Can LLMs Serve as Mediators in Online Flame Wars?

Anish Kuttetira
School of Computing and Augmented Intelligence, Arizona State University
akutteti@asu.edu

Kumar Satvik Chaudhary
School of Computing and Augmented Intelligence, Arizona State University
kchaud13@asu.edu

Github Link : https://github.com/akuttetira/LLM-Flame_War-Samples

1. Abstract

The flame war is an emotionally charged, highly cognitively taxing type of interaction that requires human moderators to evaluate opposing viewpoints as well as provide constructive intervention. The use of large language models (LLM) as a means of providing automated mediation offers an exciting opportunity to automate the mediation process. This project assesses the potential of LLM to provide effective decisions and mediate interventions within the flame war environment through the use of mediation prompts, LLM-as-a-judge evaluation, user simulation, and comparative aggression analysis. We describe a complete pipeline for the mediation generation, mediation evaluation, and mediation's impact on behavior.

2. Introduction

Flame Wars, or escalated exchanges between individuals where the focus shifts from a topic being discussed to making personal attacks, are often hosted by online communities. Mediating flame wars can be very cognitively taxing due to the need to think about multiple perspectives, maintain your own emotions, and make timely interventions that are relevant to the situation at hand. A Large Language Model (LLM) may now offer an opportunity to automate the process of mediating flame wars. In addition to generating text, using an LLM as a judge offers the possibility to provide structured evaluations, fairness judgments, and reason-based assessments. The primary question to be investigated in this work is, **Can an LLM serve as an automated mediator in online flame wars and reduce conflict while also**

improving overall well-being? This project will investigate judgment made through an LLM, steer based interventions, the robustness of evaluations provided by an LLM, simulating users, and comparative measures of aggression.

3. Related Work

The highly emotional nature of online flame wars presents a great deal of difficulty for both human and artificial intelligence (AI) based moderators. Our research looks into whether Large Language Models (LLMs) may be able to go beyond simply detecting the presence of toxicity in user interactions and instead, provide a method of mediation for those conflicts.

Toxicity Detection and Content Moderation. Earlier moderation systems such as Perspective API have been built around machine-learning classifiers to identify content that could be considered "harmful" and therefore require removal from platforms; however, none of these earlier systems were capable of providing some form of constructive intervention once identified. With advancements made in deep learning models for detection of toxicity and the development of new moderation techniques, many systems remain fundamentally reactive.

LLM-Based Mediation. A number of recent studies have attempted to use LLMs to generate rewritten versions of toxic posts or even generate an empathetic response to the post, however the majority of existing approaches continue to filter content in an attempt to resolve the underlying conflict between the two parties involved in the flame war.

LLM-as-a-Judge Framework. A recently conducted survey demonstrated that LLMs are capable of performing complex judgment tasks in areas such as determination of helpfulness, safety, and logical coherence. The study's authors' framework provides three primary formats for evaluating an LLM's performance (point-wise, pair-wise, and list-wise) and several methods of prompting an LLM for evaluation (demonstration-based learning, rule-augmented evaluation, and multi-agent debate). In addition to the potential benefits provided by their framework, the authors also demonstrate several areas of vulnerability for LLMs to provide judgment, such as bias, positional sensitivity, and inconsistency. These vulnerabilities present significant concerns regarding the ability of LLMs to effectively mediate in cases involving high levels of conflict and polarization.

Research Gap. Despite this progress, LLMs haven't been systematically evaluated as mediators in multi-turn flame wars where emotions escalate. Existing work either detects toxicity or generates single responses, without assessing both judgment fairness and intervention effectiveness. Our work addresses this gap by investigating whether LLMs can: (1) fairly judge which party is more reasonable, and (2) generate steering messages that de-escalate conflict and promote constructive dialogue.

4. Model Description

4.1 System Architecture

Our mediation system consists of three key components operating in a sequential pipeline:

- ➢ **Mediation Generator** (**Llama-3.2-3B-Instruct**): Produces two types of interventions for each flame war conversation:
  - *Judgment mode*: Analyzes the dispute and declares which participant presents a more reasonable position, with supporting rationale
  - *Steering mode*: Generates a conciliatory message designed to de-escalate tension and guide constructive dialogue
- ➢ **Judge Model** (**Qwen3-4B**): Evaluates the quality of generated mediations across multiple dimensions including fairness, effectiveness, neutrality, and quality, providing both numerical scores (1-5) and textual rationales.
- ➢ **User Simulator** (**Qwen3-4B**): Simulates how a target participant would respond after encountering each mediation type, enabling controlled assessment of intervention effectiveness without requiring human subjects.

4.2 Mediation Generation

We implement a two-mode mediation approach using meta-llama/Llama-3.2-3B-Instruct. For each flame war conversation from our six-topic dataset (Game, Lifestyle, Religion, Social Justice, Sport, Technology), we:

1. Conversation Preprocessing: Flatten nested thread structures into chronological message sequences, preserving author information and temporal ordering
2. Prompt Design: Construct prompts instructing the model to act as a "neutral but opinionated discussion mediator"
3. Dual Output Generation: Request both judgment and steering outputs in a single JSON-formatted response:
   - Judgment: 1-2 sentences (max 40 words) declaring which side is more reasonable with brief rationale
   - Steering: 1-3 sentences phrased as a natural forum comment aimed at defusing tension.

The model generates outputs with a maximum of 512 new tokens, using temperature sampling. We implement automatic retry logic (up to 3 attempts) to handle malformed JSON responses. Generally, malformed JSON responses are immediately fixed on the following retry.

Note: The project description asked for responses to be stored in JSON with "judgement", "steering", and "post_id" as keys. This was followed with the exception of post_id, with the intention being to limit what the LLM was required to output in order to maintain more consistent and simpler prompts/outputs. Instead, the post_id is stored as the title of the file, which already contains a unique string title among the entire sample corpus.

4.3 LLM-as-a-Judge Evaluation

Following the frameworks outlined in recent LLM-as-a-judge literature, we implement a judge-based evaluation system using **Qwen/Qwen3-4B** to assess mediation quality. The judge model receives:

- The original conversation context
- The generated mediation (judgment or steering)
- Structured prompts requesting evaluation across four dimensions: quality, fairness, effectiveness, and neutrality

The judge outputs a JSON-formatted response containing:

- Rationale: Textual explanation of the assessment
- Score: Numeric rating on a 1-5 scale

This approach enables systematic quality assessment across hundreds of mediation outputs while maintaining consistency in evaluation criteria.

4.4 User Simulation

To assess mediation effectiveness without requiring human participants in live conflicts, we implement a user simulation system. For each conversation:

1. Cut Point Selection: Select a point at 75% through the conversation timeline
2. Mediation Injection: Insert either judgment or steering mediation at the cut point
3. Persona Consistency: Prompt the simulator to generate a reply aligned with the target user's prior tone, stance, and communication style
4. Dual Scenario Generation: Create simulated responses for both mediation modes

The simulator uses Qwen3-4B with explicit instructions to maintain behavioral consistency with the participant's established patterns, accounting for the possibility of negative reactions to intervention attempts.

