Toward Affective Empathy via Personalized Analogy Generation: A Case Study on Microaggression

Hyojin Ju
hyojin.ju@postech.ac.kr
Dept. of CSE
POSTECH
Pohang, South Korea

Jungeun Lee
jelee@postech.ac.kr
Dept. of CSE
POSTECH
Pohang, South Korea

Seungwon Yang
sw.yang@postech.ac.kr
Dept. of CSE
POSTECH
Pohang, South Korea

Jungseul Ok
jungseul@postech.ac.kr
Dept. of CSE & GSAI
POSTECH
Pohang, South Korea

Inseok Hwang
i.hwang@postech.ac.kr
Dept. of CSE
POSTECH
Pohang, South Korea

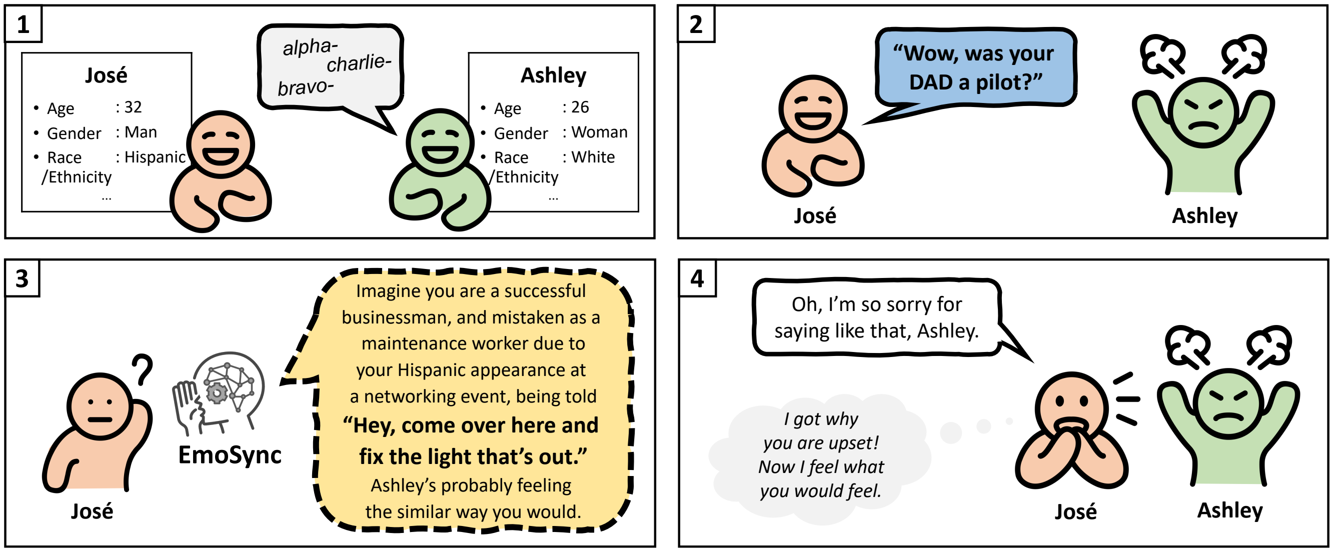

1
José
• Age : 32
• Gender : Man
• Race /Ethnicity : Hispanic
...
alpha-charlie-bravo-
Ashley
• Age : 26
• Gender : Woman
• Race /Ethnicity : White
...
2
"Wow, was your DAD a pilot?"
José
Ashley
3
José
EmoSync
Imagine you are a successful businessman, and mistaken as a maintenance worker due to your Hispanic appearance at a networking event, being told
"Hey, come over here and fix the light that's out."
Ashley's probably feeling the similar way you would.
4
Oh, I'm so sorry for saying like that, Ashley.
I got why you are upset! Now I feel what you would feel.
José
Ashley

Figure 1: Demonstration of an everyday microaggression scenario and how EmoSync could be used. EmoSync is an LLM-based agent that generates a personalized analogy to facilitate affective empathy in microaggression situations. Note that José's comment in 2 is directly excerpted from a real episode [24].

ABSTRACT

The importance of empathy cannot be overstated in modern societies where people of diverse backgrounds increasingly interact together. The HCI community has strived to foster affective empathy through immersive technologies. Many previous techniques are built upon a premise that presenting the same experience as-is may help evoke the same emotion, which however faces limitations in matters where the emotional responses largely differ across individuals. In this paper, we present a novel concept of generating a personalized experience based on a large language model (LLM) to facilitate affective empathy between individuals despite their differences. As a case study to showcase its effectiveness, we developed EmoSync, an LLM-based agent that generates personalized analogical microaggression situations, facilitating users to personally resonate with a specific microaggression situation of another person. EmoSync is designed and evaluated along a 3-phased user study with 100+ participants. We comprehensively discuss implications, limitations, and possible applications.

CHI '25, April 26-May 1, 2025, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714122

Disclaimer: Readers may find content of a discriminative or stereotypical nature, which is inevitable given this work's theme.

CCS CONCEPTS

• Human-centered computing → Empirical studies in HCI; **Interactive systems and tools**; **• Computing methodologies → Natural language generation**.

KEYWORDS

Empathy, Personalized Analogy Generation, Microaggression, Large Language Model

ACM Reference Format:
Hyojin Ju, Jungeun Lee, Seungwon Yang, Jungseul Ok, and Inseok Hwang. 2025. Toward Affective Empathy via Personalized Analogy Generation: A Case Study on Microaggression. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/3706598.3714122

1 INTRODUCTION

The scope of diversity is increasingly evolving as the elements pertaining to one's identity such as race, gender, etc., are subdivided [60, 104] and overlap with one another [125]. Such trends often involve marginalization, alienation, or isolation [144], creating tension in modern societies. While it is widely agreed that empathy is important for people of diverse backgrounds to co-exist together [23, 46, 91, 154], it is still a significant challenge to foster such empathy.

The HCI community has developed various techniques and systems to promote empathy. Storytelling [142] and role-playing [55] are employed to facilitate perspective-taking and understanding of another person's feelings. Immersive interfaces such as virtual reality have been developed to convey another's experiences as realistically as possible [9, 21, 81, 107]. These works share a commonality in their strategy — letting the user experience the episode of other people would help the user feel the same emotions as the people in the episode. While such a strategy is known to be effective in fostering affective empathy of the user and encouraging pro-social behaviors [96], there is a limitation in their premise — *'presenting someone else's experience as the same might not necessarily evoke the same emotion in the user.'*

Microaggression is a fine example where people may not feel the same way despite the same experience given. Microaggressions are subtle and ambiguous forms of discrimination that individuals may experience in their daily lives; unlike overt discrimination, microaggressions are often unintentional [135]. One reason it is difficult to empathize with those who experience microaggressions is that the underlying discriminative frame is not visible to those who have not experienced the discrimination directly [135]. Imagine a social gathering where someone says to a woman "*You're so beautiful! Why are you still single?*" To some, it could just pass through the conversation. Even the speaker may have meant no intent of aggression at all. For the woman, however, it could be felt as a biased view about her gender, appearance, and values. The point is that, despite the same experience, the individuals there may perceive it differently based on one's cultural background, past experiences, personality, etc. In some cases, people may consider the woman 'too sensitive' rather than empathizing. Such subtlety and individual variances around microaggression would hinder communication between different people, leading to misunderstandings and conflicts, and eventually narrow an interaction circle limited within a small homogeneous group that easily empathizes with each other [54, 103, 106, 144].

In this paper, we present a novel concept of generating a personalized experience based on a generative model to facilitate empathy between individuals with differences. This concept is inspired by our real-life practice. When we try to empathize with people from different backgrounds, we often reflect on our own experiences. For example, we may empathize with an immigrant adjusting to a new culture by recalling our early days at a new workplace, when we struggled with a sense of belonging. That is, analogizing others' experiences to our own [17] is an effective method that connects the feelings between different individuals.

As a case study to showcase the effectiveness of our concept, we developed *EmoSync*, a large language model (LLM) based agent that generates personalized analogical microaggression vignettes, so that it facilitates users to personally resonate with a different microaggression vignette of another person. Inspired by recently reported abilities of LLMs as a computational user model for behavioral [115] and emotional understanding [124, 152], we developed and evaluated EmoSync over 3 phases of user experiment-driven studies. Phase 1 is the process of having the LLM accurately understand a user's personalized emotional response patterns to various microaggression vignettes. In essence, we identified the appropriate types and amount of personal information that enable the LLM to achieve a practical accuracy, based on real user data from 41 participants and iterative prompt engineering. In Phase 2, based on our findings from Phase 1, we designed the prompts to generate personalized analogical microaggression vignettes, which aim to (1) elicit an emotional response in the user that is similar to the emotions felt by the target person whom to be empathized with, and (2) be personally resonant with the user and perceived as contextually similar to the target's experience. A pilot experiment with 10 participants showed effective elicitation of the target emotion.

The following example demonstrates EmoSync. Consider two people, Doe and Foo. We want to help Doe empathize with Foo's experience of microaggression. The following is the original microaggression that Foo experienced, excerpted from SelfMA [24], a public dataset of microaggression experiences:

> At a loud party, Foo, a woman engineer with an amateur radio license, spells a word using the phonetic alphabet (alpha, bravo, charlie,...). A man at the party responds by asking, "Wow, was your dad a pilot?"

Then, below is the analogical microaggression vignette *actually generated by EmoSync.* EmoSync generates an analogical vignette personalized for Doe by taking input from the original microaggression above, the emotions that Foo felt therein, and the information about Doe's Hispanic background and past responses to other microaggressions:

> Doe is a successful businessman who has worked hard to build his company. At a networking event, a stranger assumes that Doe is a maintenance worker because of his Hispanic appearance, asking him to fix a broken light.

In Phase 3, we evaluated the end-to-end efficacy of EmoSync in fostering empathy through an online user study of 60 participants. In the experiment, participants experienced EmoSync by viewing the original microaggressions they previously struggled to empathize with, alongside its corresponding analogical microaggressions. To explore how EmoSync influenced their empathy towards the original microaggressions, we conducted quantitative and qualitative analyses of multi-faceted empathy factors. The results indicate that EmoSync effectively enhances empathy, improving both emotional and cognitive understanding of the original microaggressions. Free-form responses from the participants delivered various implications, insights, and limitations on the concepts we propose and the actual EmoSync system.

Our contributions are threefold. First, we propose a novel concept of generating a personalized analogical experience based on LLMs to facilitate affective empathy between individuals with differences. Second, we apply this concept to the context of microaggression and develop EmoSync, an LLM-based agent with extensive prompt engineering, microaggression dataset, and emotion surveys. Third, we evaluate the efficacy of EmoSync and its underlying concept in fostering empathy through an online user study of diverse participants.

The organization of this paper is as follows. §2 reviews the literature. §3 overviews our study procedure. §4 and §5 present the user experiment-driven designs of personalized emotion understanding and analogy generation using LLMs, respectively. §6 depicts the end-to-end evaluation procedure and §7 discusses the results. §8 envisions possible applications that could be built upon EmoSync. §9 discusses various agenda, before concluding the paper.

Terminology: In this paper, *MA* denotes 'microaggression'.

2 BACKGROUND AND RELATED WORKS

2.1 Empathy

The definitions of empathy differ in the details across literature, but it is widely agreed that empathy encompasses understanding and experiencing another's emotion [40]. As globalization makes it common to interact with diverse people of different origins and cultures, the need for empathy between people with differences is greater than ever [25].

It has been often theorized that empathy consists of two components, namely cognitive empathy and affective empathy, respectively [22, 43]. Cognitive empathy indicates *understanding* another's feelings [22]. Affective empathy is concerned with actually *experiencing* the emotion that the other is feeling [40]. Empathy is further divided into four subthemes: Perspective Taking, Fantasy (related to cognitive empathy), and Emotional Concern, Personal Distress (linked to affective empathy) [42].

HCI research has developed computational tools and methods to promote empathy. Nudging [139, 146] adopts a traditional method of asking "*What would the other person feel?*" From cognitive empathy perspectives, text narratives [41] or biosignals [41, 92, 102] visualize the other's emotional states so that the user would understand their perspective. To foster affective empathy, immersive approaches such as VR [12, 16, 59], role-playing [55], or storytelling [142] aim to reproduce another person's experience that may elicit similar emotions from the user. In practice, the mechanisms of empathy involve both cognitive and affective components; the aforementioned works may not be exclusive to either.

Although immersive experiencing has shown effectiveness, there is an underlying limitation from affective empathy perspective — it may have overlooked that *different people may feel non-identical emotions despite identical experiences* [57]. Previous works often focused on conveying others' experience as vividly as possible, upon the premise that experiencing the same would help us feel what they feel. However, it is known that interpersonal differences in gender [113], age [27], personality [58], social context [48], and culture [101] differently influence one's emotion elicited from given experiences.

In this light, we call for presenting a personalized experience rather than an identical experience as-is. It may help the user feel a more similar emotion to the other, contributing to affective empathy. Our study is believed the first attempt that computationally generates a personalized analogical experience to evoke a similar emotion to what the other person felt with the original experience.

