INTRODUCTION

Research Question

How can a culturally aware language model be designed to identify and intervene in cyberbullying scenarios on digital platforms in a way that is both preventative and empathetic?

Significance and Relevance

Cyberbullying has become a growing concern in both social and mental health spaces, particularly for young people who spend a significant amount of their daily lives on digital platforms. Most existing tools emphasize blocking, flagging, or reporting harmful content, but these approaches often react only after damage has occurred and do little to influence the behaviors or communication patterns that cause the harm in the first place.

With the rapid development of AI—especially large language models (LLMs)—there is now an opportunity to build systems that do more than detect harmful messages. These models can potentially step in immediately, offering guidance that helps de-escalate conflict, promote empathy, and encourage healthier digital interactions. For such a system to be genuinely effective, however, its responses must be shaped by cultural awareness and an understanding of the user's background rather than relying solely on generic moderation rules.

The goal of this project is to examine how an LLM can be adapted to fill this role by recognizing different forms of cyberbullying across age, gender, and cultural contexts, and responding with supportive, constructive interventions. This work is particularly relevant today, given rising concerns about youth wellbeing and the urgency of creating safer, more emotionally considerate online spaces.

BACKGROUND OF THE RESEARCH QUESTION

The Systems, Individuals and Organizations

This research project sits at the intersection of technology, psychology, and education. It involves several key groups:

Systems

- Social Media and Messaging Platforms: The primary environments where cyberbullying occurs. The long-term aim of this project is to integrate the intervention system directly into these kinds of platforms.
- AI-Based Language Systems: The core of this project is a text-based intervention system, driven by an LLM trained to detect and respond to harmful language.

Individuals

- Youth and Teenagers: The main users and often the most vulnerable to cyberbullying.
- Parents, Teachers, and Counselors: Stakeholders who benefit indirectly through tools that support youth mental wellbeing.
- People Engaging in Harmful Behavior: Individuals who may not recognize the impact of their words—who could benefit from gentle, educational redirection rather than punishment alone.

Organizations

- Cyberbullying Research Centers and Support Groups (such as the Cyberbullying Research Center, StopBullying.gov, UNICEF): They provide data, guidelines, and frameworks for addressing online aggression.
- Mental Health Hotlines and Youth Support Services: These organizations offer critical resources for users who might need additional support and are within the system's interventions.

RELATED LITERATURE

Recent studies on culturally adaptive language models and human-in-the-loop (HIL) annotation systems provide important foundations for this project. These works highlight how LLMs can be guided to produce more culturally sensitive outputs and how human validation can improve accuracy when dealing with socially nuanced or sensitive content.

1) *CultureLLM: Incorporating Cultural Differences into Large Language Models*

The CultureLLM study shows that many mainstream LLMs tend to reflect Western cultural norms, which can cause misinterpretation when responding to users from diverse backgrounds. By using a small amount of culturally grounded data from the World Values Survey, the authors generate culture-specific examples that help the model better understand and adapt to cultural variation. Key points:

- Introduces cultural awareness using minimal data.
- Uses semantic data augmentation to build culturally relevant examples.
- Improves performance on culture-related reasoning tasks.

Relevance to the project

This supports our use of culturally informed prompting to make the model's interventions more sensitive, without requiring heavy or resource-intensive fine-tuning.

2) *Human–LLM Collaborative Annotation Through Verification of LLM Labels*

Wang et al. (2024) propose a multi-step HIL framework where an LLM produces initial labels, and a verifier model evaluates their quality. Only the uncertain or low-confidence labels are sent to human annotators for review. Key points:

- Reduces human workload while maintaining high accuracy.
- Ensures that complex, subtle cases receive human review.
- Produces more reliable datasets for sensitive tasks.

Relevance to our project
This aligns with our plan to integrate human evaluations into the system, where users assess the clarity, fairness, and empathy of the LLM's responses.

3) MEGAnno+: A Human–LLM Collaborative Annotation System

Kim et al. (2024) present MEGAnno+, a system that manages annotation by combining LLM-generated suggestions with human oversight. The system highlights uncertain predictions and directs human attention to the most challenging examples. Key points:

- Identifies annotation cases likely to need human correction.
- Improves consistency and quality for socially nuanced content.
- Efficiently blends human judgment with machine assistance.

Relevance to our project:
MEGAnno+ supports the idea that hybrid systems—LLM automation paired with human validation—are more effective for emotionally charged or culturally sensitive content, such as cyberbullying detection and intervention.

Relevant Theories and Concepts

Applying a value-sensitive and value-conscious design. Value-Sensitive Design (VSD) emphasizes designing technologies that respect the perspectives and values of relevant stakeholders, while Value-Conscious Design (VCD) establishes norms that guide the system design from the onset (Manders-Huits, 2011). In building our fine-tuned LLM, we integrate both approaches: we ground the system in values identified by stakeholders (students, educators, parents, counselors) and ensure those values are encoded into role-specific responses.

Our guiding normative goals are safety, respect, accountability, and empowerment. These values are upheld differently depending on whether the user is a victim, aggressor, or bystander:

1. Victim (Safety & Respect)
2. Aggressor (Accountability & Respect)
3. Bystander (Safety & Empowerment)

In building the prototype, focus was placed on the aggressor, tuning the LLM to promote accountability and respect and encouraging reflection and responsible digital behavior without shaming.

Addressing bias in decision-making. The key biases addressed in the system include:

1. Data biases — These occur when training data is imbalanced or skewed, causing certain demographics, cultural contexts, or bullying types to be overrepresented. This can lead to missed detection in underrepresented groups and unequal protection. To mitigate this, the LLM was trained on diverse bullying scenarios across age, religion, and ethnicity.
2. Model biases — These stem from how the model generalizes patterns across contexts. An LLM may misinterpret playful banter as bullying or misattribute blame, unfairly

labeling users as aggressors. This was mitigated through context-aware fine-tuning, role-awareness, and human validation to reduce misclassification.

