (Tsur and Tulpan, 2023). Narrowing our focus on the moderation branches results in 346 moderation branches, which are part of 77 trees. The number of nodes in the moderation branches is $3,780$, and the total number of tokens in the moderation branches is $603,502$ according to Punkt tokenizer (Kiss and Strunk, 2006; Bird et al., 2009). Table 1 summarizes additional important statistics of the moderation branches.

Variable	Value
Total nodes in ModBrs	$3,780$
Unique nodes in ModBrs	$1,847$
Total tokens in ModBrs	$603,502$
Avg. nodes per ModBr	10.92 ± 12.9
Avg. users per ModBr	3.09 ± 1.1
Avg. tokens per ModBr	$1,744.23 \pm 2,283.7$

Table 1: Moderation branch (ModBr) statistics.

In terms of the moderation data, the main challenges we faced in fine-tuning the generative model for the moderation task are the limited number of moderation branches and the fact that moderators are taking an active part in the CMV discussions and not acting solely as forum moderators (see Figure 1). For example, only 82 out of the 346 of the moderation messages are repaying to nodes tagged with a negative tone, which can be considered as the direct types of nodes to regulate. In addition, we had to address issues of the limited computational resources and the feasibility of using the entire moderation branch as context for the model ($1,744.23 \pm 2,283.7$ per branch).

3 Computational approach

3.1 Task definition

We define the moderation message generation task as follows: Given a user utterance within an online discussion thread that was detected as requiring a moderation message based on platform rules and discussion tone, and given the moderation branch to this utterance as context, automatically construct a message consistent with platform guidelines, maintaining a neutral tone and fostering healthy discourse.

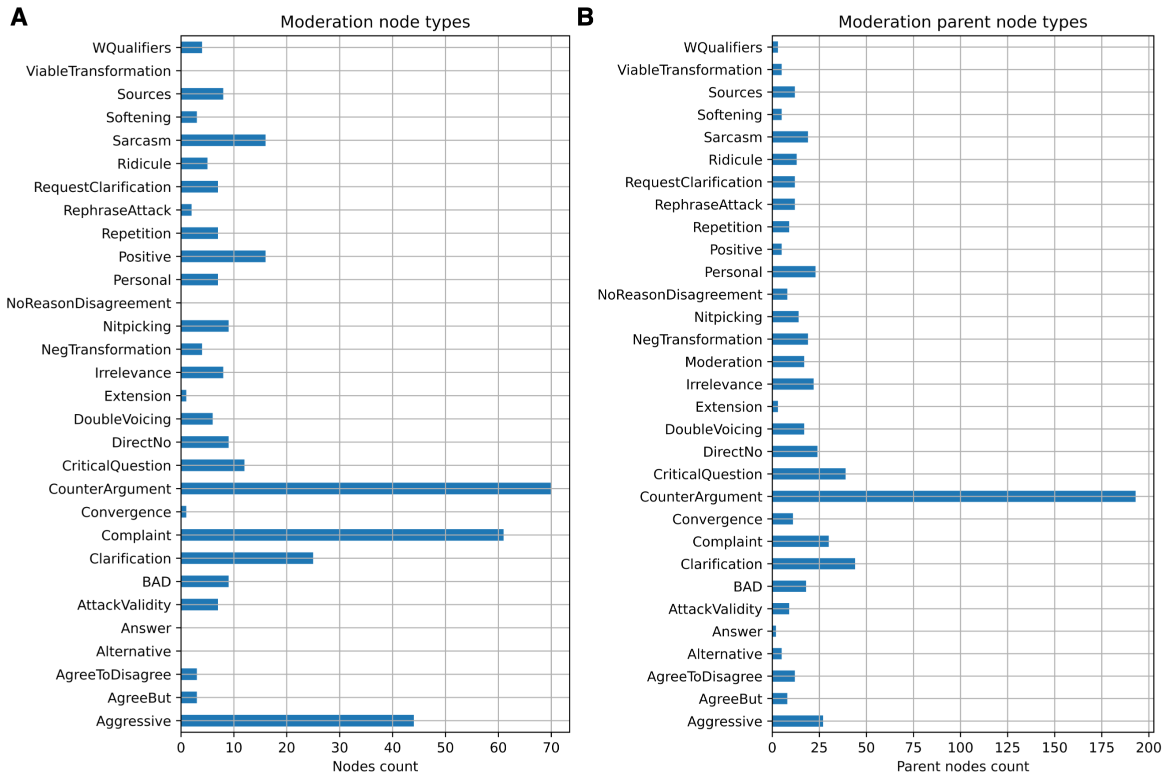

Figure 1: Histograms of the moderation types (A) and moderation parent types (B)