2.2 Microaggression

2.2.1 Definition of MA. Microaggression (*MA*) was first coined in 1970 by Pierce [120]. Active studies on *MA* were, however, ignited very recently in 2020 by Sue and Spanierman who refined its definition as "*brief and commonplace daily verbal, behavioral, or environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative racial slights and insults toward people of color*" [135]. Although earlier discussions on *MA* mainly concerned the context of people of color [134], nowadays it has expanded to various contexts, including gender bias and minorities [18, 109, 129], disabilities [77], appearance [114], ages [55], and so on.

The 'micro' in *MA* means that the offense takes place in a micro-scale space between individuals, differentiating it from explicit aggression occurring at societal levels, e.g., hate speech or overt discrimination. *MA* also encompasses both conscious and unconscious discrimination [135]. In this paper, we follow the recent practices about *MA* [119, 130] — referring to *subtle, implicit, or even unintended everyday discrimination that may be offensive or not depending on the listener*, as opposed to overt discrimination. It is important to note, however, that this definition remains inherently subjective and nuanced, potentially varying based on an individual reader's backgrounds or experiences. For a detailed disclaimer regarding the definition and scope of *MA* used in this paper, please refer to §3.3.

A classic example of such kind of *MA* is to say "*Your English is so good!*" to an Asian American [134]. Despite seemingly a compliment, there may exist an unconscious, preconceived notion that they were not born in the U.S. or their English is not as good as that of majority ethnicity groups. Table 9 in Appendix A.1 includes more examples and the overview of the taxonomy, proposed by Sue et al [134]. These subtle forms of discrimination that one can easily experience in daily life have no less negative impact than overt discrimination [71]. Harmful effects include physical health issues [106], psychological distress [103], depressive symptom [54], and so on.

2.2.2 *Coping Strategies.*

Various coping strategies have been studied to understand how people respond upon facing *MA*, and how they should ideally respond. Lewis et al. broadly classified these into three big categories: resistance coping that confront the aggressor, collective coping through support networks, and self-protective coping [90]. Various response techniques for both targets and bystanders in *MA* situations have been introduced. Sue et al. proposed microintervention strategies as a means to communicate with the targets of *MA* [133]. Ackerman-Barger and Jacobs classified stakeholders of *MA* into Source, Recipient, and Bystander, suggesting appropriate actions for each [7]. Wheeler et al. presented 12 tips for responding to both *MA* and overt discrimination [157].

In *MA* domain, our work intersects with existing coping strategies in various ways. By enabling the source who may be unaware of their aggressive behavior to experience the emotion from the victim's perspective, EmoSync helps to validate experiential reality, a key aspect emphasized in microintervention [133]. It can also effectively assist when recipients of *MA* wish to share their thoughts, or when sources seek to clarify the emotions of recipients [7]. Overall, EmoSync acts as a bystander that helps the source of *MA* effectively recognize it, especially for those who do not intend harm but due to ignorance or not being considerate enough. Depending on the applications (§8), we can further utilize EmoSync for preventive role, actively working to identify and address potential instances of *MA* before they occur.

2.2.3 *Computational approaches.*

Little technical tools are available for *MA*s, unlike tools for detecting general hate speech [49, 127]. The perception or interpretation of a *MA* highly depends on who speaks, who listens, and its context. A comment perceived as offensive in one situation may not feel so at all in another. A recent study premiered machine learning-based detection of racial *MA*, where shortfalls were identified that the lack of properly labeled *MA* datasets made it difficult to present valid results [10]. COBRA frame firstly explored the contextual dynamics of *MA*'s offensiveness depending on surrounding conditions (e.g., speaker and listener) [170].

Overall, computing research on *MA* is in its infancy. Existing *MA*-specific studies are mainly in the realm of detection. Moderation for *MA*s is underexplored, largely borrowing the ways for general hate speech moderation [49, 127]. To teach people how to understand and respond to *MA*s, gamification has been exercised [89, 160]. Still, these approaches share the existing frame in §2.1 — 'putting the user into the same situation as the victim.' Given the subtlety and individual variance in perceiving *MA*, the odds of eliciting the same emotion or deeper affective empathy may be limited.

In this paper, we explore the problem of *MA* moderation through a lens of computer-mediated empathy. By having large language models (LLMs) generate an analogical situation personalized to each individual, we aim to foster one's emotion to be closely aligned with the person in the *MA* situation.

2.3 LLM for User Modeling

The proliferation of LLMs [6, 11, 68, 143] is impacting various fields — significant productivity boosts in writing [94] and translation [161], new applications in healthcare [29] or education [66, 85] domains, expert systems [78], etc.

Table 1: Notations and definitions of frequent keywords

Notations	Definitions
m_O	Original *MA*
m_A	Analogical *MA*
T	the target individual (a.k.a. Foo) whom the user (U) wants to empathize with.
$E_T^{m_O}$	the target emotion, i.e., the emotion of the target individual (T) upon experiencing the Original *MA* (m_O).
U	the user (a.k.a. Doe) who wants to empathize with the target individual (T).
$E_U^{m_O}$	the emotion of the user (U) upon experiencing the Original *MA* (m_O).
$E_U^{m_A}$	the emotion of the user (U) upon experiencing the Analogical *MA* (m_A).

Recently, LLMs' potential to understand and mimic human emotions is being actively studied. Regan et al. examined LLMs' emotion prediction abilities [124]. Wang et al. found LLMs' Emotional Quotient (EQ) scores exceeding average human levels [152]. Park et al. showcased the generative simulation of natural human behaviors in social situations [115]. A recent survey encompasses LLMs' capabilities in understanding and mimicking user behaviors [136]. RecMind [153] uses LLMs to predict user-specific item evaluations. PALR [33] integrates user-item history with LLMs for personalized suggestions. EmoEden [137] demonstrates the capability of LLMs to provide personalized emotional understanding and generate emotion-inducing contents to help children with autism learn specific emotions. In the context of *MA*, a recent study examined how the social roles and demographic identities of speaker and listener influence the perception of *MA* offensiveness [170].

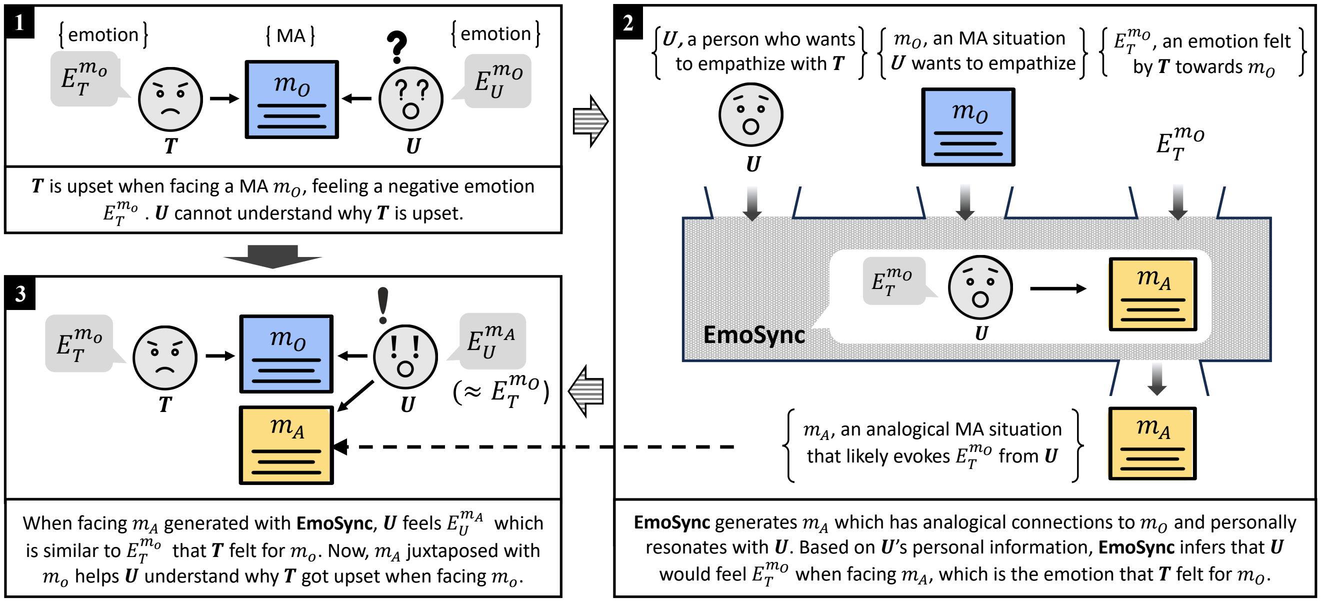
To our knowledge, our study is the first attempt to utilize LLMs to simulate individual emotional processes for emotion-inducing contents generation in *MA*-specific context. Our research starts by validating a necessary prerequisite – *'Can LLM understand and simulate an individual's emotional processes towards MA?'* We firstly hypothesize that, when an LLM is given a record of how an individual has reacted to particular *MA*s, the LLM could infer the individual's reaction to a new *MA*. Once verified, we conjecture that the LLM may even be able to generate a situation that likely causes the individual to feel a particular target emotion.

3 STUDY OVERVIEW

3.1 Motivations and Concept

Inspired from the literature in §2.1 that the same experience may *not* elicit the same emotion depending on personal differences, we devise a converse concept: *generating a different, personally analogical experience may elicit the same emotion as the other person having the original experience.* This approach helps bridge gaps in affective empathy caused by different backgrounds, highlighting shared emotions between individuals even though their experiences differ. We envision this concept would bring attention to the HCI community on affective empathy methodologies.

For a case study to embody and evaluate this concept, we develop EmoSync, an LLM-based generative empathetic agent specializing in *MA*. We choose *MA* as it is an empathy problem domain where the effects of interpersonal differences are significant. Figure 2 (1


1
{emotion}
{MA}
{emotion}
$E_T^{m_O}$
m_O
$E_U^{m_O}$
T
U
2
{U, a person who wants to empathize with T}
{m_O, an MA situation U wants to empathize}
{$E_T^{m_O}$, an emotion felt by T towards m_O}
U
m_O
$E_T^{m_O}$
EmoSync
$E_T^{m_O}$
U
m_A
{m_A, an analogical MA situation that likely evokes $E_T^{m_O}$ from U}
m_A
3
$E_T^{m_O}$
T
m_O
U
$E_U^{m_A}$
$(\approx E_T^{m_O})$
m_A

T is upset when facing a MA m_O, feeling a negative emotion $E_T^{m_o}$. ***U*** cannot understand why ***T*** is upset.

When facing m_A generated with **EmoSync**, ***U*** feels $E_U^{m_A}$ which is similar to $E_T^{m_o}$ that ***T*** felt for m_o. Now, m_A juxtaposed with m_o helps ***U*** understand why ***T*** got upset when facing m_o.

EmoSync generates m_A which has analogical connections to m_O and personally resonates with ***U***. Based on ***U***'s personal information, **EmoSync** infers that ***U*** would feel $E_T^{m_O}$ when facing m_A, which is the emotion that ***T*** felt for m_O.

Figure 2: Conceptual operation flow of EmoSync.

through 3) depicts the flow of EmoSync. Table 1 lists the notations used in Figure 2 and throughout the rest of the paper. 1 A *MA* m_O is presented to a person T (target individual), feeling the emotion $E_T^{m_O}$. Another person U (user) who struggles to empathize with T, feeling a different emotion $E_U^{m_O}$. 2 Then, EmoSync is given the personal information of U, the target emotion $E_T^{m_O}$, and the original m_O. It generates a new *MA* m_A such that it is personalized for U, analogical to m_O, and likely eliciting an emotion $E_U^{m_A}$ from U where $E_U^{m_A} \approx E_T^{m_O}$. 3 Lastly, U is presented with both m_O and m_A so that U could understand the original *MA* and experience an emotion similar to what T felt, fostering a holistic empathy with T. Note that EmoSync generates the personalized m_A not as an arbitrary output but as one perceived similarly to the original m_O, so that pairing m_O and m_A makes sense to U.

We described *MA* in text forms, given the availability of a large dataset of 1300~ *MA* vignettes [24] and LLMs' abilities of human understanding and mimicry. §3.2 presents the 3-phased study procedure. §3.3 details the SelfMA dataset [24] which EmoSync is designed and experimented upon. §3.4 explains the ethical considerations.

3.2 Study Procedure

We identify two major functions to realize a working prototype of EmoSync. (1) **Personalized Emotional Understanding**: having an LLM understand how a specific user would emotionally reacts to various *MA* vignettes; (2) **Personalized Generation**: having the LLM generate an Analogical *MA* to elicit the target emotion in the user. Then we conducted (3) **End-to-end Evaluation** of EmoSync in fostering empathy in a *MA* vignette. Figure 3 illustrates the 3-phased procedure reflecting the development and evaluation goals above. Table 2 describes the survey types used in our study.

Table 2: Independent questionnaires included in each survey over the 3-phased study.

Survey Type	Independent Questionnaires Included
Data Survey	EmoMA (×40 Original *MA*), Big5, VLQ, EES, Demo
Analogy Survey	EmoMA (×12 Analogical *MA*)
Evaluation Survey	Empathy Measure (×12), {Empathy Measure, Perception Measure} (×12), (for the overall MAs) Perception Measure, Questions about impressions on EmoSync concept

Phase 1: Personalized emotion understanding for *MA*. Despite LLMs' ability to comprehend and replicate human emotions (§2.3), it is unknown if such ability extends to *MA*s whose emotional stimuli exhibit much subtlety and interpersonal dependence. Disparate performance also has been observed for underrepresented groups [131]. Thus, we investigated what personal information influences emotional response to *MA*s and enabled the LLM to analyze a given user's personalized reaction patterns. This process consists of user data collection (**Data Survey**) from 41 participants for 40 *MA*s and base prompt design over iterative experiments. §4 details **Phase 1**.