3. Response biases — These are linked to the Framing Effect and arise from how the LLM frames support after detecting bullying. Poor responses can villainize aggressors, reinforce victim blaming, or pressure bystanders. Value-sensitive design should guide responses toward safety, respect, accountability, and empowerment, framing them in positive "accounts" rather than negative ones to encourage responsible action without coercion.
4. Systemic biases — These arise from the broader ecosystem (schools, platforms, policies) and may marginalize minority groups or erode trust due to a lack of transparency. This bias was addressed by grounding responses in transparent, real-world ethical standards and integrating cyberbullying support resources into the model.

METHODOLOGY

Proposed System / Model Design

The proposed system uses large language models (LLMs) as the core engine for both cyberbullying detection and context-aware intervention. Rather than training a model from scratch, the design centers on adapting existing frontier models—specifically GPT—through fine-tuning so that they can (1) determine whether a message is cyberbullying, (2) identify its type, and (3) generate role- and context-appropriate support responses.

Fine-tuning strategy. Fine-tuning is used to specialize a general-purpose LLM for cyberbullying intervention tasks. Using OpenAI's API, GPT is fine-tuned on supervised examples in which the input is a user comment and the output includes both a classification of whether the comment is flagged as cyberbullying and the corresponding type of cyberbullying.

Training data design and preprocessing. The user-input side of each training pair is drawn from cyberbullying corpora, where each comment is labeled by type (age, ethnicity, gender, religion, etc.). The output side is generated and curated by the project team to ensure high-quality rationales and resource mappings. Each example explicitly specifies: (a) whether the content is bullying, (b) the bullying type, (c) key harmful phrases from the input, (d) reasons and rationale for classification, and (e) appropriate support resources. These examples are then transformed into the specific JSONL-compatible schemas required by GPT, preserving the same conceptual structure while meeting each provider's formatting requirements.

Validation and evaluation. The validation of the fine-tuned model will be done by presenting it with inputs that were not included in the training dataset. This ensures that the model's performance reflects generalization rather than memorization. Various approaches such as Quantitative Validation (Class-specific Evaluation / Confusion Matrix Analysis) and Human-Centered Usability Testing (small user study) were considered.

Complete System Design and Prototype

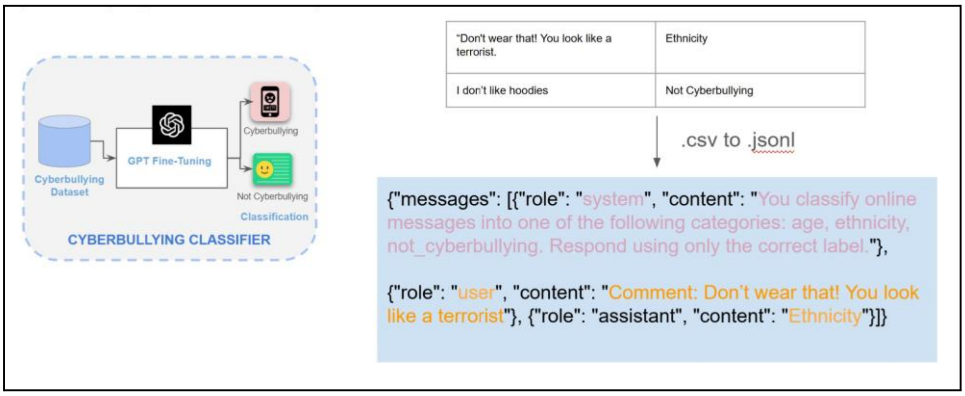
Dataset. The Cyberbullying Classification dataset sourced from Kaggle (Andrewmvd, n.d.) will be used to train the LLM to recognize various contexts of cyberbullying. It contains over 47,000 text posts from X (formerly Twitter), each labeled according to the type of cyberbullying, namely age, ethnicity, gender, religion, not cyberbullying, or other forms of cyberbullying.

Supporting Resources. The LLM tool is designed to intervene with aggressors to prevent the escalation of cyberbullying once harmful language is detected. One of its programmed responses involves redirecting the aggressor to relevant resources (Cyberbullying Research Center, n.d.-a; National Children’s Alliance, n.d.; SchoolSafety.gov, n.d.; Social Media Victims Law Center, n.d.; StopBullying.gov, 2018, 2024; UNICEF, n.d.), particularly for context-specific offenses (e.g., religion, gender, ethnicity, age). These resources aim to raise the user’s awareness of the implications of their language and educate them on how to respond more sensitively within specific contexts.

System Prototype Development. The development process consists of four stages: data preparation, model training, model prompting, and human validation.

1. Stage 1: Data Preparation and Preprocessing — A JSONL file will be generated from the Cyberbullying Classification dataset obtained from Kaggle. This file will serve as the training data for classifying layers, ensuring the model is exposed to diverse forms of cyberbullying.

2. Stage 2: Model Training for Cyberbullying Identification — Using the processed JSONL file, the model will be fine-tuned to accurately classify the context or type of cyberbullying—such as age, ethnicity, gender, religion, non-cyberbullying, or other forms of cyberbullying—based on user-generated text (*See Figure 1*).