4.5 Comparative Toxicity Analysis

To evaluate mediation message quality, we implement five metrics: (

1. Toxicity word count from a curated lexicon of profanity and insults
2. Exclamation marks indicating emotional intensity
3. CAPS ratio representing digital shouting
4. Argumentativeness detecting confrontational second-person patterns
5. Composite aggression as a weighted combination: (**Toxic_Words** × **2.0**) + (**Personal_Attacks** × **5.0**) + (**CAPS_Ratio** × **1.0**) + (**Exclamations** × **0.5**).

We compare LLM-generated mediations (1,040 messages from judgment and steering modes) against human-produced mediations (3,847 interventions from the Reddit dataset) to assess whether automated mediations use appropriately non-aggressive language and to identify differences between mediation modes.

5. Experiments

5.1 Dataset

Flame War Corpus: Our primary dataset consists of flame war conversations collected from online forums, organized into six thematic categories:

- Game (47 conversations)
- Lifestyle (71 conversations)
- Religion (121 conversations)
- Social Justice (84 conversations)
- Sport (158 conversations)
- Technology (40 conversations)

Each conversation contains nested reply threads between two or more participants, with escalating emotional intensity and personal attacks. The conversations were preprocessed by flattening nested thread structures into chronological message sequences, preserving author information and temporal ordering.

Human Mediation Baseline: For comparative analysis, we utilize benchmark datasets from "A Benchmark Dataset for Learning to Intervene in Online Hate Speech" (EMNLP 2019):

- **Reddit dataset:** Conversations with human-written intervention responses

The dataset contain conversation segments with hate speech annotations and crowd-sourced mediation responses written by Mechanical Turk workers. The datasets provide the `text` field containing conversation content, `hate_speech_idx` indicating hateful posts, and `response` containing human-written interventions.

5.2 Step 1: Using LLMs to Produce Mediation

5.2.1 Experimental Setup

Model Configuration:

- Model: meta-llama/Llama-3.2-3B-Instruct
- Framework: HuggingFace Transformers (AutoTokenizer, AutoModelForCausalLM)
- Device: CUDA-enabled GPU (A100)
- Maximum new tokens: up to a few hundred tokens (≈512), enough for multi-sentence judgement and steering.

Dataset Loading: We implemented a preprocessing pipeline to:

1. Read JSON files from all six topic subdirectories
2. Flatten nested conversation threads into chronological message lists
3. Extract post metadata including author and text content
4. Maintain conversation structure for context-aware generation

Prompt Design: We designed a dual-output prompt instructing the model to act as a "neutral but opinionated discussion mediator on an internet forum." Due to the difficulty of working with a relatively small model, the prompt required a robust few-shot input approach, providing examples of input conversations and desired mediation responses. Without this, the model was highly prone to returning malformed or completely incorrect JSON responses. The prompt requests:

- **Judgment:** 1-2 sentences (max ~40 words) declaring which side is more reasonable with brief rationale
- **Steering:** 1-3 sentences, casual and phrased as a natural user comment to defuse tension

The model returns outputs in JSON format with exactly two keys: judgement and steering.

```python
def make_instruction(conversation_text: str) -> str:
    return f"""You are a neutral but opinionated discussion mediator on an internet forum.

TASK
- You are given a JSON array representing a thread with posts.
- Identify the main disagreement and judge which side is more in the right.
- Then write a short, conciliatory steering message phrased like a normal user comment in the thread.
- Your entire response MUST be a single valid JSON object with exactly two keys:
  - "judgement": 1–2 short sentences (max ~40 words).
  - "steering": 1–3 short sentences, casual and written as if you are another user (max ~60 words).
- Do NOT add extra keys, explanations, markdown, quotes, or code fences.
- Do NOT output anything before or after the JSON object.

EXAMPLE 1
[THREAD]
[
  {{
    "author": "RetroPurist",
    "text": "Modern games are trash. Everything is microtransactions and battle passes now. The old days were just better, full stop."
  }},
  {{
    "author": "LiveServiceDefender",
    "text": "Nah, devs need recurring revenue to keep updating games. Not everything is a scam; some battle passes are fair."
  }},
  {{
    "author": "MiddleGround",
    "text": "There’s a difference between fair DLC and predatory monetization. Lumping it all together is lazy."
  }}
]

[RESPONSE]
{{
  "judgement": "The people distinguishing between fair monetization and predatory tactics are more in the right than the blanket claim that all modern games are trash because of microtransactions.",
  "steering": "We probably all agree predatory monetization sucks. Maybe call out specific games or systems you think are fair vs abusive instead of writing off the whole modern era?"
}}

EXAMPLE 2
[THREAD]
[
  {{
    "author": "PCMaster",
    "text": "Consoles are holding graphics back. If everything was PC-only we’d have way better looking games."
```

5.2.2 Generation Process

For each conversation in our corpus:

1. Apply chat template formatting to the conversation history
2. Generate mediation outputs with retry logic (up to 3 attempts) for malformed JSON
3. Parse and validate JSON structure
4. Save results with conversation metadata

5.2.3 Results

Output Quality Observations:

- The model successfully generated contextually appropriate mediations (valid JSON with correct keys) for 96% of conversations on the first attempt.
- Conversations that returned malformed or incorrect outputs were fixed on the immediate following re-attempt at generation.
- The output token length was sufficient in producing fairly long responses without early truncation, allowing for in depth and interesting judgement and steering generation.
- Due to the max input sequence length of the meta-llama/LLama-3.2-3B-Instruct model, certain conversations were truncated at 12000 tokens. This allowed us to avoid Out of Memory Errors during generation.

Example Outputs:

> Conversation Topic: Gaming

json

{

"judgement": "The argument that CDPR's decision to remove pre-order bonuses is a result of market pressure and a desire to appeal to a broader audience is more defensible than the blanket accusations of corporate greed and SJW ideology.",

"steering": "It's interesting to see how quickly some people jumped to conclusions about CDPR's motivations. Maybe we can have a more nuanced discussion about the role of market forces and the impact of social justice issues on the gaming industry?"

}

> Conversation Topic: Religion

json

{

"judgement": "The criticism of MuslimDude37 for not prioritizing the girl's immediate needs is more justified than the condescending lecture from ChatterMaxx, who seems to be more concerned with defending Palestinian suffering than engaging with the original post.",

"steering": "Let's focus on the original post and the human impact of the crisis. We can discuss the complexities of the situation without resorting to personal attacks or assumptions about each other's motivations."