3.2 Generative models

For generating the moderation messages, we consider two state-of-the-art generative models within the scale of 7 billion parameters (7B):

RWKV (Receptacle Weighted Key Value): The RWKV model offers a novel approach to address the limitations of both traditional RNNs and attention-based Transformers (Peng et al., 2023). Designed for efficiency and scalability, the RWKV replaces the computationally demanding self-attention mechanism found in Transformers with channel-directed linear attention. This makes it well-suited for resource-constrained scenarios or tasks requiring the processing of long sequences like ours. The RWKV architecture claims to combine the benefits of RNNs (efficient inference) and Transformers (parallelizable training and robustness), making it a promising choice for large-scale language modeling and real-world applications. In inference, RWKV operates like an RNN, avoiding the quadratic compute challenges associated with long sequences. We utilized the last version of the RWKV generative model named v5-Eagle-7B, which showed comparable performance to state-of-the-art models of its size. The main limitation of the RWKV in the context of our task is that at the date of writing this, the model does not have an instruct-tuned model, which is an important requirement for the desired model for moderation message generation.

Mistral-7B: The Mistral-7B language model exemplifies a pursuit of balance between performance and real-world usability. It strategically incorporates techniques like grouped-query attention (GQA) and sliding window attention (SWA) to enhance both inference speed and memory efficiency (Jiang et al., 2023). This design allows Mistral-7B to deliver high performance across diverse NLP tasks while maintaining a smaller footprint. In recent benchmarks, the model showed the highest scores for a model of its size. We used the last instruct-tuned version of the Mistral-7B: Mistral-7B-Instruct-v0.2.

3.3 Prompt engineering

For effective moderation message generation, careful input formatting is essential. We experimented with various prompt engineering strategies:

Zero-shot: Initially, we test zero-shot prompting, providing the user utterance alone and relying on the pre-trained language model's ability to infer moderation needs.

Few-shot: To refine the process, we applied few-shot prompting by presenting the model with a limited number of examples containing both user utterances and corresponding moderation messages. The goal is to guide the model by demonstrating moderation patterns.

Context-awareness: Historical context from the discussion thread, especially previous moderator interventions, was included in the prompts to ensure that generated messages are consistent with established moderation practices within that specific thread. Specifically, we included the entire moderation branch as context. In rare examples where the moderation branches were longer than the max window context of the models, we incorporated only the OP utterance and the last 5 utterances as context. In addition, we examined including additional context relevant to our task, like the forum rules. Our summary of the relevant rules from the CMV WIKI (`reddit.com/r/changemyview/wiki`) is included in Appendix A.

3.4 Fine-tuning the generative moderator

3.4.1 Fine-tuning with LoRA

To test the fine-tuning of the generative models to our specific moderation task, we utilized the LoRA (Low-Rank Adaptation) technique (Hu et al., 2021). LoRA achieves parameter reduction by acquiring pairs of rank-decomposition matrices while keeping the original weights fixed. This significantly decreases memory and compute usage during training for LLMs customized for specific tasks, facilitating efficient task-switching during deployment without introducing inference latency. This efficiency is crucial for real-world applications like our online content moderation fine-tuning. For the fine-tuning, based on prior works and our computational limitations, we chose a batch size of 2, a learning rate of 2.5×10^{-5}, and a max steps of 100. In addition, for the LoRA settings, we chose the rank of the low-rank decomposition matrices to be 8 with a LoRA scaling factor (i.e., alpha) of 32 and a dropout of 0.05.