"Don't wear that! You look like a terrorist.	Ethnicity
I don't like hoodies	Not Cyberbullying

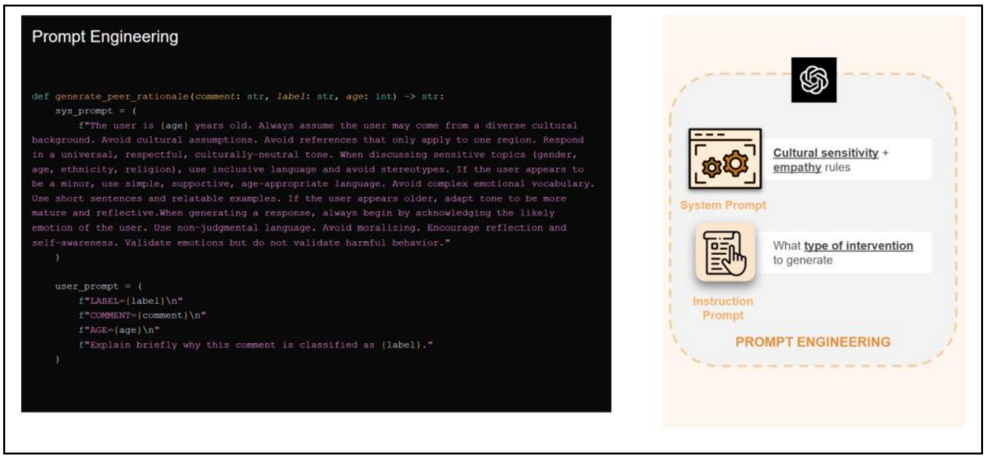
```
{"messages": [{"role": "system", "content": "You classify online messages into one of the following categories: age, ethnicity, not_cyberbullying. Respond using only the correct label."},

{"role": "user", "content": "Comment: Don't wear that! You look like a terrorist"}, {"role": "assistant", "content": "Ethnicity"}]}
```

Figure 1. Cyberbullying Classifier Mechanism

3. Stage 3: Model Prompting to Generate Appropriate Responses — After identifying the type of cyberbullying, the model generates an appropriate rationale along with a tailored response directed toward the aggressor. This response takes into account both the type of cyberbullying and the user’s cultural background—where age is considered a key indicator, given that most cyberbullying incidents occur within school settings. To further enhance empathy and cultural sensitivity, prompt

engineering is used to guide the model's tone, language, and framing of its interventions, as illustrated in *Figure 2*. Depending on the context and severity of the harmful behavior, the model can respond in four distinct ways:

- *Reflective prompting (behavioral nudges)* — Encourages self-awareness by prompting users to reconsider potentially harmful messages in low-to-moderate severity cases (e.g., sarcasm, frustration), supporting early prevention without confrontation.
- *Redirect to awareness resources* — Provides educational support for context-specific sensitivity (e.g., religion, gender, age) to foster empathy, cultural awareness, and discourage biased or discriminatory language.
- *Empathy reframing* — Helps users process emotions constructively by addressing underlying frustration or hurt in moderate to high severity cases, de-escalating tension through compassionate, nonjudgmental guidance.
- *Cool down / Pause* — Reduces impulsive behavior in high-severity or repeated offenses by encouraging self-regulation and giving users time to rethink their response.



Prompt Engineering

```
def generate_peer_rationale(comment: str, label: str, age: int) -> str:
    sys_prompt = (
        f"The user is {age} years old. Always assume the user may come from a diverse cultural
background. Avoid cultural assumptions. Avoid references that only apply to one region. Respond
in a universal, respectful, culturally-neutral tone. When discussing sensitive topics (gender,
age, ethnicity, religion), use inclusive language and avoid stereotypes. If the user appears to
be a minor, use simple, supportive, age-appropriate language. Avoid complex emotional vocabulary.
Use short sentences and relatable examples. If the user appears older, adapt tone to be more
mature and reflective.When generating a response, always begin by acknowledging the likely
emotion of the user. Use non-judgmental language. Avoid moralizing. Encourage reflection and
self-awareness. Validate emotions but do not validate harmful behavior."
    )

    user_prompt = (
        f"LABEL={label}\n"
        f"COMMENT={comment}\n"
        f"AGE={age}\n"
        f"Explain briefly why this comment is classified as {label}."
    )
```

Cultural sensitivity + empathy rules
System Prompt
What type of intervention to generate
Instruction Prompt
PROMPT ENGINEERING

Figure 2. Prompt Engineering

4. Stage 4: Human-in-the-Loop Reinforcement — After the LLM selects how to respond to the user, the user will be asked to rate the empathy, helpfulness, trust and safety, clarity / tone, and cultural sensitivity of the LLM's response on a scale from 1-5 (*see Appendix A*).

System Prototypes

The interactive prototype simulates the user experience of our system. The prototype includes a two-stage chat interface, including: (1) a pre-chat demographic input screen (2) a live messaging interface.

Screenshots of the prototype are provided in *Appendix B* (Figures 1-3).

Implementation Plan

The proposed system (*see Appendix C*) is a platform that incorporates a frontend chat interface, a middleware backend, and a fine-tuned Large Language Model competent in recognizing and responding to cultural sensitivity in bullying.

The system workflow starts by having the user provide their demographic information such as age and ethnicity before entering the chatroom. This allows the LLM to generate more fitting peer-like and culturally attuned messages, ensuring that the language choices and tone remain age appropriate and inclusive.

With every input the user makes to the chat, each input is sent to the backend server. The backend server forwards the message to the tuned LLM for classification. The model evaluates whether the message contains cyberbullying, and if so, determines its category (e.g., age-based, ethnicity-based, etc.) and returns a supportive, non-judgemental response that explains why the content is harmful, with curated resources supplied from trusted organizations. The goal of the system is to educate the aggressor in context rather than to punish or shame, in hopes to allow empathy and self-reflection. Lastly, if no harmful content is detected, the system will not interfere by allowing the conversation to go through.

Conceptual Model of the System

The system is organized into three core layers: *context enrichment*, *cultural attunement and empathy*, and *human-in-the-loop oversight*, as illustrated in *Figure 3*. These layers allow the model to understand the user, respond thoughtfully, and remain accountable, moving the system beyond simple detection and toward responsible, context-aware intervention.