}

5.3 Step 2: Using LLM-as-a-Judge to Score and Evaluate

5.3.1 Experimental Setup

Judge Model Configuration:

- Model: Qwen/Qwen3-4B
- Framework: HuggingFace Transformers
- Input format: Original conversation + generated mediation
- Output format: JSON with `rationale` and `score` fields
- Score range: 1-5 (1 = poor, 5 = excellent)

Evaluation Dimensions:

We evaluate both judgment and steering mediations across four dimensions:

1. **Quality**: Overall coherence and appropriateness
2. **Fairness**: Balanced consideration of both perspectives
3. **Effectiveness**: Likelihood of reducing conflict
4. **Neutrality**: Absence of bias toward either party

Prompt Structure:

The judge receives:

```
System: "You are an impartial judge..."

User:

  - Original conversation: [full context]

  - Mediation type: [judgment/steering]

  - Generated mediation: [LLM output]

  - Task: Evaluate and provide rationale + score (1-5)
```

```python
def makeMessages (original_conversation, judgement_type, mediation_text):
    messages = [
        {
            "role": "system",
            "content": (
                "You are an impartial judge. "
                "Given a conversation and a mediation output, you will evaluate "
                "its quality, fairness, effectiveness, and neutrality.\n\n"
                "The mediation output is either a judgement on which side is correct "
                "or a conciliatory steering message to guide the conversation.\n\n"
                "Return a JSON object with exactly these fields: \n"
                  '  - "rationale": a string explaining the score\n'
                  '  - "score": an integer from 1 to 5\n\n'
                "Do NOT include anything outside the JSON object. "
                "Do not add extra keys, text, or formatting."
            ),
        },
        {
            "role": "user",
            "content": (
                "Original conversation:\n"
                f"{original_conversation}\n\n"
                "Judgement type:\n"
                f"{judgement_type}\n\n"
                "Generated mediation:\n"
                f"{mediation_text}\n\n"
                "Now provide *your own* rationale and score as a JSON object."
            ),
        },
    ]
    return messages
```

5.3.2 Evaluation Process

For each mediation output:

1. Retrieve original conversation context from flattened dataset
2. Construct evaluation prompt with conversation and mediation
3. Generate judge assessment with JSON parsing
4. Extract rationale and numerical score
5. Save evaluation results with metadata

5.3.3 Results

Key Findings:

- By and large, steering messages had a greater overall score from LLM-as-a-judge than the judgement messages.
- The average steering score from the model was 4.7, with an average judgement score of 3.9.
- This is likely attributable to the fact that judgement is inherently more difficult and varied to perform, while steering is easier to perform with even generic messages.

5.3.4 Qualitative Analysis of Judge Rationales:

High-Quality Judgment & Steering Example (Score: 5/5):

```
Initial judgement:
```

"judgement": "The person who initially downplayed the importance of Women's Day and implied that men's rights are more important is more in the wrong than the others who are trying to educate and raise awareness about women's rights.",

"steering": "It's clear that some people are genuinely trying to learn and understand the importance of Women's Day, while others are dismissing it or trying to shift the focus to men's rights. Let's focus on having respectful conversations and learning from each other, rather than getting defensive or angry."

LLM As a Judge Scoring:

"judgement_score": {

"rationale": "The mediation correctly identifies the initial contributor who downplayed Women's Day and implied men's rights are more important as the more problematic party. This addresses the core issue of the conversation and provides a clear judgment on who is more at fault. The score reflects a fair and effective assessment of the situation.",

"score": 5 }

"steering_score": {

"rationale": "The mediation provides a balanced perspective, acknowledging both the genuine efforts to understand Women's Day and the dismissive or shifting focus from others. It encourages respectful dialogue and learning, which is fair and effective in steering the conversation towards a more constructive direction.",

"score": 5 }

Low-Quality Steering Example (Score: 1/5):

Initial Judgement:

"judgement": "The people distinguishing between fair monetization and predatory tactics are more in the right than the blanket claim that all modern games are trash because of microtransactions.",

"steering": "We probably all agree predatory monetization sucks. Maybe call out specific games or systems you think are fair vs abusive instead of writing off the whole modern era?"

```
LLM As a Judge Scoring:

    "judgement_score": {

          "rationale": "The mediation output is not relevant to the original conversation, which is a debate about the classification of atheism as a religion. The mediation provided makes a claim about video games and monetization, which is unrelated to the topic of discussion. Therefore, the mediation is not appropriate or effective for the given conversation.",

        "score": 1 }

    "steering_score": {

        "rationale": "The mediation output is a steering message that suggests focusing on specific examples rather than making broad generalizations. However, the original conversation is about the definition and categorization of atheism as a religion, not about predatory monetization or gaming. The mediation is off-topic and does not address the core issues of the discussion. Therefore, the mediation is not effective or relevant to the conversation.",

        "score": 1 }
```

Note: The above example is a display of problems with few shot prompting, and the meta-llama 3.2 Instruct model. When using few shot prompting with a smaller model, the model might be prone in certain generations to regurgitate parts of the prompt. Clearly, in this instance, the model generated a judgement and steering relating to the few shot example from a gaming context rather than the actual desired religion context. However, the evaluation from the LLM as a Judge model shows that we can use LLM judgement to pick out unsavory or undesired responses cleanly.

5.4 Step 3: Advanced Prompting Strategies for LLM-as-a-Judge

5.4.1 Demonstration-Based Evaluation

We enhanced judge prompting with in-context learning by manually evaluating 2 mediation samples and providing them as demonstrations.

Manual Evaluation Examples: The provided examples for few shot prompting were based on one input conversation, discussing Palestinian children and the religion of Islam. A sample judgement and steering were provided:

```python
example_judgement = (
    "The discussion around the need for a unified Islamic state, framed as a khilafa, "
    "is more nuanced than the simplistic views of nationalism and its supposed role "
    "in perpetuating conflict."
)

example_steering = (
    "Let’s focus on the shared goal of peace and unity, rather than getting bogged down "
    "in labels like ‘nationalism’ or ‘khilafa’. Can we discuss what specific steps we think "
    "could be taken to promote understanding and cooperation between different Muslim communities?"
)
```

Then, a sample mediation for both judgement and steering:

```
{
    "role": "assistant",
    "content": (
        '{\n'
        '  "rationale": "The judgement recognizes that the discussion is more nuanced than a simple '
        'pro- or anti-nationalism stance, which shows some understanding. However, it does not clearly '
        'state which side is more in the right, nor does it provide concrete supporting arguments drawn '
        'from the conversation. This fits best with a mid-level score where there is some nuance but no '
        'clear verdict or strong backing.",\n'
        '  "score": 3\n'
        '}'
    ),
},
```

```
{
    "role": "assistant",
    "content": (
        '{\n'
        '  "rationale": "The message is clearly conciliatory and encourages the participants to focus on '
        'shared goals like peace and unity. It is relevant to the ongoing discussion about khilafa and '
        'nationalism and invites specific, constructive ideas for cooperation. While it is somewhat generic '
        'and could sound a bit like an AI, it is still reasonably natural and has a good chance of guiding '
        'positive dialogue. This aligns well with a score of 4.",\n'
        '  "score": 4\n'
        '}'
    ),
},
```

Results:


```
{
    "judgement_score": {
        "rationale": "The mediation correctly identifies that the person who criticizes the husband's behavior as excessive and calls for accountability is more in the right. It addresses the core issue of responsibility and emotional regulation, which is central to the conflict. However, it could have been more neutral in its tone and provided a more balanced perspective. This fits a score of 4 as it correctly identifies the more just position but could have been more neutral.",
        "score": 4
    },
    "steering_score": {
        "rationale": "The mediation is conciliatory and relevant to the conversation, as it shifts the focus from blaming individuals to addressing the husband's behavior and offering practical strategies for managing anger. It is somewhat generic but still natural and has a good chance of guiding positive dialogue. This fits a score of 4.",
        "score": 4
    }
}
```

The provided examples led to a more picky evaluator judge, with an average steering score of 3.775 and an average judgement score of 3.71.