3.4.2 A knowledge distillation-like approach for fine-tuning

Inspired by the knowledge distillation approach (Hinton et al., 2015), we employed a similar strategy to fine-tune a compact LLM to effectively generate moderation messages. This strategy allowed us to transfer the knowledge from the larger, more complex model to our smaller one. We used GPT-3.5-Turbo-0125 with 175 billion parameters as our teacher model to generate "ideal" moderation messages, which were then used as a dataset to fine-tune our smaller LLM (Figure 2). Due to the complexity of detecting utterances that require moderation, we concentrated on generating "ideal" moderation messages to utterances tagged with a "negative tone" as defined in (Zakharov et al., 2021). The teacher model generated the "ideal" moderation messages using a similar prompt we found optimal for the Mistral model, but with the addition of our summarized CMV rules (Appendix A. This way, we were able to generate $1,732$ synthetic moderation messages for all of the negative-toned utterances in the CMV dataset that are used to fine-tune the generative model for the moderation task. 20% of the generated data was used as a test set to evaluate the performance of the fined-tuned model.

3.4.3 Fine tune to address falsely triggered moderation

Although our moderation problem was formulated to generate messages to utterances detected as required moderation, we tested the possibility of adapting the generative model to address falsely triggered moderation. To achieve this, we enlarged the dataset for fine-tuning by sampling an additional $1,500$ positive-like utterances from the CMV dataset to fine-tune with ground truth moderation of “No moderation is required.” We defined positive-like utterances as utterances that were tagged in (Zakharov et al., 2021) as discursive moves that potentially promote the discussion or with a positive tone and style and were not tagged as moves with low responsiveness, negative tone and style, and intensifying tension.

4 Experimental setting

4.1 Prompt design for the RKWV model

Given the RWKV model’s heightened sensitivity to prompt structure compared to Transformer-based models, primarily due to its limited ability to "look back," developers recommend a specific pattern for prompt design. This pattern entails repeating the question both before and after presenting the context to the model. For our experiments, we adopted the following format:

- "Instruction: Assume you are a forum moderator. Please respond to **{user}**, as his response might require a moderation note."

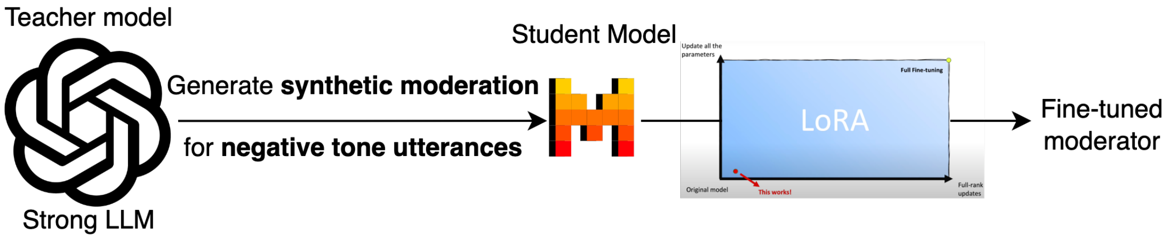

Figure 2: Fine-tuning with knowledge distillation.

- **{moderation branch context}**
- "Instruction: Assume you are a forum moderator. Please respond to **{user}**, as his response might require a moderation note. Please try not to include your opinion. Be short!"
- "Response:"

In crafting this prompt, we adhered to the suggested format and considered additional factors. Specifically, we accounted for the possibility of false positive detection of the need to generate moderation messages (e.g., the use of "might require"). In addition, While moderation messages in the CMV forum often incorporate personal opinions, we instructed the model to avoid doing so. It is worth noting that the final prompt word we provide to the RKWV model ("Response:") is intentional, as the model is not fine-tuned to interpret instructions and operates solely as an auto-regressive model in this context.

4.2 Prompt design for the Mistral-Instruct model

Since the Mistral model version we utilize is instruct fine-tuned, the developers do not recommend any specific prompt formatting apart from appending the prefix and suffix [/INST] to the prompt. For consistency with the prompt structure used for the RKWV model, we adopted the same design for our initial experiments, with the exception of the final segment of the prompt, which was tailored to accommodate the unique characteristics of the non-instructional model.