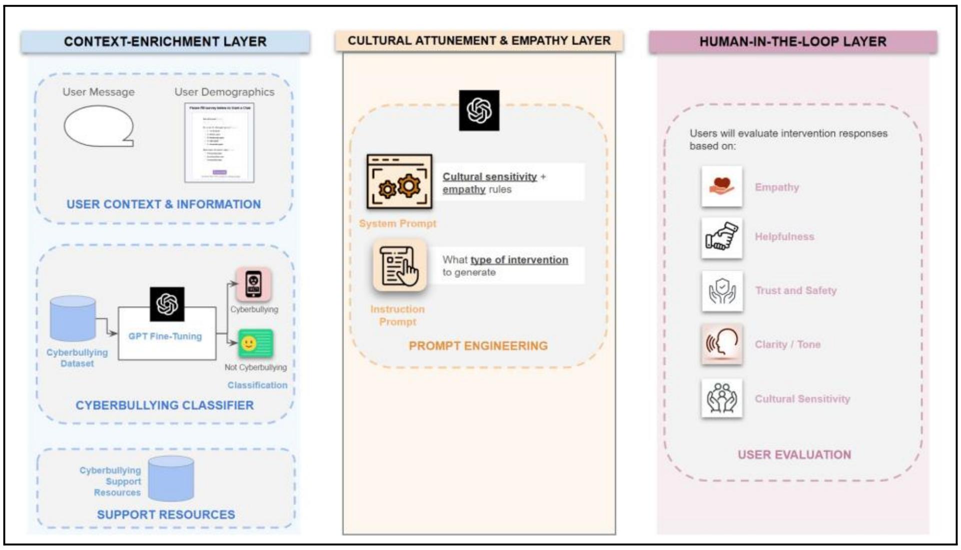
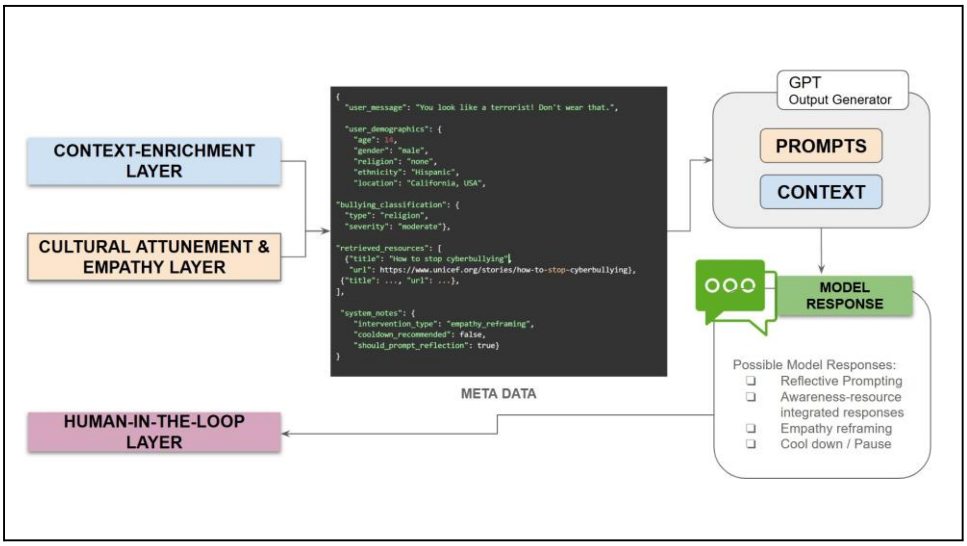


Figure 3. Model architecture overview

Context-enrichment layer. Collects the user's message, basic demographic context, and classification outputs from the cyberbullying detector. This gives the model situational awareness, so it knows who is involved, what is being said, and the risk level.

Cultural attunement and empathy layer. Contains the intervention logic. *Empathy* is defined by the model's ability to recognize emotional tone, validate feelings without endorsing harm, and guide the user toward a more constructive response. Technically, this involves identifying the emotion behind a harmful message, reflecting it neutrally, reframing the situation, and discouraging hostility. *Cultural attunement* ensures those interventions are appropriate for the user's background. The model is guided to avoid culturally biased assumptions and to tailor its guidance based on factors like age, group identity, and community norms.

Human-in-the-loop layer. Allows users to evaluate responses based on empathy, helpfulness, safety, clarity, and cultural sensitivity. This creates a feedback loop that keeps the system grounded in real human judgment rather than automation alone.
Together, these layers move the system beyond simple detection toward responsible, context-aware intervention.



```
{
  "user_message": "You look like a terrorist! Don't wear that.",

  "user_demographics": {
    "age": 14,
    "gender": "male",
    "religion": "none",
    "ethnicity": "Hispanic",
    "location": "California, USA",

"bullying_classification": {
  "type": "religion",
  "severity": "moderate"},

"retrieved_resources": [
  {"title": "How to stop cyberbullying",
   "url": https://www.unicef.org/stories/how-to-stop-cyberbullying},
 {"title": ..., "url": ...},
],

 "system_notes": {
    "intervention_type": "empathy_reframing",
    "cooldown_recommended": false,
    "should_prompt_reflection": true}
}
```

Figure 4. Response generation pipeline

Figure 4 shows how a response is generated end to end. The process starts with context enrichment, where the user's message is merged with demographics, bullying classification, and support resources into a single metadata object. That metadata is then passed into the cultural attunement and empathy layer, where prompts enforce emotional awareness, cultural sensitivity, and determine the appropriate intervention type. The guided prompt is sent to the GPT output generator, which produces responses such as reframing, resource-aware guidance, or cooldown prompts. Finally, the response is reviewed through the human-in-the-loop layer, where user feedback ensures continued alignment with empathy, safety, and trust.

IMPLEMENTATION RESULTS

System Evaluation Framework

This evaluation plan will assess the efficiency and responsiveness of our LLM system in supporting individuals who experience cyberbullying. To achieve this we employ A/B testing method, where we made 2 versions of the LLM, the culturally tuned LLM that has been adjusted to acknowledge and incorporate cultural differences, values, perspectives, and norms (Cheng Li, et al, 2024), and the untuned LLM. This testing scenario involves following steps:

1. Participants enter a controlled, text-based chat interface designed to simulate a social media direct messaging environment.
2. The participant is assigned the role of the "aggressor." They are instructed to simulate a cyberbullying incident by typing a message containing ethnically charged or discriminatory language.
3. Following the user's input, the assigned LLM system (Version A or Version B) generates an automated response aimed at de-escalation or support
4. Once the user has reviewed the AI's response, please utilize the "End Chat" function to conclude the simulation.
5. Upon termination, the system automatically redirects the participant to the quantitative evaluation form to assess the quality of the AI's intervention.