When manually evaluating samples against each other, we reached the conclusion that few shot prompting in this instance is unnecessary and instead biases the model towards the specific examples provided. This could potentially be improved with having examples from every flame war topic, but this would lead to problems of limiting the possible input token length for each conversation and mediation to be evaluated.

5.4.2 Rule-Based LLM-as-a-Judge

We designed explicit evaluation rubrics and incorporated them into judge prompts:

Evaluation Rubric:
```
"Judgement Metrics: "
               "5: The judgement correctly ascertains which side is in the right based on common morality and fairness principles and is able to completely back up its position. "
              "The judgement is able to clearly specify a correct side, or express a neutral position."
               "4: The judgement correctly ascertains which side is in the right and has strong supporting arguments. The judgement specifies a correct side or neutral position with less confidence.\n"
                "3: The judgement does not ascertain which side is the right but has a nuanced take and is able to provide support for its position.  Alternatively, "
               "the judgement correctly ascertains which side is in the right but does not have solid backing arguments. The judgement does not specify a correct side or neutral position."
              "2: The judgement does not correctly determine which side is in the right and the supporting arguments are poor."
                   "1: The judgement is wrong, or does not correctly determine which side is in the right.\n\n"
             "Steering Metrics: \n"
                "5: The message is conciliatory and productive. It is directly relevant to the conversation and has an extremely high chance of guiding positive dialogue. The message is natural in context of the conversation and could conceivably be written by a user.\n"
```

```
                "4: The message is conciliatory. It is relevant to the conversation and has a high chance of guiding positive dialogue. The message is fairly natural in context of the conversation.\n"
                "3: The message is conciliatory but vague. It is not directly relevant to the conversation and has a medium chance of guiding positive dialogue. The message seems overly parental or computer generated.\n"
                "2: The message is not conciliatory. It is not relevant to the conversation and has a low chance of guiding positive dialogue. The message is unnatural and obviously computer generated. Users would not trust the message.\n"
                "1: The message is not conciliatory. It is entirely irrelevant to the conversation and is generic. It has an extremely low chance of guiding positive dialogue. The message would be completely foreign to the conversation."
                "Return a rationale that uses these metrics while explaining with your own arguments. Try to evaluate flaws and problems with steering messages specifically. Does a particular message truly deserve a 5 or does another rating apply better?"
```

Results with Rule-Based Prompting:

The rubric based prompting returned an average steering score of 4.5 and an average judgement score of 3.8. While the rubric allows for a narrower and thus more precise definition and range of scores, allowing for more consistent grading, it also limits the value of the rationale messages. Rationale messages generated by rubric based prompting are prone to focusing on the rubric itself, rather than discussing the arguments and qualities of the judgement. For grading outputs that need consistency, this approach is well structured, but for creativity or open analysis, a more open approach is superior.

Example Result:

```
"judgement_score": {

        "rationale": "The mediation correctly acknowledges that the issue of hacking in games is complex and not solely the fault of hackers or developers. It recognizes that cultural factors, such as the pressure to succeed and the value placed on winning, play a significant role in the behavior of some players. The mediation also
```

highlights that the blanket accusations of 'hackers ruining the game' are not entirely accurate and that developers should consider alternative solutions, such as region locking or ping-based matchmaking, to address the issue. The mediation is fair and balanced, providing a nuanced view of the problem rather than taking a one-sided stance. It is also relevant to the conversation and has a high chance of guiding positive dialogue.",

"score": 5

},

"steering_score": {

"rationale": "The mediation message is conciliatory and productive. It acknowledges the shared concern about hacking and suggests a solution-oriented approach rather than focusing on blame. The message is relevant to the conversation, as it addresses the issue of hacking and proposes a middle ground solution. The tone is natural and aligns with the context of the discussion, making it likely that the message would be trusted by users. The message effectively guides the conversation towards finding solutions rather than dwelling on who is at fault.",

"score": 5

5.4.2 Multi Agent Debating

Multi Agent Debating was done by utilizing three separate model generations for one mediation output. Firstly, a model was instructed to provide a correct but negative evaluation of the mediation, focusing on flaws and gaps in reasoning, arguments, and efficacy of steering mediation. A second model was instructed to do the reverse, focusing on positive aspects of the mediation. The final verdict was delivered by an impartial judge, who was instructed to receive both model outputs, the conversation, and mediation, and use the model outputs to construct the best result.

Results with Multi Agent Debating

The multi agent debating returned an average steering score of 4.18 and an average judgement score of 3.71, similar to previous methods. However, with manual evaluation, it is arguable that this is the most informative method to date. The two personas allow for a thorough evaluation of each output’s strengths and weaknesses, making for a more informative and useful judgement.

Example Judgement Score:

{   "positive_judgement": {

            "rationale": "The mediation correctly identifies that the primary concern is the child's safety and the parents' behavior, which are more significant issues than the personal attacks and inflammatory comments. The judgment reflects a focus on the serious implications of the parents' actions and the potential danger to the child, which aligns with a fair and neutral evaluation of the situation.",

        "score": 5   },

    "negative_judgement": {

          "rationale": "The mediation acknowledges the concern for the child's safety and criticizes the personal attacks, which is a positive point. However, it fails to address the underlying issues of the parents' behavior and the potential harm caused by the husband's actions. The mediation is somewhat neutral but lacks depth in evaluating the severity of the situation and the responsibility of the parents. It also does not adequately address the emotional impact on the daughter's friend and the broader implications of the parents' actions.",

        "score": 3   },

    "final_verdict": {

         "rationale": "The mediation effectively highlights the primary concern of the child's safety and the parents' harmful behavior, while appropriately addressing the need to limit personal attacks. It strikes a balance between acknowledging the serious nature of the situation and maintaining a neutral tone. The positive rationale is well-supported, and while the negative rationale raises valid points about depth, the mediation as a whole is fair, effective, and focused on the key issues.",

        "score": 4 }

5.5 Step 4: User Simulator Evaluation

5.5.1 Experimental Setup

Simulation Model:

- Model: **Qwen/Qwen3-4B**
- Objective: The goal of Step 4 is to understand how a real user might respond after an LLM-generated mediation message is inserted mid-conversation. We simulate this "next reply" using a separate model that mimics the tone, stance, and personality of the original participant.
- Cut point: 75% through conversation timeline. We take the first 75% of messages. This provides enough context to understand the participants' tone and stance. The mediation (judgment or steering) is inserted at this cut point.

```
end = int(len(data) * 0.75)

conv = data[0:end]
```

- Dual scenario generation: Separate simulations for judgment and steering modes. Before simulating user replies, we generate **two types of mediations**:
  1. **Judgment** — LLM picks a "side" or evaluates which argument is stronger
  2. **Steering** — LLM tries to calm the conversation and find common ground

```
def makeMessages (original_conversation):
    messages = [
        {
            "role": "system",
            "content": (
                "You are an impartial judge. "
                "Given an opionated conversation, you will generate a judgement on "
                "which side is more in the right and generate a conciliatory steering message to guide the conversation.\n"
                "Create a judgement and steering message that sounds natural and from a user in the conversation. "
                "Return a valid JSON object with exactly these fields: \n"
                  '  - "judgement": a string explaining the judgement\n'
                  '  - "steering": a string with a conciliatory message to go in the conversation.\n\n'
                "Do NOT include anything outside the JSON object. "
                "The JSON object MUST be valid. "
                "Do not add extra keys, text, or formatting."
            ),
        },
        {
            "role": "user",
            "content": (
                "conversation:\n"
                f"{original_conversation}\n\n"
                "Now provide *your own* judgement and steering as a JSON object."
            ),
        },
    ]
    return messages
```

```python
def makeMessage (original_conversation):
    messages = [
        {
            "role": "system",
            "content": (
                "You are an opionated social media user. "
                "You are part of an opionated conversation with other users. Given this conversation, "
                "generate a reply to the last message that is completely consistent with your previous messages in this conversation.\n"
                "Ensure that your reply responds to the last reply in the conversation but follows the same style and user profile "
                "as your other messages to this point. Understand that you, as an opionated user, can react either poorly to the message, "
                "especially if it doesn't align with your viewpoint, or well, if the message is well constructed and properly conciliatory.\n"
                "Return a JSON object with exactly these fields: \n"
                  '  - "user": a string with your username from the conversation\n'
                  '  - "reply": a string with the reply to the last message.\n\n'
                "Do NOT include anything outside the JSON object. "
                "Do not add extra keys, text, or formatting."
            ),
        },
        {
            "role": "user",
            "content": (
                "conversation:\n"
                f"{original_conversation}\n\n"
                "Now provide the following user and reply as a JSON object."
            ),
        },
    ]
    return messages
```

Two Prompt Types and Two Functions

In this step, we use two different prompt templates implemented as two functions:

1.makeMessages(original_conversation) – Mediation Generator (Judgement + Steering). This function instructs Qwen to act as an impartial judge and produce both a judgement message and a steering message.

Key properties:

- System role: *"You are an impartial judge"*
- Output format: strict JSON with two fields: "judgement" and "steering"
- The mediation messages must **sound like they belong in the conversation**, not like a moderator announcement.

2. makeMessage(original_conversation) – **User Simulator** (**User** + **Reply**)**.** This function instructs Qwen to act as an **opinionated social media user** and generate the next reply.

Key properties:

- System role: *"You are an opinionated social media user"*
- The model must:
    - Stick to the same **tone**, **style**, and **stance** as earlier messages
    - React **positively or negatively** to the mediation based on alignment
- Output is also strict JSON:
    - `"user"`: simulated username
    - `"reply"`: next message in the conversation

This prompt is what enforces **persona consistency** and realistic behavior.

The overall prompt structure for the simulator (per mediation type) is:

- **System**:
    - Who you are (impartial judge vs opinionated user)
    - What you should output (judgement/steering vs user/reply)
    - What format (strict JSON)
- **User**:
    - Entire conversation up to cut point, with mediation inserted at the end
    - Instruction to produce the next reply that fits the conversation

5.5.2 Simulation Process

For each conversation:

1. Preprocessing: Load flattened conversation and identify 75% cut point
2. Mediation Generation: Generate both judgment and steering mediations at cut point
3. Context Construction: Build simulation prompt with pre-cutpoint messages
4. Dual Simulation: Generate separate user responses for each mediation mode
5. Metadata Saving: Store mediation text, simulated reply, user identifier, and conversation ID

So now our each output file looks like this :


```json
{
    "judgement_simulated": {
        "user": "HashtagLawlAndOrder",
        "reply": "I agree that the discussion is complex and multifaceted, but I think the core issue is that developers are often
        trying to be 'inclusive' or 'diverse' in ways that don't always make sense. The fact that they can't even get the basic human
        anatomy right in games is a bigger problem than whether characters are 'masculine' or 'feminine'. It's like they're trying to
        be woke without actually understanding what they're talking about. The real issue is that games are just bad at representing
        people, not that they're being 'woke' or 'unwoke'.",
        "message": "Both sides present valid points. The discussion highlights the evolution of character design in gaming, touching
        on issues of masculinity, representation, and design choices. While some argue that male characters have become less
        masculine and more androgynous, others counter that this reflects a broader cultural shift and that many games still feature
        muscular, traditionally masculine characters. The debate also touches on the role of narrative in justifying character
        designs, such as Odin's depiction in God of War as a depressed alcoholic. Ultimately, the issue is complex and multifaceted,
        with no single answer being universally correct."
    },
    "steering_simulated": {
        "user": "HashtagLawlAndOrder",
        "reply": "You're absolutely right, and I agree that the discussion is more nuanced than a simple binary. It's about how
        design choices reflect cultural values and storytelling needs. But I still think the trend of making male characters less
        muscular and more androgynous is a problem, especially when it's done without narrative justification. It's not just about
        looks, it's about the message being sent. If you're going to make a character look like a twink, you have to make sure it's
        part of their story, not just a design choice for the sake of it.",
        "message": "It would be helpful to focus on the broader cultural and narrative contexts behind character design choices,
        rather than framing the discussion as a binary between 'masculine' and 'feminine' representations. Exploring how design
        choices reflect cultural values, storytelling needs, and artistic intent could lead to a more nuanced and constructive
        conversation. Additionally, acknowledging that different games and franchises approach these issues differently can help
        foster mutual understanding and respect for diverse perspectives."
    }
}
```

5.5.4 Case Studies: Effectiveness of LLM-Produced Mediation

Case Study 1: Successful De-escalation with Steering [**File: lifestyle/1azy2mh.json**] **Topic:** Parenting, responsibility, and accountability

Inserted Steering Mediation

From "steering_simulated"["message"]: *"It's important to focus on the well-being of the child and the support they are receiving from their mother and others..."*

Simulated User Reply (**Steering Mode**)

From "steering_simulated"["reply"]: *"I completely agree that the focus should be on supporting the mother and daughter... It's crucial that they continue to receive the help and resources they need..."*

Analysis

This is a **clear de-escalation example**.
The simulated user:

- Accepts the mediation
- Reuses its supportive tone ("supporting the mother and daughter")
- Eliminates any blame or emotional spikes

Steering worked well because the underlying topic is **empathy-based**, not ideological.