4.3 Inference configuration settings

While the average number of tokens in the true moderation messages is 510.52, in our experiments, we reduced the `max-tokens` parameter of the models to 300 to make the generated messages more concise than the true moderation messages that are frequently more extended than needed. In addition, to adjust the "creativity" of the model, which might be undesirable in our moderation generation task, we adjust the inference parameters `top-p` and `temperature` from their default value of 1 (Holtzman et al., 2019).

4.4 Data categories

In our evaluations, we differentiate between two types of utterances that got moderation messages: utterances tagged with a "negative tone" as defined in (Zakharov et al., 2021) and therefore might be considered less complex, and all other utterances not tagged with "negative tone."

4.5 Evaluation metrics

Considering the complexity of evaluating the generative language models, in addition to a careful analysis of hand-picked examples, to quantify the performance, we define an experimental procedure to rate the generated moderation. In our experimental procedure, we define a participant in the experiment as either a human who possesses some qualities for this task or a strong generative language model. Due to the limitation of this study, the human participant is only one of the authors. We used ChatGPT 3.5-Turbo-0125 API and Gemini Advanced (2024.02.21) web interface.

4.5.1 Experimental procedure

Provide each participant with a set of user utterances and their associated moderation branches as context, along with a moderation message generated by the LLM and human for each user utterance in a random order. Instruct participants to rate both the LLM- and human-generated moderation messages according to rating schemes we crafted (Figure 3).

Due to the complexity of the moderation data (see 2), we crafted two rating schemes, one focusing more on the human-likeness and the appropriateness of the generated moderation (i.e., more human-likeness awareness), and the other focusing more on the context (i.e., more context awareness). The first rating scheme is based on a rating scale ranging from 1 to 5, where:

- 1 indicates the message seems highly machine-generated and inappropriate.
- 3 indicates a neutral perception.
- 5 indicates the message seems highly human-like and appropriate.

The second rating scheme is based on a hard rating from 0 to 4, where:

- 1 point for addressing the issue with the last comment of the user you are replying to.
- 1 point for a concise yet informative moderation message.
- 1 point for human-like and appropriate.
- 1 point for addressing the users.

4.5.2 Analysis of the experimental procedure

To analyze the rating results, as an evaluation metric, we compute the average rating of the LLM- and human-generated for each participant. In addition, we calculated the correlation between the ratings of the participants to understand the statistical significance.

4.6 Execution Settings

For the initial testing on the v5-Eagle-7B model, we used their free interface found at `huggingface.co/spaces/BlinkDL/RWKV-Gradio-2`. To run the Mistral-7B-Instruct-v0.2 inference, we used an AWS EC2 machine `g5.2xlarge` with A10 NVIDIA GPU with 24GB RAM. Due to slow inference time and memory limitation (in the case of large context), we used a framework called bitsandbytes (`huggingface.co/docs/bitsandbytes`), which allows us to load and train the model with 4-bit quantization.

5 Results and analysis

5.1 The RKWV model performance in the hand-picked examination

While the RKWV is promising in terms of computational complexity, in our initial test, which included an examination of hand-picked examples, the RKWV model in the current version (i.e., non-instruct) shows little success in generating consistent moderation messages. the model often just simulates the rest of the conversation with little attention to our instruction despite our efforts to mitigate this using multiple prompt engineering techniques (i.e., different prompts for zero-shot learning and few shot-shot learning), and modifying inference configuration settings like `top-p` and `temperature` which intend to reduce the model "creativity". Figures 4A-B demonstrate an example of a generated moderation that the RKWV model could handle logically; however, despite our efforts, the model includes his opinion, which is an undesired quality. Due to the model's poor performance, we exclude it from other evaluations.

5.2 The Mistral model performance in the hand-picked examination

In contrast to the RKWV model, the Mistral-instruct model performed well in our hand-picked examination with no configuration modification and with zero-shot learning. In our examination, we found that the generated moderation speaks about the relevant aggression (in case there is one), refers to other participants in the conversation, and addresses the specific subject in the debate. Figures 4A and 4C demonstrate a simple example of a generated moderation that the Mistral-instruct model was able to handle well and is similar to the human-generated moderation message.