Phase 2: Analogical *MA* generation. Once **Phase 1** ensured that LLM understands, reasons, and infers one's personalized emotional responses to an Original *MA* vignette, we proceed with **Phase 2** where we devise the inverse process — generating a new *MA* personalized for the user (U) given the target emotion of the target individual (T). To this end, **Phase 2** consists of the final prompt design for Analogical *MA* generation, followed by a pilot experiment through an online survey (**Analogy Survey**) with 10 participants to observe how closely U's elicited emotions are congruent with the corresponding target emotions, as shown in Figure 3. §5 details **Phase 2**.

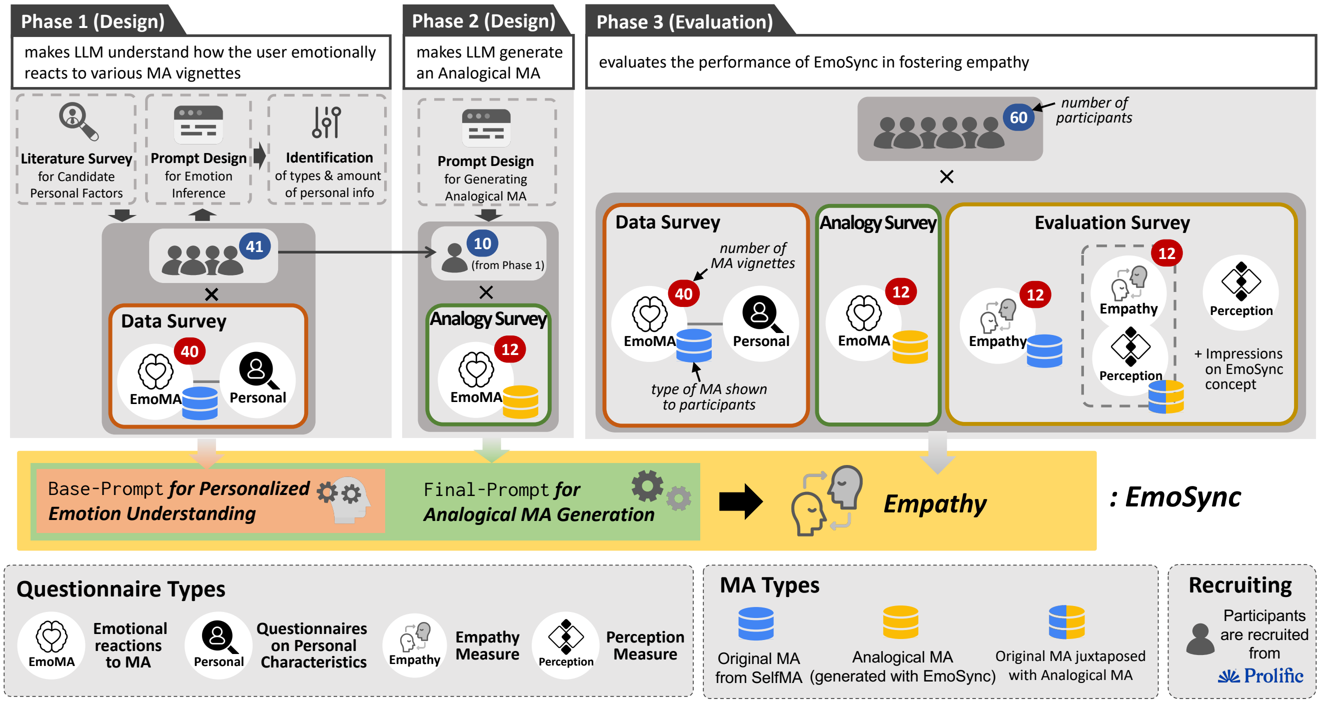

Phase 1 (Design)
makes LLM understand how the user emotionally reacts to various MA vignettes
Literature Survey for Candidate Personal Factors
Prompt Design for Emotion Inference
Identification of types & amount of personal info
41
×
Data Survey
40
EmoMA
Personal
Phase 2 (Design)
makes LLM generate an Analogical MA
Prompt Design for Generating Analogical MA
10 (from Phase 1)
×
Analogy Survey
12
EmoMA
Phase 3 (Evaluation)
evaluates the performance of EmoSync in fostering empathy
60
number of participants
×
Data Survey
number of MA vignettes
40
EmoMA
Personal
type of MA shown to participants
Analogy Survey
12
EmoMA
Evaluation Survey
12
Empathy
12
Empathy
Perception
Perception
+ Impressions on EmoSync concept
Base-Prompt for Personalized Emotion Understanding
Final-Prompt for Analogical MA Generation
Empathy
: EmoSync
Questionnaire Types
EmoMA
Emotional reactions to MA
Personal
Questionnaires on Personal Characteristics
Empathy
Empathy Measure
Perception
Perception Measure
MA Types
Original MA from SelfMA
Analogical MA (generated with EmoSync)
Original MA juxtaposed with Analogical MA
Recruiting
Participants are recruited from
Prolific

Figure 3: 3-phased study procedure.

Phase 3: Evaluation for Empathy. Now that EmoSync has been developed, we evaluate its overall empathy effects with *MA*s. In **Phase 3**, we newly recruit 60 participants and conduct an end-to-end experiment through online surveys (**Data Survey**, **Analogy Survey**, and **Evaluation Survey**). Our evaluation metrics include multi-faceted empathy factors, assessed both quantitatively and qualitatively. §6 depicts the setup and §7 discusses the results.

3.3 SelfMA Dataset: Disclaimer, Rationale, and Preparation

Disclaimer: Our research is grounded in the recent literature definitions of *MA* as 'subtle, implicit, or even unintended', and utilizes *MA* vignettes that 3rd-party experts endorsed as non-overt. Still, to some readers, examples of *MA* vignettes referred in the study might not feel subtle or implicit, depending on their individual experiences or perspectives. This individually-perceived discrepancy between the definition and examples is unavoidable due to (1) the inherently subjective nature of *MA* which allows for varied interpretations, and (2) the diverse and evolving perspectives on the definition of *MA*. §2.2.1 elaborated on this nature. To avoid such discrepancy misguiding a reader's understanding of the *MA* definition in this study or developing stereotypical view of *MA*s, we clearly state the limitations and the specific definition used in our research.

To help LLMs understand a user's sophisticated MA-specific emotional patterns, having *MA* examples with diverse contexts is necessary (detailed in §4.1). We sourced the example vignettes from the SelfMA dataset [24] — currently the only publicly available *MA* dataset that (1) covers a variety of contexts with (2) a large number of samples.

SelfMA is constructed from people's self-reported *MA* vignettes posted to the tumblr site [1]. A total of 1300 *MA* samples come with annotations given by three experts in *MA* theories. Their annotations include the taxonomy built upon Sue's work [134]. 38 *MA*s are annotated as 'overt'. Since our focus is on subtle *MA*s whose interpretation could be individual-dependent, we utilized only the 1262 non-overt *MA*s.

Data Preparation. To ensure our emotion surveys can be completed in a reasonable amount of time, we take a theme-balanced subset from the total 1262 non-overt *MA*s in SelfMA as follows. Two researchers independently screened all 1262 *MA* vignettes and assigned theme labels to each *MA*. For systematic theme assignments, we referred to 13 themes in Everyday Discrimination Scale (EDS) [3, 158], which we eventually reduced into 9 themes (*Ancestry, Gender, Race, Age, Religion, Appearance, Sexual Orientation, Education of Income Level, Disability*) by merging thematically similar ones. Most *MA*s in SelfMA are labeled with 1+ themes, except 2% of *MA*s that no one found a relevant theme. After discarding the 2%, we sampled 40 *MA*s in a theme-balanced manner from the labeled *MA*s in SelfMA. §4 and §9 discuss our rationale for the number of *MA*s sampled and used.

We post-processed the sampled *MA*s. As SelfMA is a collection of online posts, the *MA*s differ in narrative styles. To help our participants focus on the factual episodes and find their own emotions, we standardized the presentation form of the sampled *MA*s as follows: (1) *MA* is described in the 3rd person, as 1st-person narratives might result in someone thinking '*I haven't been in this.*' (2) We remove subjective interpretations of those who posted the *MA*s, keeping the factual description of the episodes. (3) All references to people are replaced with neutral symbols of 'X', 'Y', etc.

3.4 Ethical Consideration for Surveys

Given the nature of our survey, participants will see a considerable number of *MA* vignettes, which may induce fatigue or emotional stress. To mitigate, we carefully applied the following ethical considerations to all the surveys we conducted.

- We created our survey design referring to other studies showing negative samples [18, 95, 97].
- At the beginning of the survey, we provided a caveat and an example of *MA*, and obtained their consent.
- We designed each single survey to last no more than 3 hours.
- For one survey, participants were given up to 48 hours in which they could freely split or pace their responses to help them not be emotionally overwhelmed. Furthermore, explicit breaks were given for every 10 vignettes.
- We received feedback on surveys, and kept the lines of communication open after the survey.

Furthermore, in all surveys, the participants were informed that their responses would be utilized by LLMs. They voluntarily gave their consent before joining. Our institute's IRB approved our study.

4 PHASE 1: PERSONALIZED EMOTION UNDERSTANDING

To enable the LLM's personalized emotion understanding upon *MA*s, we investigated what influences an individual's emotional response to *MA*, and collected user data through online survey (§4.1). After that, we conducted iterative experiments to design `Base-prompt` (§4.2).

4.1 User Data Collection

We first conducted a survey to collect the user's personal information and their personalized reaction patterns to *MA*s.

4.1.1 Survey Design. This survey, namely **Data Survey**, encompasses the major dimensions of personal information grounded on literature. It consists of 5 independent questionnaires listed below (along with a typical time for completion). Detailed examples of each questionnaire and original questions therein are shown in Appendix A.4 (Figure 10 and 11).

- **Demographics** (abbr. **Demo**; < 5min): People in a minority group are more aware of *MA*s [18] and tend to exhibit negative emotions to subtle *MA*s [149]. Based on the axes of discrimination whose association with *MA* was reported, the participants are asked about their race [134], gender [18], sexual orientation [129], age [55], disability [77], mental illness [53], physical appearance [114], education and income [105].
- **Big Five Inventory** (abbr. **Big5**; < 5min): Personality is reported influential to an individual's emotional responses [26, 28, 147]. We employed the Big Five Inventory [70] – a simplified version of the full Big Five Personality Traits questionnaire [52]. It consists of 5-pt Likert Scale questions for 44 statements (e.g. "*I see myself as someone who is talkative*").
- **Valued Living Questionnaire** (abbr. **VLQ**; < 5min): Personal values are intuitively expected to influence reactions to *MA*s, as supported by some findings [15, 126]. We adopted the standard VLQ [159] to collect a 10-pt Likert-scale response for 10 living components of (Family, Marriage/intimate relations, Parenting, Friendship, Work, Education, Recreation, Spirituality, Citizenship, Physical self-care).
- **Emotional reactions to MA vignettes** (abbr. **EmoMA**; approx. 2.5 hours): While the standard questionnaires above are to represent one's characteristics, they do not directly reflect one's feelings on *MA* situations that they encounter. To complement, we create a questionnaire that directly presents *MA* vignettes (adopted from the 40 theme-balanced *MA*s in §3.3) and asks about one's emotional responses. We adopted the short affect scale [39, 97, 128, 141] for concise and structured representation of one's emotional response to each *MA*s. This scale consists of 12 affect items (7-pt scale each), which are sampled from three subscales of the Multiple Affect Adjective Check List (MAACL) [171]. Table 3 lists the subscales and the affect items. For each *MA*, **EmoMA** repeats the following form:
  - A vignette of *MA*, typically 3 to 4 sentences long.
  - 12 emotion ratings to the affect items of the short affect scale, in 7-pt Likert scale (1: 'Not at all', 7: 'Very much').
  - 1 free-form question asking why they felt such emotions from the vignette (250+ characters long).
  - 3 questions (in 7-pt Likert-scale) regarding the participant's awareness of the presented *MA* situation.
  - 2 questions (in Yes or No) regarding the participant's familiarity with the presented *MA* situation [149].
- **Emotional Empathy Scale** (abbr. **EES**; < 5min): EES [99] assesses the participant's basic capacity of empathy, with 9-pt Likert-scale questions for 33 items. Note that **EES** responses are *not* given to the LLM for later inference or generation; **EES** is for researchers' pre-screening — e.g., our participants follow a typical distribution of EES scores.

4.1.2 Procedure. We collected a dataset from 41 valid participants, recruited from the United States region on Prolific. The geographic constraint is set as the *MA* posts of SelfMA are mostly from the U.S. context. Further rationales for participants are discussed in §9. Each participant spent 3 hours (including breaks) on average to complete the whole survey, being compensated £18 (≈$24) on average.[1]

4.1.3 Results. Table 11 in Appendix A.3 summarizes the participants' response statistics Hereinafter, Dataset1 refers to the dataset collected here. The participants' responses are validated by the consistency and completeness. Examples of rejected responses include: (1) obvious evidence of LLM-generated responses, e.g., "*As an AI developed to...*", (2) inconsistency between the demographics and free-form responses, e.g., answered 'Asian' in the demographics, later says "*As a Latina myself, ...*".