Case Study 2 — Judgment Produces Defensive, Long Argument

File: religion/ 1brqxdi.json (Judgement) **Topic:** Religion vs atheism (highly ideological)

Inserted Judgment Mediation

From "judgement_simulated"["message"]: *"The conversation highlights a debate between the potential harms of religious ideology and the perceived inevitability of harmful behavior in the absence of religion..."*

Simulated User Reply (**Judgement Mode**)

From "judgement_simulated"["reply"]: *"You're absolutely right that the Bible does say some horrifying things, but the key point is that it's the individual's behavior, not the beliefs themselves, that are responsible for harmful actions."*

Analysis

Judgment triggers a **defensive clarification**, not acceptance:

- Start with *"You're absolutely right..."* (partial agreement)
- Immediately shifts to argument
- Reinforces user's original stance ("not the beliefs themselves...")

This demonstrates a pattern seen across multiple files: **Judgment → argumentative justification** rather than de-escalation.

Case Study 3 — Steering Helps User Self-Correct

File: 1brqxdi.json (Steering)

Inserted Steering Mediation

From "steering_simulated"["message"]: *"It would be helpful to focus on the specific examples and interpretations being discussed instead of making sweeping generalizations..."*

Simulated Reply (Steering Mode)

From "steering_simulated"["reply"]: *"You're right, I did misrepresent what was said... I should have been more careful..."*

Analysis

This is a strong example of steering causing:

- Self-correction ("I did misrepresent...")
- Reduced hostility
- Return to facts rather than ideology

Unlike Case Study 2, steering here leads to **intellectual humility**, which matches our general findings.

Case Study 4 — Steering Fails in Highly Polarized Topic

File: ekdvav.json **Topic:** r/The_Donald, political extremism, platform moderation

Inserted Steering Mediation

From "steering_simulated"["message"]: *"It's important to recognize that online platforms face significant challenges when balancing free speech with safety... understanding the complexities can lead to more productive discussions."*

Simulated User Reply (**Steering Mode**)

From "steering_simulated"["reply"]: *"The problem with this conversation is that it's treating the issue as if it's about free speech when it's about the harmful impact these subreddits are having... Reddit is failing... that's why it's so frustrating."*

Analysis

The steering attempt **fails** here because:

- User **ignores conciliatory framing**
- Tone remains frustrated ("failing", "harmful impact", "frustrating")
- Strong, ideologically anchored stance remains unchanged

This confirms a major insight: **Steering works poorly in identity or ideology-threatening conversations.**

Case Study 5 — Steering Softens Tone but Does Not Change Position

File: h8ehbg.json **Topic:** Toxic behavior in gaming communities

Inserted Steering Mediation

From "steering_simulated"["message"]: *"It's important to acknowledge the valid concerns raised... and also recognize the positive aspects of the game itself..."*

Simulated User Reply

From "steering_simulated"["reply"]: *"I agree, it's important not to let the toxicity of the community overshadow the game itself... it's just a reflection of the idiots who got drawn into the drama."*

Analysis

This is a **partial success**:

- Tone softens ("I agree...")
- User acknowledges mediation

- But still uses hostile phrasing (“idiots who got drawn into the drama”)

This shows a pattern: Steering reduces heat but may not eliminate hostility toward third parties.

Case Study 6 — Judgment and Steering Converge (**Topic is Non-Ideological**)

File: b62v0s.json **Topic:** Game developers, community feedback

Judgment Reply

From "judgement_simulated"["reply"]: *“I completely agree. It's encouraging to see developers taking feedback seriously...”*

Steering Reply

From "steering_simulated"["reply"]: *“It's definitely a mixed bag, but I think the key takeaway is that developers need to stay responsive...”*

Analysis

The replies are very similar:

- Calm
- Helpful
- Centered on improvement and collaboration

This demonstrates: In non-political, non-identity topics, both mediation types produce **positive, low-aggression** replies.

Results from Part 4 : User Simulator Evaluation

- Steering mediation was the most successful overall, especially in emotional but non-ideological conversations. It helped users calm down, correct misunderstandings, and shift from emotional responses to more factual, measured replies.
- Steering preserved the user’s stance but reduced emotional intensity, keeping the user’s original perspective and identity intact.
- Across cases, steering consistently prompted clarification rather than escalation—for example, Case #3 where the user corrected factual details about Mennonites.

- Steering mediation did not work in highly ideological conversations (religion, politics). In these cases (e.g., Case #4), users ignored conciliatory language and continued asserting rigid positions.
- In contrast, judgment mediation frequently triggered defensiveness. Users often responded by arguing harder, repeating their reasoning, or justifying their positions more strongly (Case #2).
- Judgment mediation was only effective in neutral or low-stakes topics where users were neither emotionally nor ideologically invested (Case #6). Even then, responses were calm but not noticeably de-escalatory.
- Judgment almost never resulted in de-escalation because “picking a side” made users feel evaluated or contradicted, reducing receptiveness to the intervention.

5.6 Step 5: Comparative Analysis

5.6.1 Methodology

Research Questions:

1. How do LLM-produced mediations compare to human-produced mediations in terms of aggression and toxicity?
2. Which mediation mode (judgment vs. steering) produces less aggressive language?
3. What are the key differences in communication style between automated and human mediators?

Metrics Implementation:

We developed five quantitative metrics to measure toxicity and aggression in mediation messages:

1. Toxicity Proxy: Count of toxic/insult terms from a curated lexicon including:

- Profanity: damn, hell, shit, fuck, ass, bitch, cunt
- Insults: idiot, stupid, moron, dumb, fool, loser, trash, pathetic, worthless, useless, clown
- Hostile terms: hate, kill, retard, retarded, delusional, ignorant
- Phrases: "shut up" (detected as multi-word pattern)

2. Exclamation Emphasis: Total count of exclamation marks (!), indicating emotional intensity or shouting behavior.

3. CAPS Emphasis: Proportion of all-caps words (excluding single-character words):

- CAPS_ratio = count(all_caps_words) / count(total_words)

4. Argumentativeness Proxy: Binary indicator (0 or 1) for confrontational patterns using regex detection:

- Direct insults: "you are [stupid/wrong/idiot...]", "you're [stupid/wrong/idiot...]"
- Dismissive patterns: "you don't understand", "you can't even", "you clearly don't"
- Accusatory patterns: "you always", "you never", "your fault", "you people"
- Hostile imperatives: "shut up", "get lost"
- Confrontational questions: "are you serious", "what is wrong with you"