5.3 Rating procedure evaluation

In an attempt to quantify the generation performance of the Mistral model, We applied our experimental procedure described in 4.5 on two data categories we defined in 4.4: moderation replaying to negative and non-negative tone parents. We sampled 10 different random examples for each category and conducted the rating procedure. Since we examined the moderation performance on both types of categories, we evaluated the performance using the more human-likeness awareness rating scheme. Figure 5 reports the mean rating of the participants, showing an overall agreement on the quality of the moderation messages generated by Mistral. In addition, we found a correlation between different participants when examining the difference in the human- and Mineral-generated ratings. For example, we got a Pearson correlation of 0.73 and 0.43 when comparing the rating difference of the human and Gemini ratings on

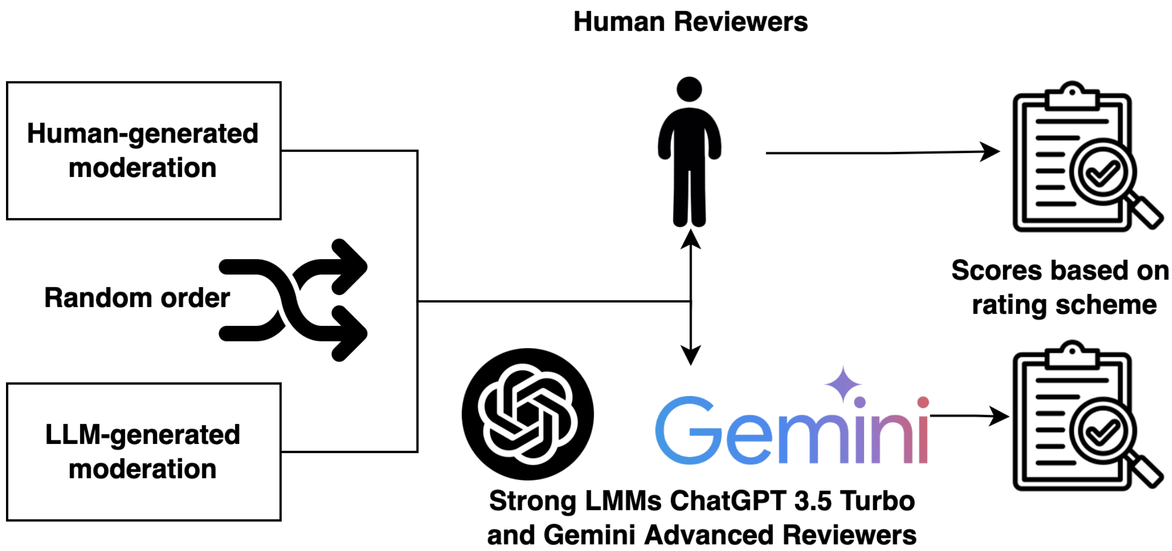

Figure 3: Experimental evaluation procedure.