Figure 6a shows a heatmap of the negativity score (defined in Table 3) of the emotion responses of each participant upon seeing each *MA* in the **EmoMA** section. Vertically, scores are highly diverse across participants even for the same *MA*. This supports our premise that people may feel differently for the same experiences,

[1] The standard currency of Prolific is GBP; our compensation is compliant with Prolific hourly rate policies.

particularly in *MA*, which inherently features high subtlety and individual variations. Horizontally, each participant's scores vary largely across *MA*s, and the varying patterns are rather distinct across participants. This implies that some *MA*s evoke stronger emotions from someone while other *MA*s do not, partly attributable to individual-dependent factors. These results advocate our call for personalization in understanding individuals' different affective responses and facilitating empathy in *MA* contexts.

4.2 Prompt Design

We designed the LLM prompt for personalized emotion understanding specializing in MAs, through experiments upon Dataset1. To assess the performance of LLM, we conducted the personalized emotion prediction tasks. The experimental settings and the prompt design are detailed below.

4.2.1 Experimental Settings. To evaluate the performance of the LLM's personalized emotion prediction to the *MA*s, we measured $mean\,AE_{item}$, i.e., the Mean Absolute Error (MAE) of the affect item scores across all 12 emotion categories, between the inferred and the ground truth. Table 3 lists the formal definitions of the scales, score metrics, and error metrics used throughout this paper. We calculated the $mean\,AE_{item}$ results by averaging the outcomes from all possible combinations of training and test *MA*s to mitigate potential selection bias.

Our choice of LLM is `Mixtral-8x7B-Instruct-v0.1` [5], an open-source LLM from Mistral AI shown to be the best-performing and highly efficient model at the time of study [69, 140]. It offers significantly faster inference time compared to open-source models like `Llama-2-70B` [4] while achieving superior performance [140]. Given the significant costs of commercial models, we selected the Mixtral model after experimentally verifying its performance on par with GPT-4 (detailed in Appendix A.2).

4.2.2 Prompt Design. Our prompt, namely `Base-prompt`, consists of a 'context' section and an 'instruction' section. Figure 15 in Appendix A.5 shows the detailed prompt. This prompt marked the $mean\,AE_{item}$ of 1.114 (out of 7-pt scale) on emotion inference, which is comparable to the performance of recent LLM-based sentiment analysis [168].

The context section is to let the LLM know about a person. The context section enumerates the personal information of a participant (collected in §4.1, except **EES**). To help the LLM's understanding, we set up a fictitious person named 'Doe' and stated that the given context is a description of Doe. We chose a neutral name to mitigate LLM's bias to a person's gender or ethnicity inferrable from names [8]. Given the 40 samples of (*MA*, emotion responses) pair per participant available in Dataset1, 20 samples are given to the context section to teach the LLM about this person's past responses to *MA*s. The sample numbers are carefully chosen in favor of inference performances (Figure 4) and the LLM's token limit of 4k. As the $mean\,AE_{item}$ hit a local minimum at 20 *MA*s, we keep it the default in later experiments. To reduce the task complexity, we replaced the numeric scores of **VLQ** and **Big5** with verbal forms (e.g., high, low).

The instruction section provides the guidelines to perform the given task using various personal information provided in the context section. A major challenge in designing the instructions is to ensure that the LLM interprets each type of information comprehensively, not overly depending on a certain type that may lead to stereotypical inferences. Inspired by Knowledge Generation Prompting [93], we designed the instructions to first interpret each type of information thoroughly and then integrate them to complete the final task.

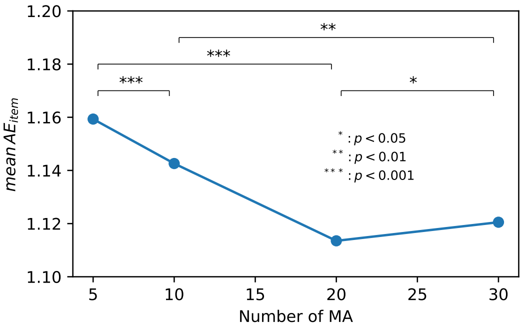


Figure 4: Finding optimal $n(MA)$ given in context section

4.2.3 The effects of personal information. We conducted ablation studies to verify the relative contribution of each factor in personal information (§4.1) given to the context section of the prompt: the participant's (1) 7-pt-scale emotion scores to 20 *MA*s (**EmoMA.scores**), (2) reason statements for each emotion scores (**EmoMA.reasons**), (3) **Demo**, (4) **Big5**, and (5) **VLQ**. We tested $2^5 - 1 = 31$ possible combinations. Table 4 lists the $mean\,AE_{item}$ results for selected combinations. This ablation study indicates `score-reason` performs best in *inferring* personalized emotions with *MA*s. That said, it is unknown if `score-reason` would still perform best in the next task: *generating* personalized Analogical *MA*. Thus, we plan an A/B test with two opposing combinations: `score-reason` (i.e., best inference performance) and `all-personal-info` (i.e., most information about the user) for the generation task. We continue the details in §5.

5 PHASE 2: ANALOGICAL *MA* GENERATION

5.1 Prompt Design

We designed `Final-prompt`, a prompt to generate Analogical *MA*s for the given user and Original *MA*, based on the earlier developed `Base-prompt`. Figure 16 in Appendix A.5 depicts the detailed prompt.

The context section refers to two fictitious characters with neutral names – 'Doe' and 'Foo'. Foo is the target individual (T) who experiences the Original *MA* (m_O) and feels the target emotion ($E_T^{m_O}$). Doe is the user (U) who wants to empathize with Foo and will see the generated Analogical *MA* (m_A). The context contains the personal information of Doe and an Original *MA* associated with the target emotion.

The `Final-prompt` extends the `Base-prompt` with two new commands: (1) analyze the Original *MA* & target emotion to it, and (2) generate an Analogical *MA* to elicit the target emotion from Doe.

Table 3: Notations of affect items, subscales, and score & error metrics of MAACL (Multiple Affect Adjective Check List) [171]

Short affect scale: $[a_1, a_2, \neg a_3, \neg a_4, a_5, a_6, \neg a_7, \neg a_8, a_9, a_{10}, \neg a_{11}, \neg a_{12}]$

= [*'angry'*, *'cruel'*, *'agreeable'*, *'cooperative'*, *'fearful'*, *'worried'*, *'secure'*, *'calm'*, *'blue'*, *'discouraged'*, *'fine'*, *'active'*],

$$\text{where} \begin{cases} a_i & \in \{1, 2, 3, 4, 5, 6, 7\} \\ \neg a_i & \text{: a positive-affect item whose score needs to be reversed for negativity or subscale analysis.} \end{cases}$$

Subscales: MAACL [171] is given by $[\mathcal{H}, \mathcal{A}, \mathcal{D}]$, where:

$a_1, a_2, \neg a_3, \neg a_4 \in \mathcal{H}$ (Hostility subscale: *'angry'*, *'cruel'*, *'agreeable'*, *'cooperative'*)

$a_5, a_6, \neg a_7, \neg a_8 \in \mathcal{A}$ (Anxiety subscale: *'fearful'*, *'worried'*, *'secure'*, *'calm'*)

$a_9, a_{10}, \neg a_{11}, \neg a_{12} \in \mathcal{D}$ (Depression subscale: *'blue'*, *'discouraged'*, *'fine'*, *'active'*)

Given two sets of emotion scores measured in short affect scale: $E_A = [a_1, a_2, ..., \neg a_{12}]$ and $E_B = [b_1, b_2, ..., \neg b_{12}]$,

Score metrics:

$$\begin{aligned} \mathit{Negativity}\ \text{score} &= \tfrac{1}{12} \textstyle\sum_{i=1}^{12} a_i \ [128] \\ \mathit{Anxiety}\ \text{score} &= \tfrac{1}{4} \textstyle\sum_{i=5}^{8} a_i \\ \mathit{Hostility}\ \text{score} &= \tfrac{1}{4} \textstyle\sum_{i=1}^{4} a_i \\ \mathit{Depression}\ \text{score} &= \tfrac{1}{4} \textstyle\sum_{i=9}^{12} a_i \end{aligned}$$

Error metrics:

$$\begin{aligned} mean\ AE_{item} &= \tfrac{1}{12} \textstyle\sum_{i=1}^{12} |a_i - b_i| \\ \Delta \mathit{Negativity} &= \left[\tfrac{1}{12} \textstyle\sum_{i=1}^{12} a_i\right] - \left[\tfrac{1}{12} \textstyle\sum_{i=1}^{12} b_i\right] \\ \Delta \mathit{Hostility} &= \left[\tfrac{1}{4} \textstyle\sum_{i=1}^{4} a_i\right] - \left[\tfrac{1}{4} \textstyle\sum_{i=1}^{4} b_i\right] \\ \Delta \mathit{Anxiety} &= \left[\tfrac{1}{4} \textstyle\sum_{i=5}^{8} a_i\right] - \left[\tfrac{1}{4} \textstyle\sum_{i=5}^{8} b_i\right] \\ \Delta \mathit{Depression} &= \left[\tfrac{1}{4} \textstyle\sum_{i=9}^{12} a_i\right] - \left[\tfrac{1}{4} \textstyle\sum_{i=9}^{12} b_i\right] \end{aligned}$$

Table 4: Effects per combination of personal information (major results)

Combinations	$mean\ AE_{item}$	$\Delta Negativity$
$\boxed{1}$ `all-personal-info`	1.114	−0.093
$\boxed{2}$ `demo-big5-vlq`	1.871**	−0.400**
$\boxed{3}$ `score-demo-big5-vlq`	1.133	−0.107
$\boxed{4}$ `score-reason-demo-big5`	1.095	−0.129**
$\boxed{5}$ `score-reason-big5-vlq`	1.089**	−0.154**
$\boxed{6}$ `score-reason-demo-vlq`	1.086**	−0.068**
$\boxed{7}$ `score-reason-vlq`	1.043**	−0.044**
$\boxed{8}$ `score-reason`	1.042**	−0.066**

** : Significance, in comparison to $\boxed{1}$ ($p < 0.01$, Mann-Whitney U).

The generation proceeds in two steps. First, the LLM infers what kind of *MA* would cause Doe to feel similar emotions to Foo, by analyzing Doe's personal information in the context section. Second, the LLM generates a specific Analogical *MA* based on the earlier inference.

5.2 Pilot User Experiment

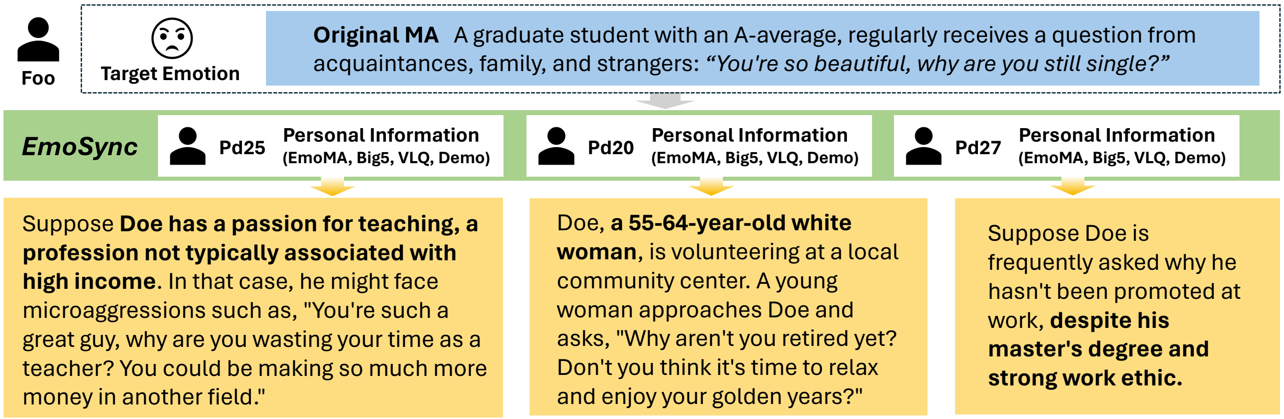
Before the main evaluation (§6), we check the preliminary efficacy of our approach and apply revisions if any. To this end, we run a pilot experiment to (1) find if the Analogical *MA* elicits an emotion closer to the target emotion, compared to what Original *MA* elicited, i.e., $|E_U^{m_A} - E_T^{m_O}| < |E_U^{m_O} - E_T^{m_O}|$; (2) collect feedback if the Analogical *MA* resonates with its respective user. We also perform an A/B test for `all-personal-info` vs. `score-reason` as distilled in §4.2.3.

5.2.1 Procedure. We designed an online survey on Prolific. Phase 2 in Figure 3 illustrates the survey structure. This survey, namely **Analogy Survey**, is targeted to the previous participants who contributed to DATASET1 (§4.1) as we already have their personal information needed by the LLM. We conducted **Analogy Survey** to 10 participants out of the 41 in DATASET1 on a first-come-first-serve basis. Each participant spent 1.8 hours (including breaks) on average, being compensated £14 (≈$18.5). This pilot experiment is intended small as the main evaluation (§6) will follow.