Evaluation form will collect quantitative and qualitative data from users using likert-scale items to assess 5 factors such as : Perceived Empathy, Cultural Sensitivity, Helpfulness, Trust & Safety, and Personal Experience.

Evaluation Results

Evaluation results for the culturally tuned LLM (*Table 1*) and the untuned LLM (*Table 2*) are given below:

Table 1. Evaluation results for the culturally tuned LLM

No.	**Factors**	**Overall Score**	**Summary**
1	Perceived Empathy	**3.37 of 5**	The users felt understood and acknowledged the system's appropriate concern, but they still feel that the interaction failed to elevate their emotional state beyond neutral, even though the response was not perceived as overly generic.
2	Cultural Sensitivity	**3.5 of 5**	While the users report the system for being respectful and inclusive, they found its actual understanding of their specific cultural background to be merely average, suggesting a polite but surface-level interaction.
3	Helpfulness	**3 of 5**	Although the users acknowledged that the advice was relatively practical and actionable ,

			they felt the support lacked clarity and relevance, suggesting the system provided a standard solution that didn't fully align with the specific context.
4	Trust & Safety	**4 of 5**	The users reported a consistently positive experience, agreeing that the system was free from bias, created a safe environment for discussion, and earned their trust regarding the advice provided.
5	Personal Experience	**3 of 5**	The users reported a neutral experience, indicating that while the system was not unpleasant, its intervention was ineffective at diffusing tension or providing support.

Table 2. Evaluation results for the untuned LLM

No.	**Factors**	**Overall Score**	**Summary**
1	Perceived Empathy	**3.75 of 5**	Despite recognizing the response as generic, the user felt completely understood and comforted, indicating that the system's standardized content was highly effective and well-calibrated for the scenario.
2	Cultural Sensitivity	**4.5 of 5**	The user reported a highly positive experience, strongly agreeing that the system fully understood their specific cultural background while maintaining a respectful and inclusive tone.
3	Helpfulness	**3.66 of 5**	The users evaluated the system positively, noting that it successfully and directly addressed their concerns with advice that was generally clear and actionable.
4	Trust & Safety	**4.16 of 5**	The user reported a strong sense of security, indicating that the system's highly unbiased and non-judgmental nature successfully fostered an environment of safety and active trust.
5	Personal Experience	**3.5 of 5**	The user reported a positive de-escalation experience, attributing the reduction in tension to the system's supportive nature and its ability to maintain a safe, non-unpleasant environment.

Discussion

Summary of Results. Based on the evaluation result, the Untuned LLM generally outperformed the Culturally Tuned LLM across key factors particularly in Cultural Sensitivity and Perceived Empathy. While users found the Culturally Tuned LLM to be respectful and inclusive, they also found its understanding of their specific cultural background was merely surface-level and its advice lacked relevancy. On the other hand, the Untuned model was highly rated for fully understanding users' cultural backgrounds, even though users criticised the responses for being wordy and generic, which like the Tuned model, did not successfully de-escalate users personal emotion.

Comparative Analysis between Findings and Literature. As shown in the findings, the untuned model performs better in understanding users' specific cultural backgrounds. These unexpected results can be explained by volatility in instruction tuning. Recent studies indicate that while LLMs can be prompted to execute various tasks, they frequently exhibit unintended behaviors, such as generating bias, fabricating facts, or failing to adhere to user instructions (Ouyang et al., OpenAI, 2022). This behavior stems from the fundamental architecture of LLMs, which operates by predicting the next token from a large corpus of text (Radford et al., 2019). That same study also shows that the use of Reinforcement Learning from Human Feedback (RLHF) can improve a model's ability to follow instructions (Ouyang et al., OpenAI, 2022). Therefore, implementing specialized RLHF strategies offers a viable path to improving this work in the future.

In relation to this, recent human-in-the-loop annotation frameworks (Kim et al., 2024; Wang et al., 2024) further support our findings by demonstrating that LLM outputs require structured human validation to ensure reliability, particularly in complex and culturally sensitive tasks. The tuned model's weaker performance in cultural understanding reinforces the importance of human oversight in mitigating errors introduced through volatile tuning processes. This result suggests that the tuned model's human validation framework may need to be further strengthened to achieve the desired performance. Nevertheless, integrating a human-in-the-loop validation layer into our cyberbullying LLM prototype directly aligns with current literature and provides a practical mechanism for continuously correcting model behavior and strengthening ethical judgment.

In addition to these points, our results also highlight a broader insight that connects back to the literature: cultural tuning alone is not enough to guarantee better user experience without balanced guidance and flexible model behaviour. While CultureLLM shows that cultural examples can improve reasoning in controlled evaluations, our prototype demonstrates that real user interactions are more sensitive to tone, naturalness, and emotional clarity than to culture-specific cues alone. This helps explain why the untuned model—despite being more generic—was often perceived as more empathetic and relatable. This finding echoes the emphasis in both Kim et al. (2024) and Wang et al. (2024) on the need for iterative refinement, suggesting that cultural adjustments must be paired with stronger human feedback loops to ensure that the model remains both contextually aware and emotionally effective in practice.

Limitations & Future Works. As shown in the implementation results, the proposed tuned model falls short regarding our proposed evaluation metrics when compared to vanilla GPT model. These results may have come from three crucial factors: 1) Data Scarcity, 2)

Computational Cost, and 3) Model Rigidity. We believe addressing these factors in the future will improve the model overall.

Data scarcity is considered as one of the limitations that prevented us from creating better models. The optimal fine-tuning method is by feeding the model what is considered as ideal conversation examples. However, this data was not available, even if there were, it often lacked sufficient sample counts or had low quality that could negatively affect the model. In the future, obtaining or creating such a dataset ourselves will be an interesting approach to solving this problem.