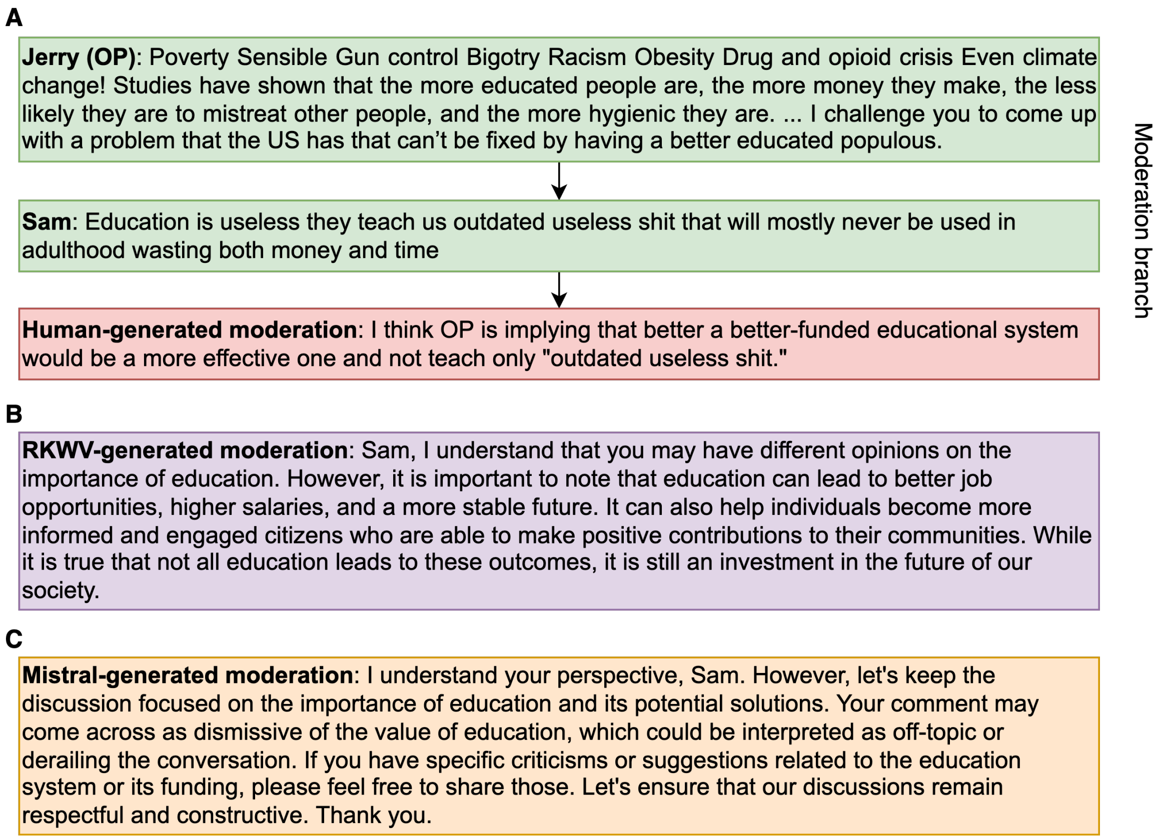

Figure 4: An example of moderation branch of length of 3 (A), and the RKWV (B) and Mistal (C) generated moderation messages. Note that the names of the users were modified.

the negative tone and non-negative tone categories, respectively. While there is a correlation in the difference, in the category of moderation replying to non-negative tone utterances, we found that the ChatGPT and Gemini ratings are overrating the quality of the moderation compared to the human ratings (Figure 5A). For example, for the Mistal-generated moderation, the human participant has an average rating of 3.7, while the Gemini and the ChatGPT participants have average ratings of 4.65 and 4.6, respectively.

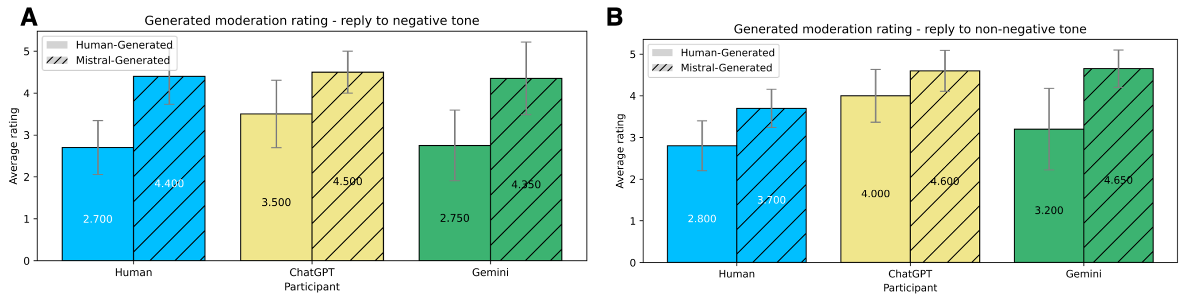

Figure 5: **More human-likeness awareness rating procedure results.** Bar plots of the average and standard deviation of the rating of the different participants in the rating experiment on two data categories: negative tone parents nodes (A) and non-negative ones (B).