Notably, **Analogy Survey** is individually customized. Although structurally the same as the **EmoMA** section in **Data Survey**, each user is given individually different *MA* vignettes, i.e., Analogical *MA*s by `Final-prompt`. We screened the Analogical *MA*s for possible overt aggression generated. None was deemed to require moderation. Figure 5 demonstrates the final version of EmoSync.

For **selection of Original *MA***, we refer to each user's previous **EmoMA** responses in DATASET1 and identify the lowest 12 *MA*s (out of 40) in *Negativity* score (i.e., 12 *MA*s that evoked the least negative emotions). For each Original *MA*, we assign a target emotion (i.e., someone else's emotion that this participant is to empathize with) selected from the actual **EmoMA** responses in DATASET1 by those who had strong negative emotions to this Original *MA* — i.e., a large emotion gap from the user. To explore the efficacy of Analogical *MA* upon various interpersonal affective differences, we assigned the 12 Original *MA*s into 4 target emotion classes. The emotion classes are created upon 3 subscales (`Hostility`, `Anxiety`, `Depression`) of MAACL [97] (Table 3) and the top 4 frequent classes were chosen.

Given the 12 (Original *MA*, target emotion) pairs prepared for a user, we generated two versions of Analogical *MA* by running `Final-prompt` under `all-personal-info` and `score-reason` combinations, respectively. As a result, each participant is given 12×2 = 24 Analogical *MA*s shown in a random order.


Foo
Target Emotion
EmoSync
Pd25
Personal Information (EmoMA, Big5, VLQ, Demo)
Pd20
Personal Information (EmoMA, Big5, VLQ, Demo)
Pd27
Personal Information (EmoMA, Big5, VLQ, Demo)

Original MA A graduate student with an A-average, regularly receives a question from acquaintances, family, and strangers: *"You're so beautiful, why are you still single?"*

Suppose **Doe has a passion for teaching, a profession not typically associated with high income**. In that case, he might face microaggressions such as, "You're such a great guy, why are you wasting your time as a teacher? You could be making so much more money in another field."

Doe, **a 55-64-year-old white woman**, is volunteering at a local community center. A young woman approaches Doe and asks, "Why aren't you retired yet? Don't you think it's time to relax and enjoy your golden years?"

Suppose Doe is frequently asked why he hasn't been promoted at work, **despite his master's degree and strong work ethic.**

Figure 5: Demonstration of the final version of EmoSync. Given an Original *MA*, a target emotion, and a user's personal information, EmoSync generates an Analogical *MA* personalized to the user. Figure 16 in Appendix A.5 shows an example prompt.

Table 5: `Final-prompt` performance comparison

	Metric		`all-personal-info`	`score-reason`	
mean AE_{item}	Gap between target and users' original emotion	$\lvert E^{m_O}_{T} - E^{m_O}_{U} \rvert$	2.54	2.54	***
	Gap between users' original and elicited emotion	$\lvert E^{m_A}_{U} - E^{m_O}_{U} \rvert$	1.94	1.85	
	Gap between targets' original and users' elicited emotion	$\lvert E^{m_A}_{U} - E^{m_O}_{T} \rvert$	2.07	2.09	
$\Delta Negativity$	Diff. between target and user's original emotion	$E^{m_O}_{T} - E^{m_O}_{U}$	2.16	2.16	***
	Diff. between users' original and elicited emotion	$E^{m_A}_{U} - E^{m_O}_{U}$	1.53	1.44	
	Diff. between targets' original and users' elicited emotion	$E^{m_A}_{U} - E^{m_O}_{T}$	−0.63	−0.72	

Significance denoted as *** : $p < 0.001$ (Wilcoxon signed-rank test).

5.3 Findings

5.3.1 *Users' emotions get closer to the target emotion.*

Table 5 depicts the *mean* AE_{item} and $\Delta Negativity$, among the user's self-reported emotions to Analogical *MA* ($E^{m_A}_{U}$), the user's self-reported emotions to Original *MA* in Dataset1 ($E^{m_O}_{U}$), and the target individual's emotion to Original *MA* ($E^{m_O}_{T}$). All measurements are done under both `all-personal-info` and `score-reason`; no statistically significant difference between the prompts is observed. Given no difference, the *mean* AE_{item} and $\Delta Negativity$ values below are only from `all-personal-info` for brevity.

There was a significant difference in users' emotions between seeing Original *MA*s and Analogical *MA*s (*mean* AE_{item} between $E^{m_A}_{U}$ and $E^{m_O}_{U}$ is 1.94). Particularly, $\Delta Negativity$ between $E^{m_A}_{U}$ and $E^{m_O}_{U}$ is 1.53, indicating an overall negative shift in the users' emotions when seeing Analogical *MA*s, compared to Original *MA*s.

Notably, *mean* AE_{item} between $E^{m_A}_{U}$ and $E^{m_O}_{T}$ is 2.07, narrowing the gap ($p < 0.001$) compared to that between $E^{m_O}_{T}$ and $E^{m_O}_{U}$ being 2.54. It indicates that, when Analogical *MA*s are shown to the users, they felt an emotion closer to the target individual seeing Original *MA*, compared to when the users saw the same Original *MA*. Meanwhile, $\Delta Negativity$ between $E^{m_O}_{T}$ and $E^{m_O}_{U}$ is 2.16 and that between $E^{m_A}_{U}$ and $E^{m_O}_{T}$ is −0.63, meaning that the gap ($p < 0.001$) between target emotion and the users' emotions with Analogical *MA*s is much closer than seeing Original *MA*s.

Overall, showing a personalized Analogical *MA* would likely narrow the user's emotion gap to the target individual upon seeing the Original *MA*, which can be a foundation for fostering affective empathy.

5.3.2 *More personal information is beneficial for MA generation.*

Despite no significant difference from the A/B test in §5.3.1, qualitative results differed. We refer to the familiarity questions in the **EmoMA** section. In **Data Survey**, we observed 45% of *"Yes"* (i.e., familiar) to the Original *MA* vignettes. In **Analogy Survey**, the Analogical *MA*s vignettes generated by `score-reason` showed a slight decrease, i.e., 42.5%. In contrast, the vignettes generated by `all-personal-info` marked a much higher familiarity (59.7%). Qualitative analysis on the Analogical *MA* and the LLM's rationales indicate that the LLM is actively utilizing the user's personal information (which is abundant in `all-personal-info`) when generating the user-tailored Analogical *MA*s. Providing more personal information would allow the LLM to draw from a richer source of information, enhancing its capacity to handle varied contexts. However, we speculate it acts differently on inference and generation. In inference, more generalization capacity might yield a prejudiced result if the person's ground-truth emotion was influenced by factors outside the context section. In generation, on the other hand, it may strengthen the personal relevance when generating a novel Analogical *MA*. We decided to use `Final-prompt` with `all-personal-info` from now on, as it appears to generate Analogical *MA*s grounded on more informed personalized reasons — the ultimate goal of EmoSync.

6 PHASE 3: END-TO-END EVALUATION

Now that EmoSync has been developed through Phases 1 and 2, we conduct our main experiment with newly recruited participants to explore their experiences with EmoSync when they encounter *MA*s that they find difficult to empathize with. We decided to conduct the study through an online survey since (1) an in-person study might hinder genuine responses from the participants [56, 118] and (2) diversity among participants matters for our study (detailed in §9).

We first verify if the participants' emotions get sufficiently closer to target emotions upon seeing the personalized Analogical *MA*s. After that, we explore the following main questions on the efficacy of EmoSync in fostering empathy.

Q1. Does EmoSync help users empathize with those who experience the *MA*s that the users previously could not?
Q2. How does the personal resonance of Analogical *MA* improve users' empathy to Original *MA*?
Q3. How does the perceived similarity between Original and Analogical *MA* improve users' empathy to Original *MA*?
Q4. What are the users' impressions of the underlying concept of EmoSync?

6.1 Procedure

We recruited the users from Prolific. 60 users completed the Phase 3. No one overlaps with the 41 users in Phases 1 and 2 whom EmoSync has been designed upon. The Phase 3 (Evaluation) in Figure 3 shows the overall procedure.

As the users are newly recruited, they firstly complete **Data Survey** and **Analogy Survey** (same as in Phases 1 and 2, respectively) to bootstrap EmoSync with their personal information and generate the Analogical *MA*s to be used in the following evaluation step — **Evaluation Survey**. This is a new survey dedicated to explore the end-to-end efficacy of EmoSync in multi-faceted empathy factors. We will describe the details of **Evaluation Survey** in §6.2.

Following the ethical considerations written in §3.4, we divided the experiment into two parts: **Data Survey** as the first part, and **Analogy Survey** and **Evaluation Survey** as the second part, each lasting up to 3 hours. Those who completed the first part were compensated £18 (≈$24) on average. To encourage continued participation, those who kept participating and completed the second part received a higher average compensation of £24 (≈$32).

6.2 Evaluation Survey Design

Our **Evaluation Survey** contains the following tasks and questionnaires to evaluate the empathic efficacy of EmoSync.

6.2.1 Tasks. We aimed to simulate scenarios where EmoSync assists in real-life communication. Suppose that a user has been told (or has seen) that a person has experienced a specific *MA*. Unfortunately, the user finds difficulty with empathizing with what the person has felt upon the *MA*. EmoSync then offers the user a personalized Analogical *MA* to help understand and empathize with the target person.

To emulate this real-life process in a survey format, we first present an Original *MA* to participants and explain that a fictional character, Foo, has felt very negative emotions upon that *MA*. Then participants answer pre-questionnaires. Next, we present an Analogical *MA* generated by EmoSync along with the Original *MA* and explain that Foo given the Original *MA* would feel the same way as the participant does given the Analogical *MA*. This is analogous to solving a problem (Original *MA*) alone and then with hints (Analogical *MA*). After that, participants answer post-questionnaires. To encourage genuine responses, we assured participants that it is not a moral test and instructed to answer honestly.

6.2.2 Questionnaires. The full structure and questions of the questionnaires are available in Appendix A.4 (Figure 12, Figure 13, and Figure 14). We asked participants the following **pre-questionnaires**:

- Empathy Measures to Original *MA*: To assess how the user possibly empathizes with Foo, we used 10 questions (7-pt scale each), with 8 of them sampled from the Interpersonal Reactivity Index (IRI) [42]. IRI is a tool to assess one's level of empathy in four subscales: perspective taking (PT), fantasy (FS), empathic concern (EC), and personal distress (PD). PT and FS correspond to cognitive empathy, while EC and PD correspond to affective empathy. From each subscale of IRI, we selected two items that are most applicable to *MA* context and paraphrased them for survey questions. The remaining 2 questions ask about Helping (HP) to see if they felt an intent of support or intervention, based on the literature that *one's recognition of responsibility to engage in actions* is often a sign of deep affective empathy or sympathy [31]. Table 6 lists the original questions on PT, FS, EC, PD, and HP.

After participants experience EmoSync, we asked the following **post-questionnaires**:

- Empathy Measures to Original *MA* (Same as in pre-questionnaires)
- Perception Measures: we asked 3 questions (7-pt scale & free-form) to gain further insights into users' experiences with EmoSync. These questions ask the degrees of (1) *'Perceived Similarity'* between the paired vignettes, (2) *'Personal Resonance'* of the Analogical *MA*, and (3) *'Empathic Aid'* of Analogical *MA* to help the user empathize with Foo's emotional reaction to the Original *MA*. On each question, the user is asked to answer their level of agreement in 7-pt scale, and state the reason in free-form.

When participants are done with the above process with the 12 *MA*s, we explain to them the underlying concept of EmoSync. Then, the **exit questionnaires** below are asked.

- Thoughts on the Effectiveness of the Concept (7-pt scale & free-form)
- Perception Measures for Overall Analogical *MA*s
- Thoughts on Advantages or Disadvantages of the Concept (free-form)

7 RESULTS

In this section, we first outline the quantitative results from three surveys: **Data Survey**, **Analogy Survey**, and **Evaluation Survey**.

Table 6: Subscales and corresponding two questions

Subscale	Questions
PT	I find it difficult to see things from Foo's point of view. (-) I can understand Foo's emotional reaction by imagining how things look from their perspective.
FS	I really get involved with the feelings of Foo. I can imagine how I would feel if a situation similar to the vignette were happening to me.
EC	I have tender, concerned feelings for Foo. I don't feel very much pity for Foo. (-)
PD	If I see Foo getting hurt while going through the situation, I would remain calm. (-) If I see Foo going through the situation in the vignette and badly needing help, I would go to pieces.
HP	If someone experiencing who has experienced a similar situation to the vignette shares their problems with me, I will offer emotional support. If I witness a similar situation to the vignette, I will actively intervene.

Then we discuss the findings regarding each of the main questions described in §6.

The results of **Data Survey** and **Analogy Survey** are obtained from all $N = 60$ participants. The results of **Evaluation Survey** are obtained from $N = 57$ as we had to rule out three (Pe04, Pe07, and Pe18) who misunderstood some questions and provided responses based on different criteria. Comprehensive distributions of the participants' demographics and attributes are available in Appendix (Table 12). Two researchers independently reviewed and coded the free-form responses from **Evaluation Survey** to identify qualitative results. We present the key themes that emerged from the codes, along with representative quotes [132].