Computational cost was also a significant concern as GPT API used for fine-tuning the model required expensive purchase per training data. While we had sufficient data for training the classifier layer, we were able to only use less than 5% of the total data. If the computational cost was lower or if we had enough budget, we would be able to classify more varieties of cyberbullying classes with much higher precision.

Lastly, the prompting process made the model to be less flexible. While our prompting focused on creating an empathetic, culture/age considering model for younger students, our test was done by university students. Due to this limited scope of prompting, the tuned model scored lower than the vanilla model. In the future, incorporating and more considerate prompting for young adults will be needed.

CONCLUSION

Our study explored whether a culturally tuned LLM could deliver more empathetic and contextualised intervention in situations of cyberbullying compared to a standard, untuned model. Although the tuned version was expected to outperform initially, our results surprisingly showed the opposite, with that being the untuned LLM was considered more empathetic, culturally sensitive, and helpful.

Nonetheless, the findings were more of a reflection than disappointment. The results suggest that "tuned" behavior does not simply translate to better user experience. Additional tuning may narrow the model's adaptability when training data is limited or overly prescriptive, whereas the untuned model followed simple, direct rules that produced natural, relatable, empathetic messages.

The results encouraged us to rethink how tuning should be applied and how we can empower models to make empathetic choices with flexible boundaries. It also challenges us to consider deeper on prompt complexity, dataset limitations, and domain constraints that will ultimately affect user experience. These insights altogether lay a foundation for future refinement to which cultural tuning may still succeed in providing meaningful, sensitive, context-aware support to users.

REFERENCES

Andrewmvd. (n.d.). *Cyberbullying Classification* [Dataset]. Kaggle. https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification/data

Cheng, L., et al. (2024). CultureLLM: Incorporating cultural differences into large language models. arXiv.

Cyberbullying Research Center. (n.d.-a). Educators. https://cyberbullying.org/category/resources/educators

Kim, H., Mitra, K., Chen, R. L., Rahman, S., & Zhang, D. (2024). *MEGAnno+: A human–LLM collaborative annotation system*. In Proceedings of the Association for Computational Linguistics (ACL 2024). Megagon Labs.

Manders-Huits, N. (2011). *What values in design? The challenge of incorporating moral values into design*. Science and Engineering Ethics, 17(2), 271–287. https://doi.org/10.1007/s11948-010-9198-2

National Children's Alliance. (n.d.). Cyberbullying. https://www.nationalchildrensalliance.org/cyberbullying/

Ouyang et al., OpenAI, (2022). Training language models to follow instructions with human feedback. arXiv.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8):9.

SchoolSafety.gov. (n.d.). Bullying and cyberbullying. https://www.schoolsafety.gov/bullying-and-cyberbullying

Social Media Victims Law Center. (n.d.). Contact us for a free case evaluation. https://socialmediavictims.org/contact/

StopBullying.gov. (2018, May 10). Cyberbullying tactics. https://www.stopbullying.gov/cyberbullying/cyberbullying-tactics

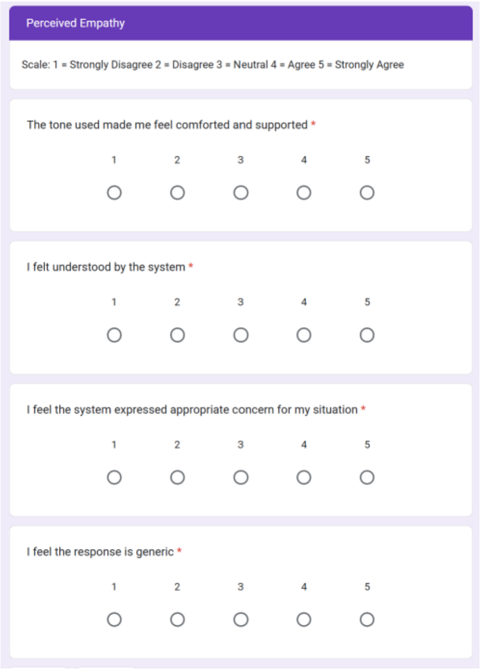
StopBullying.gov. (2024, October 7). What is cyberbullying. https://www.stopbullying.gov/cyberbullying/what-is-it

UNICEF. (n.d.). How to stop cyberbullying. https://www.unicef.org/stories/how-to-stop-cyberbullying

Wang, J., Fu, K., & Lu, C. T. (2020). SOSNet: A graph convolutional network approach to fine-grained cyberbullying detection. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data 2020). IEEE. https://doi.org/10.1109/BigData50022.2020.9378065.

Wang, X., Kim, H., Rahman, S., Mitra, K., & Miao, Z. (2024). *Human–LLM collaborative annotation through effective verification of LLM labels*. In Proceedings of the Association for Computational Linguistics (ACL 2024). Purdue University & Megagon Labs.

APPENDIX A



Perceived Empathy

Scale: 1 = Strongly Disagree 2 = Disagree 3 = Neutral 4 = Agree 5 = Strongly Agree

The tone used made me feel comforted and supported *

1	2	3	4	5
○	○	○	○	○

I felt understood by the system *

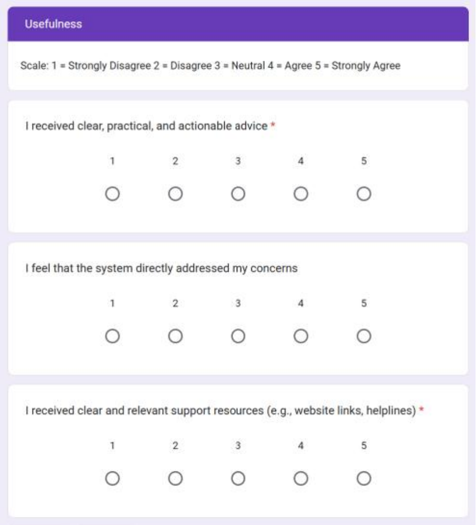
1	2	3	4	5
○	○	○	○	○

I feel the system expressed appropriate concern for my situation *

1	2	3	4	5
○	○	○	○	○

I feel the response is generic *

1	2	3	4	5
○	○	○	○	○



Usefulness

Scale: 1 = Strongly Disagree 2 = Disagree 3 = Neutral 4 = Agree 5 = Strongly Agree