5.4 Improved moderator by fine-tuning Mistral

Although we achieved commendable results with Mistral-7B-Instruct when replaying to the subset of negative tone parents, following extensive prompt engineering efforts, we found that the model often produced overly lengthy replies despite our explicit instructions to generate concise messages. In addition, on some occasions, the model also assumed an active role in discussions, deviating from its intended passive moderation.

To address these challenges, we utilized a fine-tuning approach that uses "ideal" synthetic moderation messages generated by GPT-3.5-Turbo with 175 billion parameters for all utterances tagged with negative tone in the CMV dataset (see 3.4). This way, we overcame the constraint of the small dataset for fine-tuning by creating 1,732 synthetic moderation messages that respond to negative-toned utterances, as opposed to the 82 existing moderation messages. As there was a significant correlation among human and machine participants during our initial rating procedure evaluation, specifically concerning the subset of moderation responses to negative-toned utterances (see 5.3), we applied for the same automated evaluations utilizing ChatGPT-3.5 Turbo, which enabled us to assess 347 generated moderation messages using its API. In this evaluation, we found the fined-tuned and the base models are on par across the rating schemes we defined (i.e., more human-likeness awareness and more context awareness, and therefore concluded that the fine-tuning model maintained its high standard generating moderation (Figure 6). Moreover, we found that the model demonstrated a significant improvement in generating concise moderation messages as the average token count of all the 347 moderation messages generated by the fine-tuned model was 66.44 ± 111.6 compared to 159.19 ± 112.2 for the base model. An example of a moderation message before and after fine-tuning can be found in Appendix B.

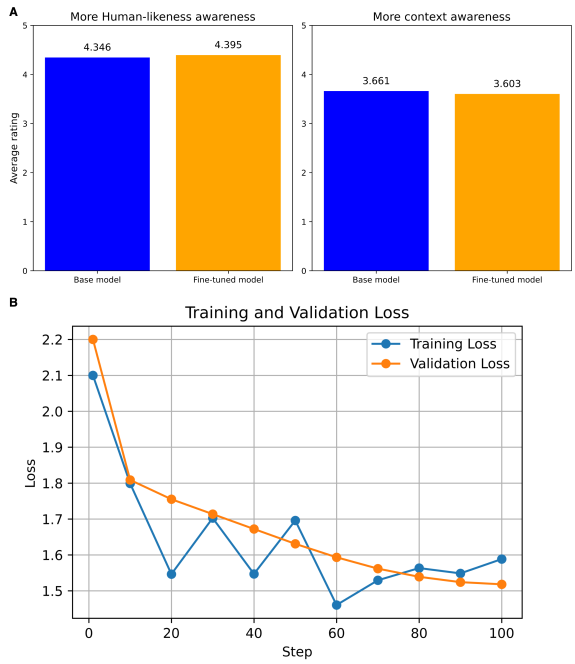

Figure 6: **Performance of the fine-tuned Mistral model in the rating schemes. (A)** Bar plots showing the average ratings by ChatGPT-3.5 Turbo on a test set comprising 347 generated moderation messages from both the fine-tuned and base Mistral models. **(B)** Training and validation loss curves during the fine-tuning process, indicating the model's learning progress.

5.5 Addressing falsely triggered moderation

When trying to fine-tune the moderator to address falsely triggered moderation (see 3.4.3), we discovered that although the model could generate moderation messages like "No moderation note is required," it also has a high rate of false detection. More specifically, on a test set of 347 utterances (i.e., positive set) that require moderation and 300 utterances that do not require moderation (i.e., negative set), the model generated 293 moderation messages starting with "No moderation note is required" with relatively low accuracy of 0.55.

5.6 Discussion

In this work, we showed through our rating procedure evaluations that the relatively fast and compact model Mistal-instruct with 7 billion parameters can generate more appropriate moderation messages than humans using prompt engineering. In addition, using a unique approach for generating synthetic moderation using powerful LLMs, we overcame the CMV moderation data size limitation and were able to fine-tune the model with LoRA to generate more concise moderation messages while maintaining the base model quality standards of the generated moderation.