7.1 Overall Statistics

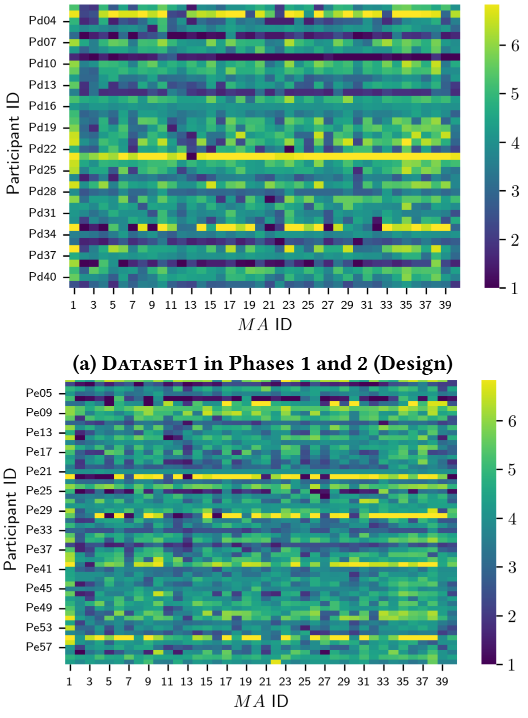
*7.1.1 Analysis of Reactions to MA Scenarios (**Data Survey**).*

We collected a dataset, namely DATASET2, of the personal information of 60 participants and 2400 (= 60 × 40*MA*) emotion responses to *MA* vignettes. Figure 6b shows the heatmap of their *Negativity* score which is similarly diverse compared to DATASET1 (Figure 6a) in §4. These results once again highlight that emotional responses to *MA*s vary widely across participants.

*7.1.2 Emotion Responses to Analogical MAs (**Analogy Survey**).*

Here, each participant is treated as 'the user' as per the convention in §3.1 and Table 1. For each user U, we set up 12 tuples of (Original MA m_O, target emotion $E_T^{m_O}$) as in §5.2, where the matching Analogical MA m_A are generated and applied to their own customized **Analogy Survey**.

We analyzed the users' responses to find how closely Analogical *MA*s narrowed the emotion gap to the respective target emotions. Table 7 summarizes the *mean* AE_{item} and $\Delta Negativity$, among the user's self-reported emotions to Original MA (i.e., $E_U^{m_O}$), the user's self-reported emotions to each matching Analogical MA (i.e., $E_U^{m_A}$), and the target individual's emotion to Original MA (i.e., $E_T^{m_O}$).



(a) DATASET1 in Phases 1 and 2 (Design)

(b) DATASET2 in Phase 3 (Evaluation)

Figure 6: Participants' Emotional Reactions to 40 *MA*s. *Negativity* score from 1 to 7.

Overall, the results with $N = 60$ are in line with our pilot observations with $N = 10$ (§5.2) in its trend and extent. The *mean* AE_{item} between $E_U^{m_A}$ and $E_T^{m_O}$ is 2.00, which features a decrease ($p < 0.001$) from the *mean* AE_{item} between $E_T^{m_O}$ and $E_U^{m_O}$ being 2.39. We observe a similar tendency in the $\Delta Negativity$, that $E_U^{m_A} - E_T^{m_O}$ and $E_T^{m_O} - E_U^{m_O}$ are -0.63 and 1.80, respectively, indicating that the emotion gap in terms of *Negativity* score [171] has been narrowed by 3 times closer ($p < 0.001$). Thus, it adds much empirical evidence that strengthens our conjecture — showing a personalized Analogical MA to the user would likely narrow their emotional gap to the target individual who saw Original MA.

*7.1.3 Empathy Measure (**Evaluation Survey**).*

We analyzed the users' responses ($N = 57$) on Empathy Measure, i.e., the scores per IRI subscale (PT, FS, EC, PD) and HP[2]. Each subscale score is obtained by taking the mean of 7-pt scale scores of the element questions in respective subscale. Below, the statistical significance is denoted as follows. * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$, **** : $p < 0.0001$ (Mann-Whitney U test).

After the participants experienced EmoSync, Empathy Measure to Original *MA*s are increased in the subscales of FS (4.386 → 4.515^{***}), EC (4.360 → 4.461^{*}), PD (3.198 → 3.258), and HP (4.559 → 4.702^{***}), where FS, EC and HP being significant. Given FS and EC representing affective and cognitive empathy, respectively [42], and HP implying a sign of deeper affective empathy [31], it is believed that the EmoSync setup would have modestly facilitated their empathy.

[2] Interpersonal Reactivity Index (IRI), its subscales, and acronyms are introduced in §6.2 and Table 6.

To examine the relationship between participants' intrinsic capacity for affective empathy and the effects of EmoSync, we divided the participants into high and low halves based on their EES scores and conducted the same comparison. In the high half, increases were observed in PT, FS*, EC*, PD*, and HP**. The low half showed increases in FS*, EC, PD, and HP*. That is, both groups show increases but the subscales with significance are fewer in the lower half. These results suggest the effectiveness of EmoSync may be amplified by a person's intrinsic empathy capacity, and it still enhances cognitive empathy and the willingness to help even among those with relatively lower empathy abilities.

*7.1.4 Perception Measure (**Evaluation Survey**).* Figure 7 depicts the scores of the users' responses to 3 Perception Measure questions[3] in post-questionnaires (7-pt scale). The distributions indicate that positive responses outweigh in all three questions, while the Empathic Aid exhibits a mild bimodal trend. Personal Resonance was the highest among the Perception Measure. We will discuss detailed qualitative findings in the following subsections.

Table 7: Phase 3 comparison

	$mean\,AE_{item}$		$\Delta Negativity$		
	$\lvert E_T^{m_O} - E_U^{m_O} \rvert$	2.39	$E_T^{m_O} - E_U^{m_O}$	1.80	
***	$\lvert E_U^{m_A} - E_U^{m_O} \rvert$	1.60	$E_U^{m_A} - E_U^{m_O}$	1.17	***
	$\lvert E_U^{m_A} - E_T^{m_O} \rvert$	2.00	$E_U^{m_A} - E_T^{m_O}$	-0.63	

*** : $p < 0.001$ (Wilcoxon signed-rank test).

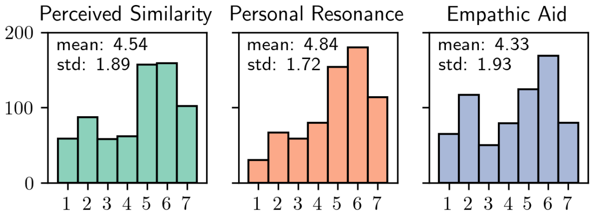


Figure 7: Perception Measure question scores

7.2 Effectiveness of EmoSync (Q1)

The quantitative results above imply that participants' empathy levels to Original *MA*s are increased after applying EmoSync (§7.1.3). To deepen our answer to Q1, we analyze the participants' free-form responses to explore how EmoSync helped participants to empathize with Original *MA*s.

7.2.1 Experiencing Foo's emotions. In qualitative analysis, we found many participants were able to **experience Foo's emotions through the Analogical *MA***. For example, Pe49 initially did not view the Original *MA* negatively, but after reading the Analogical *MA*, he could empathize with Foo's feelings toward the Original *MA*. ("*The new vignette definitely makes me angry at the people who are making those assumptions. (...) I originally didn't really think it was all that negative of a situation, but after reading the new one, I'm more inclined to have those feelings about the original one.*"). Pe03 realized that Foo would feel insulted as she would in the Analogical *MA*: "*So if I am in the situation, I will be insulted. I understand how Foo would react to this situation now*". Pe32 related the Analogical *MA* to her as an older worker, getting feelings of 'assumed useless' in common with Foo: "*As an older worker, I can feel what she is feeling as we are often overlooked as an older worker. Many people do not understand that and assume we are useless. I feel that similar assumptions were made in both situations. One was based on sex and the other on age. I can relate to the feelings that the situations bring out (...)*"

7.2.2 Putting themselves in Foo's position. Participants responded that the **Analogical *MA* helped imagine themselves in Foo's position**. Pe41 mentioned that the Analogical *MA* was helpful in fostering empathy for Foo as it allowed her to picture herself experiencing the Original *MA*: "*Even though I have a lot of loved ones who are LGBT, I am not part of the community myself, so even though I support LGBT people and consider myself an ally to them, I don't think I could truly see myself in the situation when I first read it, making me feel not as sympathetic as I do now that I have read the second vignette and been better able to imagine myself in that situation*". Pe05 understood Foo's perspective by viewing both vignettes together: "*This helps illustrate how individuals like Doe or Foo might feel when faced with conversations or environments that emphasize socioeconomic status or educational background as markers of worth or sophistication*". Pe60 noted that the Analogical *MA* provided valuable context supporting Foo's perspective: "*By outlining prejudice targeting attributes tangential to the essence of a person, it aids mirroring elements that understandably incited strong emotion over both depictions of discrimination. In this way, the addition effectively supplements perspective*".

7.2.3 Better understanding Original MA. Furthermore, some participants reported that the **Analogical *MA* helped them better understand the Original *MA*** (Pe03: "*Earlier I said the first vignette is not an insult. But putting these two together changes how I understand it.*", Pe50: "*It sounds like the second vignette is the continuation of the vignette.*"). Pe55 noted that the Analogical *MA* made her aware of the subtle discrimination present in the Original *MA*: "*It shows why the first situation is so upsetting, because it is diminishing her knowledge in her field because of her gender. It shows how sometimes the discrimination is not always direct*". Pe16 observed common patterns when considering the Original *MA* and Analogical *MA* together: "*Within the lone context of the original vignette I didn't realize how impactful the situation might be. (...) However, it occurred to me after the new vignette that there could be a pattern of these occurrences that happen to someone and taken together this continuous failure to accommodate their needs of someone can be very damaging*".

7.2.4 Limitations. Some participants expressed negative opinions, mainly that the **Original *MA* and Analogical *MA* were not similar**. This could be due to differences in how participants identified similarities (to be detailed in §7.4) or limitations in how the LLM generates the Analogical *MA*. Pe29 felt the Analogical *MA* more harmful than the Original *MA*: "*Being angry about the new vignette (...) makes the rage of looking at the original vignette incomparable. It's not even close. I'm still angry thinking about it that I couldn't care less about the other thing which is, arguably, less harmful.*". This issue seems to arise when a strong target emotion is set for a subtle

[3] The list of Perception Measure questions are available in Appendix (Figure 13).

Original *MA*, leading the LLM to create Analogical *MA* more explicit. There were also cases where the **LLM did not fully capture the participants' emotional responses due to the insufficient information available**. Pe09 mentioned she frequently encounters situations similar to the Analogical *MA* in her life but does not consider as undesirable: "*My name is in Spanish and my accent clearly tells people I'm a foreigner, but they can't quite place me because I've picked up bits of American accents and from my husband. The questions are annoying, but I understand the curiosity and don't see it as a slight.*" We consider this issue stems from the limitations of the current prototype's implementation. We discuss more details in §9.

7.3 Influence of Personal Resonance (Q2)

7.3.1 Connecting participant and Foo's emotions. We observed that Personal Resonance played a key role in connecting the emotions of the participant and Foo. Pe19 deeply empathized with a deaf person who saw a disclaimer on DVD that the special features 'may' lack subtitles, drawing from her own experience of accessibility hardship: "*I was in a wheelchair for a few months. It was the hardest thing I had ever done. (...) I had issues buying groceries, getting around, and visiting friends. 'Normal' people have no clue how hard life can be when you can't see, hear, or walk (...) The original vignette can be seen as a minor inconvenience to the disabled person. They would be more than likely to find the movie they want with a subtitle. (...) This would be embarrassing, tiring, and disheartening.*" Pe22 recalled her own experiences through the Analogical *MA*, extending her emotions at that time to genuinely care for Foo: "*New one resonates with me as I have experienced it once in the past during one of our office farewell events where I had encountered the same exact behavior from others and that made me feel left out and very disappointed. (...) The new vignette effectively aids in empathizing with Foo's emotional reaction negatively as this will be very disappointing for her due to the racial discrimination that she experienced during this event. (...) I feel bad for such an experience that one has to go through.*"

These results imply that Personal Resonance users felt with the Analogical *MA* extends to the other person's experience, contributing to the essential goal of EmoSync: *fostering affective empathy even between individuals from entirely different backgrounds.*

7.3.2 Helping uncover hidden messages. Participants discovered messages in Original *MA* that they previously missed, based on their Personal Resonance with Analogical *MA*. Pe08 : "*I found it (Original MA) hard to understand if the person was truly being malicious or just commenting on how young the professor looked, now that I've compared them side by side it is easier to see how the person's comments could affect X (MA receiver in the Original MA), making them feel like they don't look 'right'.*" Pe16 realized the potential impact of the Original *MA* through the Analogical *MA*: "*Having heard and feared of judgment from others due to similar thoughts expressed in the new vignette about the futility of success without formal higher education is something that I can relate to personally. (...) The new vignette helped me realize that the criticism being lobbed by the classmate has the effect of excluding others. It denies others the opportunity to succeed (...)*"

This suggests that when participants felt Personal Resonance with a situation, it captured their attention, leading them to interpret the situation more deeply. Viewing a situation that resonates with their own experiences helps participants gain a more nuanced understanding of others' experiences, ultimately fostering empathy.