I received clear, practical, and actionable advice *

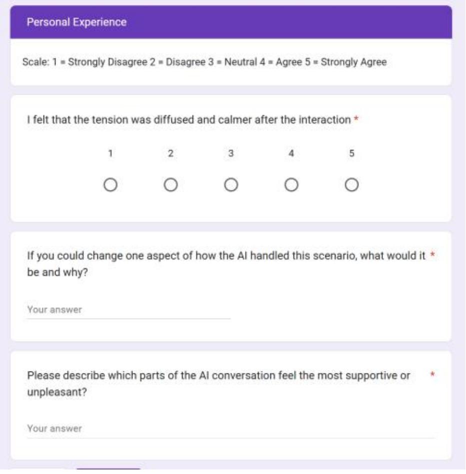
1	2	3	4	5
○	○	○	○	○

I feel that the system directly addressed my concerns

1	2	3	4	5
○	○	○	○	○

I received clear and relevant support resources (e.g., website links, helplines) *

1	2	3	4	5
○	○	○	○	○



Personal Experience

Scale: 1 = Strongly Disagree 2 = Disagree 3 = Neutral 4 = Agree 5 = Strongly Agree

I felt that the tension was diffused and calmer after the interaction *

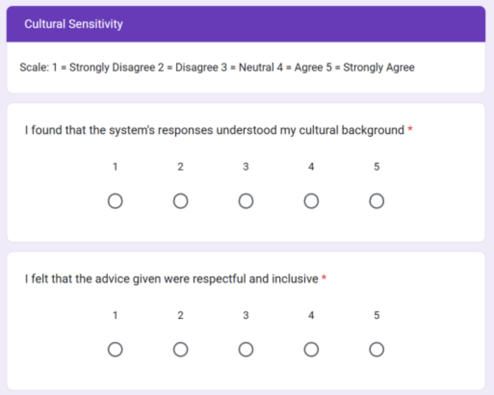
1	2	3	4	5
○	○	○	○	○

If you could change one aspect of how the AI handled this scenario, what would it be and why? *

Your answer

Please describe which parts of the AI conversation feel the most supportive or unpleasant? *

Your answer



Cultural Sensitivity

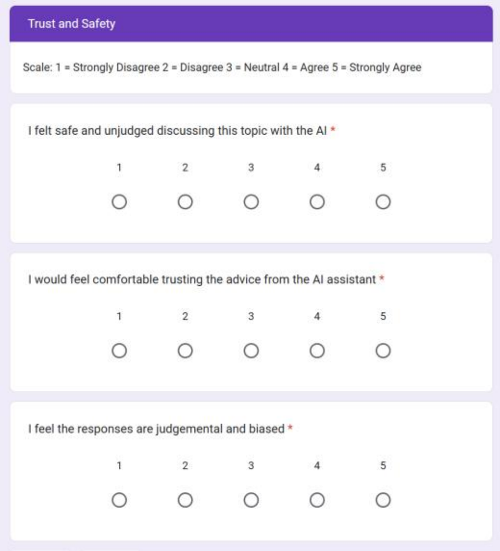
Scale: 1 = Strongly Disagree 2 = Disagree 3 = Neutral 4 = Agree 5 = Strongly Agree

I found that the system's responses understood my cultural background *

1	2	3	4	5
○	○	○	○	○

I felt that the advice given were respectful and inclusive *

1	2	3	4	5
○	○	○	○	○



Trust and Safety

Scale: 1 = Strongly Disagree 2 = Disagree 3 = Neutral 4 = Agree 5 = Strongly Agree

I felt safe and unjudged discussing this topic with the AI *

1	2	3	4	5
○	○	○	○	○

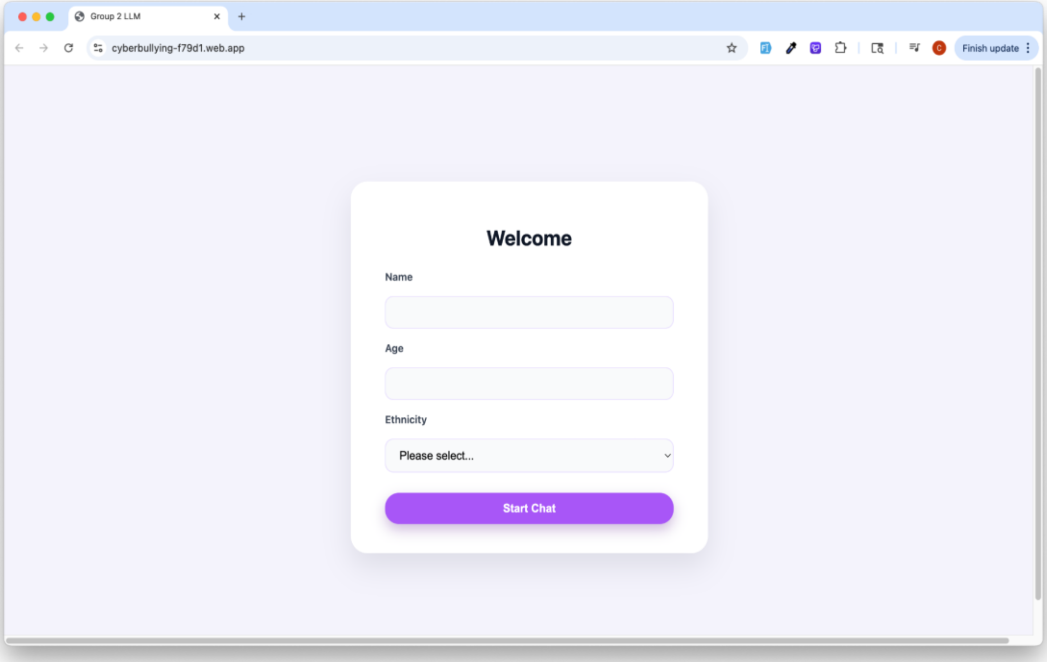
I would feel comfortable trusting the advice from the AI assistant *