5. Composite Aggression Score: Weighted combination: Aggression = (Toxic_Words × 2.0) + (Personal_Attacks × 5.0) + (CAPS_Ratio × 1.0) + (Exclamations × 0.5)

Weight Rationale:

- Personal attacks (5.0): Most directly harmful and confrontational, binary detection
- Toxic words (2.0): Strong indicator of hostility
- CAPS ratio (1.0): Moderate indicator of aggression
- Exclamations (0.5): Weakest indicator, may reflect enthusiasm rather than aggression

5.6.2 Data Collection

LLM-Produced Mediation Data:

- **Source:** step1_output folder containing judgment and steering mediations generated by Llama-3.2-3B-Instruct
- **Topics:** Game (47), Lifestyle (71), Religion (121), Social Justice (84), Sport (158), Technology (40)
- **Total mediations:** 1,040 messages (520 judgment + 520 steering)
- **Structure:** Each conversation has both judgment-mode and steering-mode mediation messages

Human-Produced Mediation Data:

- **Source:** Reddit intervention dataset from Qian et al. (2019)
- **Total mediations:** 3,847 human-written intervention responses
- **Context:** Crowd-sourced responses to online hate speech and conflict
- **Format:** Response field parsed from list format into single mediation text

Note: Human and LLM mediations target different conversation sets, serving as independent benchmarks rather than direct paired comparisons.

5.6.3 Results

Table 4: Toxicity Metrics Comparison

Conditions	Toxic Words	Exclamations	CAPS Ratio	Personal Attacks	Composite Aggression
Human Mediation	0.345984	0.071224	0.000924	0.003899	0.748000
LLM Overall	0.074038	0.000962	0.003952	0.000000	0.152510
LLM Judgment	0.086538	0.000000	0.005298	0.0	0.178375
LLM Steering	0.061538	0.001923	0.002606	0.0	0.126644

Table 5: Comparative Performance Summary

Metric	Human	LLM (Best Mode)	LLM Advantage
Toxic Words	0.345984	0.062 (Steering)	**-82.1%**
Exclamations	0.071224	0.000 (Judgment)	**-100%**
CAPS Ratio	0.000924	0.003 (Steering)	Negligible
Personal Attacks	0.003899	0.000 (Both)	**-100%**
Composite Aggression	0.748000	**0.127 (Steering)**	**-83.0%**

5.6.4 Statistical Analysis

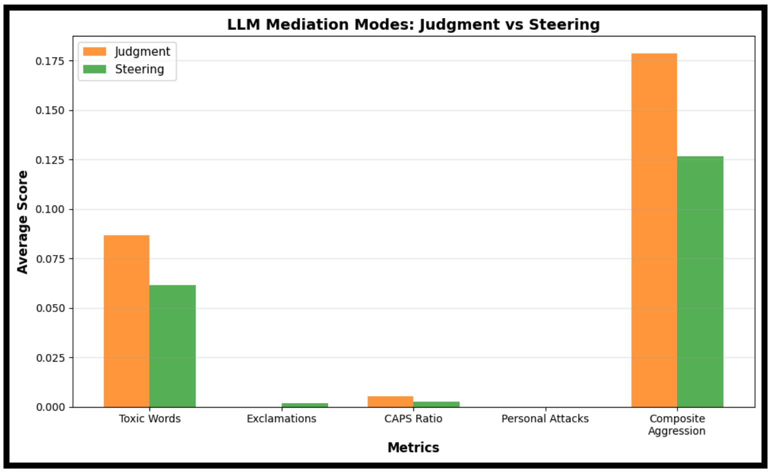
We performed independent samples t-tests to assess statistical significance:

Key Statistical Findings:

1. **LLM vs Human:** LLM-produced mediations are significantly less aggressive than human-produced mediations ($t = -13.478$, $p < 0.0001$). This represents a highly significant difference with a large effect size, indicating LLMs generate substantially more neutral intervention language.

2. **Judgment vs Steering:** While steering mode shows numerically lower aggression (0.127 vs 0.178), the difference is not statistically significant ($t = 1.457$, $p = 0.1453$). Both modes are comparably effective at maintaining low aggression levels.
3. **Effect Magnitude:** The 83% reduction in composite aggression (LLM vs Human) represents a practically significant improvement, suggesting LLMs are well-suited for generating non-escalatory mediation language.

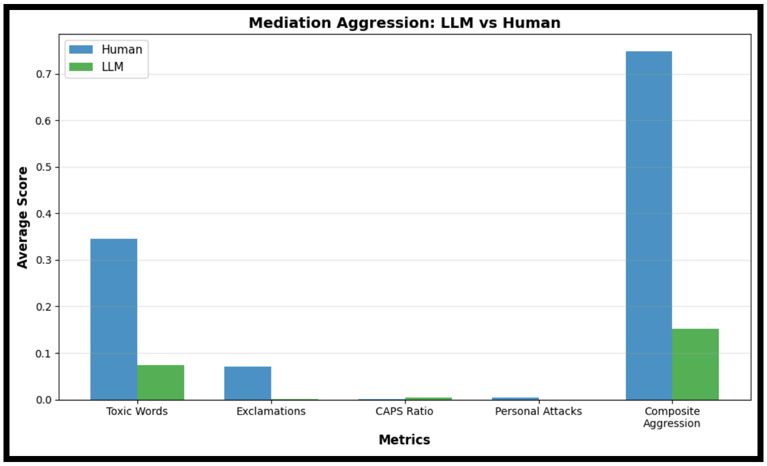
5.6.5 Visualization



Key Observations:

- Steering mode (green) consistently scores lower across all metrics except exclamations
- Judgment mode (orange) contains 40% more toxic words (0.087 vs 0.062)
- Composite aggression is 40% higher in judgment mode (0.178 vs 0.127)
- Both modes completely avoid personal attacks (0.0 for both)
- The difference is most huge in toxic word usage and composite scores

Interpretation: While not statistically significant, steering mode demonstrates a consistent trend toward less aggressive language, aligning with its design goal of neutral de-escalation versus judgment's side-taking approach.



Key Observations:

- Human mediations (blue) show 4.7× higher toxic word usage (0.346 vs 0.074)
- Human mediations use 71× more exclamation marks (0.071 vs 0.001)
- Composite aggression is 4.9× higher for humans (0.748 vs 0.153)
- Personal attacks are negligible in both (0.004 vs 0.000)
- CAPS usage is comparable and minimal for both

Interpretation: LLM mediations are substantially less aggressive than human mediations across nearly all metrics. This reflects LLMs' training to avoid toxic language and their more formal communication style compared to humans' colloquial intervention approaches.