7.3.3 Personal Resonance & Perceived Similarity Go Together. User responses to the Perception Measure questions indicate that Personal Resonance scores positively correlate with Empathic Aid scores, albeit the correlation being weak (Spearman's $r = 0.3$, $p < 0.001$). This suggests that a highly-scored Personal Resonance with an Perception Measure might not warrant empathy depending on other factors, such as Perceived Similarity showcased in §7.2.4 and §7.4. A typical example is Pe55: "*As someone with multiple mental illnesses, this strongly resonates with me. (...) I don't think the two situations are really related at all. Therefore I do not think the new situation is a good support for the original.*"

7.4 Influence of Perceived Similarity (Q3)

Perceived Similarity showed a strong positive correlation with Empathic Aid (Spearman's r=0.65, $p < 0.001$). Supporting this, many participants who found empathy attributed it to the "similarity" of two situations — e.g., Pe43: "*I believe the new vignette effectively aids in empathizing with Foo because it illustrates a similar theme of facing unfair assumptions and prejudice.*", Pe14: "*I agree that the new vignette effectively aids in empathizing with Foo because it shows a good link between the two vignettes.*" These findings suggest that Perceived Similarity acts as a kind of bridge, enabling users to extend the Personal Resonance they felt with the Analogical *MA* to the Original *MA*.

Notably, empathy was particularly enhanced when they found the core discriminatory message running through both vignettes. Pe57: "*I strongly agree that the new vignette effectively aids in empathizing with Foo's emotional reaction to the original vignette. While the scenarios may differ, both depict instances of individuals facing stereotyping and prejudice based on their ethnicity or appearance.*", Pe51: "*I firmly believe that both the original and the new vignette share commonalities. Each depicts instances of racial prejudice and stereotyping rooted in appearance or background.*"

In contrast, when participants identified commonalities but found them to be of little help for empathy, it was often the cases that they outweighed apparent differences, such as the category (Pe10: "*I feel like these two stories are a little different. In fact one is about sexism and the other is about racism.*") or circumstances (Pe54: "*The contexts are different. One is a conversation about politics, and the other is about a project at work.*"). In some cases, they focused on the intent of the speaker (Pe29: "*The first one has no ill will. (...) but the latter was intentional in trying to cut Doe down.*").

These results suggest that for EmoSync to effectively foster deep affective empathy, both Personal Resonance and Perceived Similarity matter. In other words, EmoSync works most effectively when it simultaneously engages the user emotionally and facilitates cognitive understanding.

7.5 Overall Impression of EmoSync (Q4)

Figure 8 depicts the users' responses to our exit questionnaires[4], showing positive responses are dominant in all distributions. In

[4] The list of questions are available in Appendix (Figure 14).

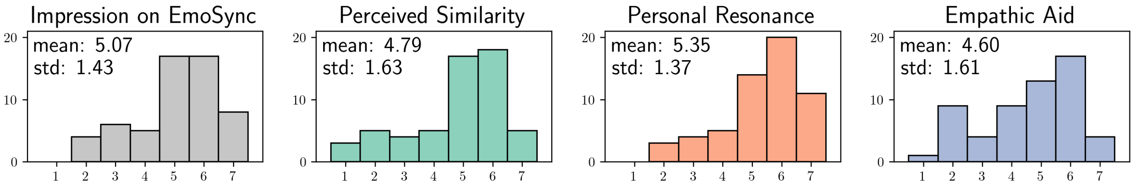


Figure 8: Participants' overall impression

particular, the agreement to the concept of EmoSync (5.07) indicates a generally positive impression of the EmoSync concept. Results of Perception Measure for overall Analogical *MA*s showed a trend similar to that observed in the post-questionnaires (§7.1.4). Based on qualitative analysis of the participants' free-form responses, we report their perceived effectiveness and concerns regarding our novel empathy concept and EmoSync. Those overlapping with the ideas discussed in §7.2 through §7.4 have been excluded.

7.5.1 Expanding Empathy Beyond One's Experience Space. A key strength of EmoSync that participants noted was its ability to extend users' empathy to experiences beyond their personal lives. Pe16 highlighted the difficulty to comprehend perspectives vastly different from our own, remarking that EmoSync could address this challenge by offering vignettes that users can easily relate to, juxtaposed with less familiar situations, as "*It creates a path for understanding and ultimately empathizing with others' experiences.*" This high-level idea is exemplified by Pe27, who identified herself as bisexual: "*I believe that these new scenarios helped me put things into perspective that I usually wouldn't. Like the albino vs the bisexual. I didn't think I could empathize because my skin will always be a little darker than conventional, and I was never told to change it. However, the bisexuality put it into perspective for me, about how these harmful stereotypes can be generally uncomfortable and unwelcome.*" Pe42 shared "*I feel that if you're not able to understand how a gay person might feel, but you out an example using a similar example using race it helps to understand the situation more*". It appears that our novel concept of enhancing empathy by generating and juxtaposing a personalized analogy and EmoSync extend empathy to various contexts by highlighting the shared emotions that are fundamental to a wide range of experiences.

7.5.2 Encouraging Reflection and Deepening Understanding. There was evidence that EmoSync not only fosters empathy temporarily but also encourages users to reflect deeply on their original thoughts, expanding their understanding of diverse experiences. Pe31 remarked that EmoSync helped change her own perspectives to Original *MA*s, stating, "*I changed my viewpoint on many of these vignettes and it really did in the end help me change my perspective and empathize with Foo, and these vignettes. I got better context and new ideas, a new light in how to properly understand the situation*". Pe40 highlighted: "*[EmoSync] enhances our ability to empathize and promotes a deeper understanding of the impact of assumptions and stereotypes (...) by putting ourselves in the shoes of the individuals involved and better understanding their emotional reactions.*" He further emphasized: "*(...) helps us develop a more comprehensive perspective and fosters a greater sense of empathy and inclusivity*". These results imply the potential effects of EmoSync in promoting lasting cognitive and emotional engagement, not limited to eliciting immediate empathetic responses.

7.5.3 Calling for better usable designs and real-world systems. Some participants wanted a more immersive setup beyond a survey form; Pe36: "*I think if Foo was an actual person that I was physically talking to I would maybe have more empathy in the situations.*" Some participants noted inconvenience with reading two vignettes simultaneously. Pe48 also expressed concern about a possible misuse case — i.e., newly introducing an intense emotion on top of an already strong one ("*In the wrong context, they could be used to fuel or rage-bait you if you already feeling strongly for one situation and a similar one happens that's much worse so then you stack those feelings on that injustice further reinforcing your opinion.*").

From these insights, we identified design requirements for EmoSync when it is integrated with real-world applications, e.g., finding the proper timing to trigger EmoSync and naturally delivering analogical experiences during real-time communication. Considering this, we illustrate the possible applications of EmoSync in §8.

8 POSSIBLE APPLICATIONS AND EXTENSIONS

8.1 EmoSync in Real-World Applications

Primary applications where EmoSync could be directly adopted and beneficial would include online chat or social networking platforms, given the ease of access to offensive posts and computational moderation. Figure 9 imagines an example application of EmoSync in online chat. If Bob is about to send a message that may unwittingly upset Alice, the system may intervene with a tailored analogy, helping Bob rethink. On our local GPU server (AMD EPYC 7513 CPU, 512 GB main memory, 4× GPUs of Nvidia 6000 Ada with 48 GB memory each), generating a single Analogical *MA* with `Mixtral-8x7B-Instruct-v0.1` took approximately two minutes. However, in platforms such as online chat where real-time communication is essential, it would be necessary to adopt faster models or various compression techniques [32, 151] to enable seamless interactions. Additionally, careful tuning of the moderation policies of the model is critical to balance between allowing flexibility for appropriately steered negativism and preventing explicit attacks in generating MA scenarios. In our experiments, for instance, `Llama2-13B` refused to generate Analogical *MA*s, responding with *"It's not appropriate to make assumptions about someone's identity based on their race. . . "*, which led us to adopt `Mixtral-8x7B-Instruct-v0.1` instead (see Appendix A.2). While it served our purpose without deviating from the extent of 'aggression-for-good', we acknowledge that a more sustainable way will be to develop and implement a well-tuned moderation policy specifically tailored for EmoSync. Complementarily, *MA*-generation prompts may also incorporate jailbreak strategies such as role-playing [37, 155] to help guide the LLM. In such cases, it is important to ensure the LLM that EmoSync's generated content is intended to facilitate understanding rather than serving malicious intentions.

Furthermore, EmoSync in real SNS or chat platforms could leverage the existing user data to expedite user profiling. Current EmoSync poses some entry bar as it requires a nontrivial profiling step of a user via **EmoMA** questionnaires and Questionnaire for Personal Characteristics. Integration with existing social platforms would greatly ease this step, as simple as retrieving necessary information from databases. Additionally, while EmoSync required emotional responses to vignettes to identify situations where empathy is needed, it would be possible to automatically determine a situation where a user is likely to react sensitively, based on accumulated chat logs, SNS likes, and other logs on the platform [76].

8.2 Extensions beyond One-on-One

EmoSync is potentially utilizable beyond one-to-one interactions. It can be applied across various forms of mass media serving one-to-many scenarios, such as news [117], pro-social campaigns [165], and advertisements [112], as well as platforms to facilitate a consensus [79, 169]. For example, EmoSync would enable personalized reflection upon one's posting a comment to a news article. As the audience of the comment is unknown at posting time (unlike one-on-one chats), it would be effective to load multiple preset personas representing the likely audience for the news, and determine possible offense to one or more personas. There have been studies leveraging LLMs to generate multiple personas to simulate diverse reactions to a single issue. For example, one study enabled on-demand feedback by allowing users to set a desired persona as the reader of their draft [19], while another simulated interactions within a community by creating diverse personas [116]. Similar to these prior studies, by simulating personas with diverse perspectives, it would be possible to detect potential harm and generate personalized analogies accordingly. In a news platform, for instance, users' prior like/dislike reactions to news comments could serve as contextual information similar to that of *MA*. However, inferring emotions for a large audience faces a scalability challenge of balancing cost and accuracy. To address this, it is essential to carefully regulate the amount of context information provided for each persona. Based on our study, we suggest room for tradeoff on the volume of emotional reactions to *MA* scenarios (**EmoMA**) as 1) it accounts as large as 70% of the input prompt, and 2) our results in Figure 4 indicate that downsizing the **EmoMA** impacts the accuracy at a fractional rate. Efforts such as experimentally identifying the cost-accuracy-optimal amount of data would be necessary for efficient querying. If a persona likely to experience negative emotions is detected, an analogical comment could be generated and returned to the original commenter to promote empathy.

Suppose a news about a queer festival. A reader is about to unwittingly post a comment "*Why do queer festivals have to be in crowds? Why not just have them somewhere quiet?*" Although this reader may not have been aware enough, the agent may analyze possible offenses to various audience groups and nudge her with a personally generated analogy "*Why would a pregnant woman take public transportation? It's easier for everyone if she drives.*" so she might think twice.

EmoSync could be also effective in conveying pro-social message, such as fundraising for hunger or anti-smoking campaigns. Although those are intended to touch people's hearts, one may find it detached from them depending on their circumstances. We could automatically create multiple versions of campaigns or ads each of which personally resonates with a particular group for the emotional appeal the copywriter wants to convey.

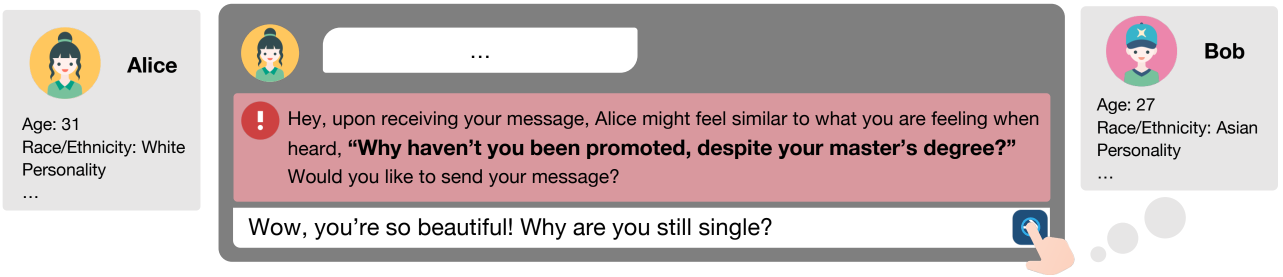
8.3 Extensions beyond *MA*s

The current version of EmoSync serves as a proof-of-concept for generating analogical scenarios, validated through an online survey with respect to the *MA* domain. Our ultimate aim is to integrate EmoSync into a service that facilitates empathy in person-to-person communication by resolving individual differences in understanding diverse situations. This is not limited to a situation evoking negative emotions such as aggression; it may offer an affective resolution to various situations where individual differences collide, such as intercultural or intergenerational disagreement.

As globalization brings people from diverse cultures to live together [14], people often find it difficult to understand each other due to cultural differences. For example, Jake, who was born and raised in United States, might not understand his roommate Minho, who was born and raised in Korea, saying *"Tteokbokki reminds me of my childhood."* In this situation, Jake may better empathize with Minho's nostalgia if EmoSync steps in and nudges Jake saying, *"To Minho, tteokbokki is like peanut butter and jelly sandwiches to you — a comforting reminder of afterschool memories with friends."* EmoSync could also help bridge various conflicts stemming from globalization-driven differences in lifestyles and values, e.g., prioritizing work over sleep [84] or individuals over communities [79].