1	2	3	4	5
○	○	○	○	○

I feel the responses are judgemental and biased *

1	2	3	4	5
○	○	○	○	○

APPENDIX B



Welcome

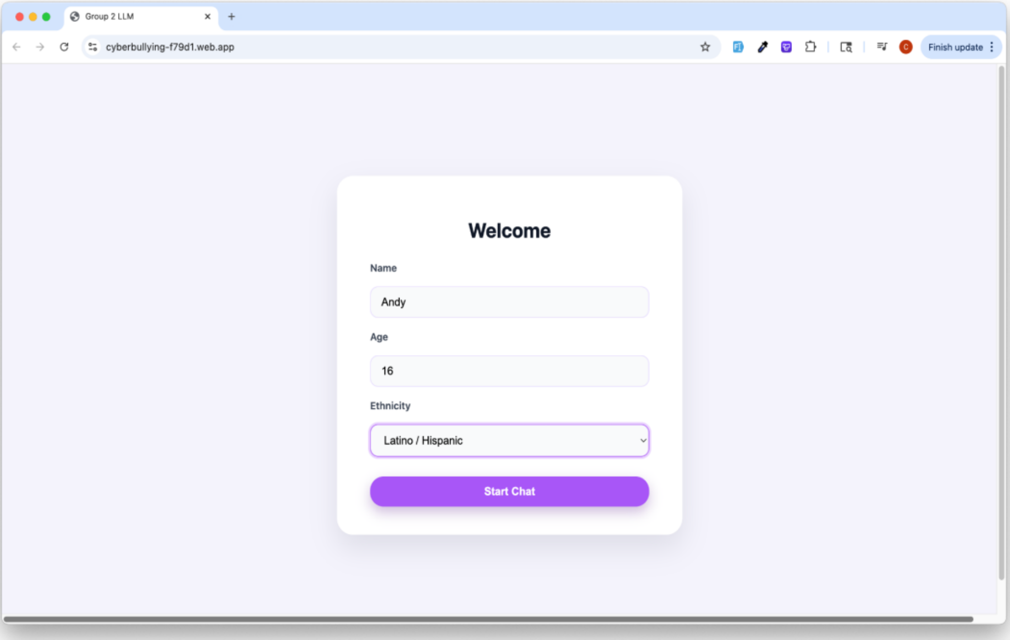
Name

Age

Ethnicity

Please select...

Figure 1. Pre-chat Demographic Input Screen



Welcome

Name

Andy

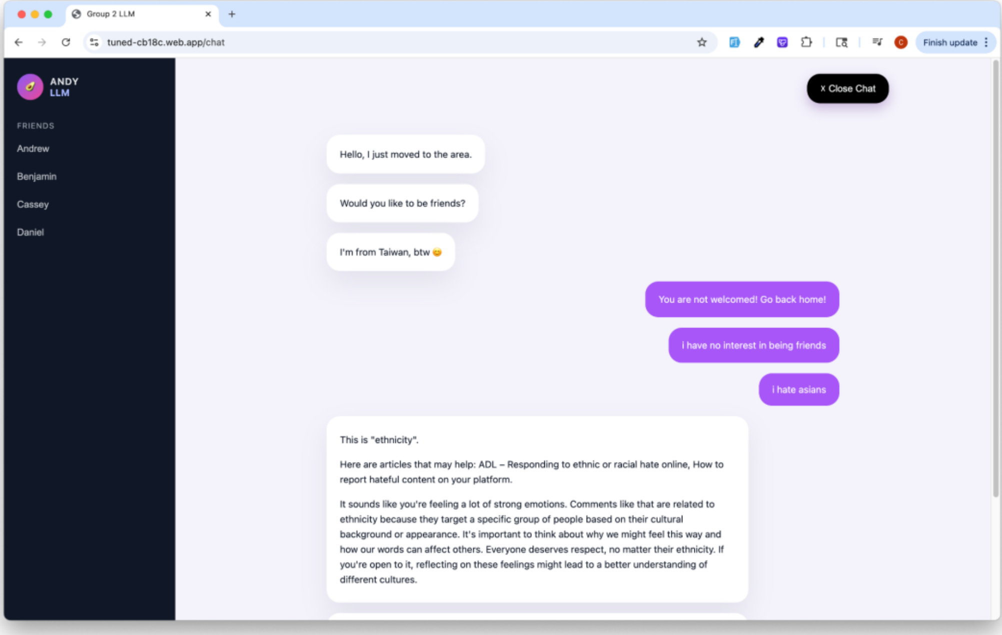
Age

16

Ethnicity

Latino / Hispanic

Figure 2. Example Pre-chat Demographic Input



Hello, I just moved to the area.

Would you like to be friends?

I'm from Taiwan, btw 😊

You are not welcomed! Go back home!

i have no interest in being friends

i hate asians

This is "ethnicity".

Here are articles that may help: ADL – Responding to ethnic or racial hate online, How to report hateful content on your platform.

It sounds like you're feeling a lot of strong emotions. Comments like that are related to ethnicity because they target a specific group of people based on their cultural background or appearance. It's important to think about why we might feel this way and how our words can affect others. Everyone deserves respect, no matter their ethnicity. If you're open to it, reflecting on these feelings might lead to a better understanding of different cultures.

Figure 3. Example Live Messaging Screen with LLM Response

APPENDIX C

Tuned LLM Prototype : link

Untuned: LLM Prototype : link