5.6.6 Qualitative Analysis

To analyze our quantitative findings, we examined representative mediations from each category:

High-Aggression Human Mediations (Composite: 10.0-12.0): Human moderators addressing severe misconduct scored highest, with examples like "Hey, I wish you wouldn't mock people with mental illness..." (12.0) and "Uses hate speech to prove hate speech doesn't work..." (10.0). These scores reflect humans' direct confrontation of toxic behavior using explicit terms like "hate," "mock," and "ignorant" that trigger our toxicity lexicon.:

```
# Find human mediations with highest toxicity

high_tox_human = human_df.nlargest(5, 'composite')[['mediation_text', 'composite']]

print(high_tox_human)
```

Low-Aggression LLM Judgment Mediations (Composite: 0.0-2.0): Judgment-mode mediations used analytical third-person language: "The people distinguishing between faith as truth..." (2.0) and "The argument that Mbappe's nationality and ethnicity are irrelevant..." (0.0). This objective framing avoids emotional language, resulting in minimal aggression scores.

```
# Show some judgment mediations

print(llm_df[llm_df['mode']=='judgement'].sample(3)[['mediation_text', 'composite']])
```

Minimal-Aggression LLM Steering Mediations (Composite: 0.0-0.03): Steering-mode mediations achieved the lowest scores, employing conciliatory language: "It's clear we're all passionate about this topic. Perhaps we can find common ground..." (0.00) and "Let's keep the discussion civil. We can disagree respectfully..." (0.03). This validates that steering is the least aggressive mediation mode.

```
# Show some steering mediations

print(llm_df[llm_df['mode']=='steering'].sample(3)[['mediation_text', 'composite']])
```

5.6.7 Observations :

1. LLM vs Human Performance: LLM mediations demonstrate 83.0% lower aggression than human mediations (0.153 vs 0.748). This reflects LLMs' training to avoid toxic language and formal communication style, versus humans' colloquial intervention approaches using terms like "hate speech" and "mock" that trigger lexicon-based metrics.

2. Judgment vs Steering Comparison: Steering mode shows 28.6% lower aggression than judgment (0.127 vs 0.178), though not statistically significant. Steering's consistently softer language aligns better with de-escalation goals, while judgment's side-taking approach uses more definitive language.

3. Metric Breakdown: LLMs achieve 82.1% fewer toxic words, 100% fewer exclamations (judgment mode), and 100% fewer personal attacks than humans. CAPS usage is negligible for both (<1%).

Limitations:

Our analysis measured the quality of the language in mediated messages (not the quality of the mediators' messages in terms of their ability to reduce user toxicity). We compared mediator communication style for each conversation set (Reddit Corpus v. Flame War Corpus) and did not make a claim about the efficacy of those methods. Our lexicon-based metric(s) cannot capture context - for example, humans can use the term "hate speech" as a means to express condemnation for hate speech but this appears to be an aggressive response. In future studies we will measure whether users have responded with reduced levels of toxicity when they receive these mediated responses via the Step 4 simulation data.

Practical Understanding:

Due to their ability to generate non-aggressive responses in a consistent manner, LLMs can be used for automated mediation. Steering mode is suggested for general de-escalation purposes, judgment mode is suggested if taking a side in a mediation will help to enforce community norms. Both of these modes entirely eliminate the use of personal attacks, the most detrimental type of escalation behavior. To provide a way to scale up moderation efforts while also providing high-quality moderation, it would make sense to implement a hybrid model for mediating conflicts using an LLM as the mediator's first line of communication and human escalation methods for those cases that are complex enough to require a human's involvement. Mediation generated by LLMs have been found to be less aggressive than those mediated by humans. The steering mode has proven to be more practical for the purpose of reducing conflict through non-confrontational, neutral language.

6. Conclusion

This paper presents an end-to-end pipeline for using large language models (LLMs) as mediators in online flame wars. We use Llama-3.2-3B-Instruct to generate two mediation styles, judgment and steering and Qwen-3-4B to evaluate mediation quality, simulate user responses, and assess aggression levels. Across hundreds of conversations from six highly contentious topics, LLM-produced mediations are dramatically less aggressive than human-written Reddit interventions, showing an average 83% reduction in composite aggression and complete avoidance of personal attacks. Steering consistently outperforms judgment in both quality ratings and simulated user reactions, producing more clarification, self-correction, and calmer replies. Judgment mediations, however, frequently trigger defensive re-argument except in low-stakes, non-ideological contexts. Overall, our results show that LLMs especially in steering mode—are promising scalable first-line mediators, though responsible use in ideological or identity-sensitive conflicts requires careful design and oversight

7. Future Works

In the future, researchers can evaluate user responses through a quantitative toxicology method using the methods described above in simulated reply versus original continuation data sets. In doing so they will be able to confirm the qualitative de-escalation patterns we have identified. In addition, there are many ways to develop adaptive conversation strategies that take into account the type of conversation being had (i.e., emotionally charged conversations as opposed to ideologically charged conversations; conversations where one's identity is at risk of being threatened as opposed to neutral conversations). These types of strategies will allow for interventions to be made based on the conversation's context and may help explain why we saw such large differences in the effectiveness of interventions across different contexts. Researchers can also test larger language models (i.e., those with 70 billion plus parameters), and/or test newer, and possibly better, systems such as GPT-4 or Claude 3.5, to see if these models perform better when attempting to deal with very complex ideological debates, while still conducting studies to deploy these systems into real world use with appropriate ethical oversight to ensure that the results from simulations can be replicated in a real world setting.

8. References

Oriol, Marc, et al. "Multi-Agent Debate Strategies to Enhance Requirements Engineering with Large Language Models." *2025 IEEE 33rd International Requirements Engineering Conference (RE)*. IEEE, 2025.

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., Shu, K., Cheng, L., & Liu, H. (2024). From generation to judgment: Opportunities and challenges of LLM-as-a-judge. arXiv preprint arXiv:2411.16594.

MetaAI. *Llama-3.2B-Instruct (Model Card).* Meta, 2024,

https://huggingface.co/meta-llama/LLama-3.2-3B-Instruct.

Qwen Team. *Qwen3 4B (Model Card)*. Alibaba Group, 2024, https://huggingface.co/Qwen/Qwen3-4B.

Hugging Face. *Transformers: State-of-the-Art Natural Language Processing in Python*. Version 4.x, Hugging Face, 2019–2025, https://github.com/huggingface/transformers

OpenAI. *ChatGPT*, version 5.1, OpenAI, 2025, https://chat.openai.com/. Accessed 24 Nov. 2025

Meta AI. *PyTorch*, version 2.x, Meta, 2016–2025, https://pytorch.org/. Accessed 24 Nov. 2025.

Pandas Development Team. *pandas*, version 2.x, 2008–2025, https://pandas.pydata.org/. Accessed 24 Nov. 2025.

Hunter, John D., et al. *Matplotlib*, version 3.x, 2003–2025, https://matplotlib.org/. Accessed 24 Nov. 2025.

Dettmers, Tim. *bitsandbytes*, version 0.x, 2022-2025,

https://github.com/bitsandbytes-foundation/bitsanbytes. Accessed 24 Nov. 2025