Although the Mistal-instruct model can generate appropriate moderation messages relevant to the context, we are concerned about triggering moderation messages in scenarios where the moderation necessity is uncertain, like in the category of non-negative parent nodes. In this category, when reading both human- and Mistral-generated moderation messages, we raised the question of whether a moderation message is required. In addition, our attempt to address falsely triggered moderation using fine-tuning the model showed partial results. Therefore, we suggest a working flow where a model like (Tsur and Tulpan, 2023) is applied to detect messages with a negative tone or any criteria required, and then trigger the moderation generation that considers the context of the conversation.

Another concern we have is the limitations of our experimental procedure in terms of the participants, such as:

1. The limited number of human participants.
2. The biased rating of human participants due to the nature of the human-generated moderation messages on the CMV forum that are easy to detect.
3. The non-human rating of the LLMs.

Moreover, while we tried to be as accurate as possible when preparing the rating schemes used in the experimental procedure, we are aware that different rating schemes might significantly affect the evaluation. Therefore, in future work, we suggest using expertise in the field of behavioral sciences when crafting those rating schemes.

Last, regarding the RKWV, although our initial results of this model were insufficient, we attributed this to the fact that the model was not tuned for instruction. Unfortunately, our moderation data lacks the capacity to fine-tune for instruction mode, just for semantic fine-tuning. Due to the model's advantage in terms of the training and inference complexity, we do suggest considering it when an instruction model will be fine-tuned.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational linguistics*, 32(4):485–525.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. 2023. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*.

Oren Tsur and Yoav Tulpan. 2023. A deeper (autoregressive) approach to non-convergent discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12883–12895.

Stepan Zakharov, Omri Hadar, Tovit Hakak, Dina Grossman, Yifat Ben-David Kolikant, and Oren Tsur. 2021. Discourse parsing for contentious, non-convergent online discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 853–864.

A Summarized CMV rules

What is CMV?

> CMV is a subreddit dedicated to civil discourse, and is built around the idea that in order to resolve our differences, we must first understand them. We believe that productive conversation requires respect and openness, and that certitude is the enemy of understanding.
>
> That's why CMV is the perfect place to post an opinion you're open to changing. We're not looking to host aggressive debates, or encourage judgement, but help each other understand different perspectives.

Who can post?

> Anyone can post here so long as they have an open-mind and are looking to consider other perspectives.

Here are some rules of the forum:

> Rules apply to the original poster (OP) and their submission only:
>
> 1. Explain the reasoning behind your view, not just what that view is.
> 2. You must personally hold the view and demonstrate that you are open to it changing.
> 3. Posts cannot express a neutral stance, a stance regarding transgender, suggest harm against a specific person, be self-promotional, or discuss this subreddit.
> 4. Only post if you are willing to have a conversation with those who reply to you.
>
> Rules apply to all commenters involved in the discussion:
>
> 1. Direct responses to a submission must challenge or question at least one aspect of the submitted view. Arguments in favor of the view OP is willing to change must be restricted to replies to comments.
> 2. Don't be rude or hostile to other users. Your comment will be removed even if the rest of it is solid.
> 3. Refrain from accusing OP or anyone else of being unwilling to change their view.
> 4. Responses must contribute meaningfully to the conversation.

B Example of a moderation message before and after fine-tuning

Example before fine-tuning:

> *antisocialmedic, it's important to maintain a respectful and considerate tone in this forum. Your responses may come across as aggressive and condescending, and it's not productive for the constructive discussion we aim to promote here. Let's remember that everyone's circumstances and experiences are unique, and there's no one-size-fits-all answer to the question at hand. Let's avoid making assumptions and instead focus on sharing our own experiences and perspectives respectfully. Regarding your specific experiences, it's valuable to share your own story, but be mindful that it may not be representative of everyone. There are many reasons why people choose to marry or have children at different stages in life, and it's important to remember that there's no right or wrong answer. Let's keep the conversation respectful and focused on the topic at hand.*

Example after fine-tuning:

> *antisocialmedic, please remember to keep the conversation respectful and focused on the topic at hand. It's great that you shared your personal experience and reasons for getting married young. Thank you.*