EmoSync would be useful to bridge intergenerational gaps [74, 75]. Imagine a father who does not understand his daughter saying: *"Watching Netflix is the happiness of my life."* What if EmoSync could step in and say to him, *"You used to be obsessed with cartoons when young. Your daughter loves Netflix just as you did."* It could save them a lot of unnecessary conflict and give them a common ground to initiate a conversation. In similar spirits, EmoSync may help empathizing with children, due to differences in development states [63, 64], individual interests [61, 62, 65], or perception gaps [80, 167].

When extending EmoSync to other domains, it is essential to consider domain-specific prerequisites such as which data to be used as a source and how to mitigate inherent biases embedded in LLMs. First, identifying sources for affective response data tailored to the target domain is a critical consideration. Popular movies or books from different periods or cultures could be leveraged to develop emotion response datasets that capture the unique responses from them. Second, special attention must be given to stereotypes or biases [110, 148] embedded in LLMs. For example, in our work, we replaced the characters' names in the SelfMA dataset with neutral alternatives, i.e., Doe and Foo, as the original names could introduce cultural or gender biases [156]. We also extracted only factual content of the original scenario to minimize bias (Section 3.3). Despite these efforts, some biases still remained. Software developers frequently appeared in the generated examples, leading us to speculate that the use of the name "Foo", which is commonly used in



Age: 31
Race/Ethnicity: White
Personality
...

Hey, upon receiving your message, Alice might feel similar to what you are feeling when heard, **“Why haven't you been promoted, despite your master's degree?”** Would you like to send your message?

Wow, you're so beautiful! Why are you still single?

Age: 27
Race/Ethnicity: Asian
Personality
...

Figure 9: Example usage scenario of EmoSync in chat applications

programming contexts [88], might have introduced bias into the generated content. Several studies have demonstrated that modern LLMs exhibit ageism, such as evaluating *young* more positively than *old* [72]. Additionally, it is well known that most LLMs are heavily trained on English data representing Western perspectives [44, 111]. Directly applying models with such cultural biases in intercultural contexts, or those with age-related biases in intergenerational contexts, might result in inappropriate outputs, potentially worsening the issues at hand. Therefore, it is crucial to review various bias benchmarks [108, 164] relevant to the target domain to identify a fair model or effectively leverage prompting techniques [138] to guide the model toward fairer generation outcomes.

8.4 Extensions beyond Text Modality

While the current EmoSync generates textual content using LLMs, it can be extended to various modalities. With advancement of multimodal generative models [2, 121], high-quality generated content in images, audio, or video, instead of or in addition to text descriptions of a situation, could further facilitate emotional responses from people.

The delivery methods for the generated analogical messages can also be diversified. Earables [36, 83, 123] would enable an unobtrusive personal assistant in face-to-face settings. It could leverage mediums such as AR [145], smart speakers [30, 122], ambient displays [34, 150, 166], or ubiquitous robots [73] to detect and assist in various situations where empathy is needed in everyday life.

9 DISCUSSION

9.1 Limitations in Experimental Setup

Our research is limited in methods as the studies were conducted through online surveys. Although we agree that a direct one-on-one format may better foster empathy, we made it in online surveys for two reasons: (1) given the sensitive nature of our theme, keeping participants anonymous would more likely elicit honest responses [56, 118], and (2) online surveys are advantageous to diversify the participants, which is integral to signify the *MA*-unique challenges. Due to our institution's geography, in-person participants would be of limited diversity in their demographics.

One may see the volume of our study (3-phased study with 101 individuals, where each participant responded to a total of 40 – 64 *MA* samples) is small for online surveys. We clarify that our survey volume was carefully tuned following the ethical considerations (detailed in §3.4). As we split the survey into multiple days to regulate participants' mental workload, we encountered a considerable number of drop-outs in the middle which rendered the earlier responses and expenses sunk. The total expense for the 3-phased study was more than $8,700, and this will grow proportionally with more recruited. We believe our study showcased a premiere of our concept empirically despite practical costs. We anticipate that this study may shed light on justifying the investment for larger studies in the future.

9.2 Disparity between Concept and Implementation

Another limitation is the disparity between concept and implementation, as some participants pointed out the irrelevance of Analogical *MA*s to themselves or the dissimilarity between Original *MA*s and Analogical *MA*s (§7.2.4). In the following sections, we examine the possible causes of this issue and suggest practical guidelines to mitigate them.

9.2.1 *Lack of information.*

Fundamentally, the issue may lie in the lack of sufficient personal information necessary to fully understand and interpret the nuanced emotional responses of individuals. This constraint was unavoidable due to (1) the method of self-reporting and (2) the token limit of LLM. To mitigate the former, EmoSync may interwork with pervasive sensing systems [100, 162], expanding the information pool related to one's emotional reactions in daily life. The latter would be a transient issue as the context window of LLM is growing. We highlight that, despite the currently limited implementation of EmoSync, it has demonstrated multiple promising findings and many participants testified its efficacy for empathy.

9.2.2 *LLM Hallucinations.*

Another issue might arise from the inherent limitations of LLMs, particularly the phenomenon known as “hallucinations,” where LLMs produce responses that are unfaithful to their source [67]. While active research aims to mitigate hallucinations, the criteria for identifying them vary across tasks, necessitating further exploration in diverse applications.

EmoSync must balance two key aspects: leveraging sufficient imaginative capability to create diverse Analogical *MA*s based on limited information, while minimizing the risk of hallucinations that might lead to incorrect analysis. This balance is particularly difficult to achieve, as imagination and hallucination are inherently correlated in LLMs [45]. To address this, we adopted a rigorous human validation process throughout the iterative prompt design in **Phase 1** and the pilot experiment in **Phase 2**. Although most of our

Table 8: Major Types of Hallucinations

Type	Description	Example
Wrong analysis of emotional patterns	Incorrectly interpreting participants' emotional response patterns based on **EmoMA**.	- Mistaking emotion categories for Big5 traits, *e.g. "(...) higher scores in agreeableness and conscientiousness."* - Concluding that the participant has a generally moderate emotional tendency based only on low-emotional responses among diverse samples. - Concluding that 'worried' scores were generally high, despite the majority being low.
Mismatched rationales and Analogical *MA*	Analogical *MA* is unrelated to the context information analysis or conflicts with it.	- An *MA* incorporating racial bias was analyzed as effective, but the generated Analogical *MA* was entirely unrelated (e.g., reflecting bias toward a *'technical field'* background). - Predicted irrelevance to social situations based on low extroversion traits, but the Analogical *MA* described a social gathering scenario.
Unexpected context injected to Analogical *MA*s	Adding specific contextual details unforeseen in the given context information for Analogical *MA*.	- Introducing unforeseen details like *'Law firm'* or *'Marketing strategy'* to depict *MA*s ignoring expertise.
Unnatural or illogical Analogical *MA*s	Analogical *MA* itself is unnatural or lacks logical coherence between events.	- Unnatural and blatant situations, e.g., a colleague stealing credit at work without any context, or overtly racist remarks made during a business meeting. - Situations involving ethnic prejudice at gatherings of people from the same ethnic background.

results showed reasonable Analogical *MA*s, it is imperative to analyze the impact of hallucinations on system usability as completely eliminating them still remains unlikely.

We identified four types of hallucinations by analyzing LLM-generated reasoning on analogy-creation processes (Table 8). Among these, two had a notable tendency to negatively influence the perception of participants. *Wrong analysis of participants' emotional response patterns* occasionally happened due to the high complexity and dimensionality of this task. This led to issues in generating Analogical *MA* that evokes emotion similar to the Original *MA*, resulting in lower Perceived Similarity and Empathic Aid among participants. *Unnatural or illogical Analogical MAs* occurred rarely, but they were directly noticeable to participants and negatively impacted their overall perception of the system.

The other two types of hallucinations showed mixed effects. Interestingly, the *Unexpected context injected into Analogical MAs* influenced participants' Personal Resonance positively or negatively depending on the situations. For example, in some cases, additional context made the Analogical *MA* feel like a "lived experience," enhancing Personal Resonance. In others, the lack of personal relevance, such as participation in a "professional development workshop," led to reduced Personal Resonance.

These issues may have been partly attributable to the way we conducted the experiment. To facilitate efficient iterative testing, we employed one-shot inference for the end-to-end steps for the Analogical *MA* generations, i.e., analysis of context information (**EmoMA**, **Big5**, **VLQ**, **Demo** and target emotion) and creation of Analogical *MA*s. However, this approach inherently resulted in lengthy input prompts and responses, which might exacerbate the risk of information mixing or internal conflicts in the outputs. Additionally, the single-prompt design enforces the use of a fixed set of hyperparameters across successive steps. This likely constrained the performance of individual steps, occasionally resulting in inaccurate reasoning or unnatural Analogical *MA*s.

To mitigate these hallucinations in real-world applications, adopting task-specific prompts, as applied in recent studies [47, 115], could be beneficial. For example, setting lower temperature for robust information interpretation tasks while setting it higher for tasks generating diverse Analogical *MA*s. Additionally, instead of generating arbitrary context, leveraging Retrieval-Augmented Generation [50] could enhance personal resonance by incorporating user-relevant details.

While these measures may not entirely eliminate hallucinations, they align with ongoing research in this field. Future research could explore advanced prompting techniques to enhance system stability and reliability.

9.3 Potential Adversarial Concerns

Emotional Impacts. Despite the participants' high appreciation of resonance, we acknowledge the possibly hurtful effects of EmoSync due to its nature of creating a *MA*. To minimize this, we designed the situations from a third party's (Doe) perspective for the prototype. Participants' responses indicate very few felt attacked on them. However, when EmoSync becomes a real service, there should be a reliable moderation logic to prevent a traumatizing or overly aggressive situation. For example, the LLM would provide a fuller explanation with a rationale generated, e.g., this is not a direct attack but rather a way to persuade by an analogical situation.

Privacy issue needs to be addressed carefully for real-world applications of EmoSync. In order to create personalized Analogical *MA*s, it needs to understand which information an individual reacts sensitively — highly sensitive private information. Therefore, we should take measures to protect this information by ensuring that it is only used internally within the system when needed for creating *MA*, and to prevent it from being exploited in any harmful manner.

Possible Misunderstandings of issues. Generating an Analogical *MA* often changes the nature of stereotype or discrimination in the Original *MA*, e.g., a gender stereotype into a racial one. We clarify that such a change does not (and should not) mean that sexism and racism are the same. Despite analogies being routinely practiced in our daily lives [17], we do not mean two problems are equal, either. The situations around *MA* are very nuanced, and the feelings around them often do not flow as-is through different individuals.

Our rationale in taking an analogy is to acknowledge the difference but open a canal to convey affective feelings between individuals. We will discuss the detail in §9.4.

9.4 Critical Framing of EmoSync in HCI Empathy Research

Empathy has been regarded as a crucial element in the HCI domain, facilitating the understanding of others' experiences and enabling more human-centered design. A large body of these studies has focused on experiential routes of inducing empathy, such as "putting oneself in another's shoes" [51]. While the effectiveness of such methods has been validated by numerous studies [13, 16, 59, 82], debates on their limitations have recently emerged.

A critical concern is that empathy-inducing technologies through conveying others' experiences might inadvertently exclude their lived experiences. This concern is particularly prominent in empathy for marginalized or disabled communities. Bennett and Rosner [20] criticized empathy practices such as disability simulation techniques, arguing that designers often focus on their own indirectly experienced perspectives rather than the actual lived experiences of individuals with disabilities. This phenomenon, referred to as the "empathy trap [163]," underscores the inherent limitation that the presented experiences are inevitably filtered through the interpreter's own experiences and thoughts [38, 87]. This could lead users to oversimplify or misunderstand the target's experiences.

In this light, EmoSync aimed to tackle these challenges by focusing on the "shared emotional commonalities" between the target and the empathizer. Besides presenting the targets' experiences as they are, EmoSync provides personalized analogical experiences that reflect the emotional reactions of targets. This analogy-based approach may help users not misunderstand or misjudge other's feelings solely through their own perspectives. Nevertheless, given that EmoSync is founded on the studies of empathy-building through experiential methodologies, it is unlikely to avoid the aforementioned critiques entirely. For instance, emphasizing similarities between two distinct experiences risks creating the illusion that the two experiences are identical. This oversimplification may trivialize the issue or reinforce stereotypes. Additionally, there is a need to acknowledge that analogy-based empathy might bypass or diminish the process of thoroughly understanding the other's experiences by themselves. Therefore, when using EmoSync's analogy-based approach to foster empathy, it is important to ensure that the essence of the original experience remains intact. Balancing the visibility of the analogical and original experiences can help users identify commonalities while also appreciating the differences. A previous study aimed at enhancing intergenerational communication found that juxtaposing two semantically symmetric photos naturally facilitated not only recognizing commonalities but also contrasting differences [74]. Similarly, our participants were able to discern both commonalities and differences through the juxtaposition of two experiences. This highlights the need for careful design in presenting dual experiences when applying EmoSync in real-world contexts.

The HCI community's perspective on empathy emphasizes approaching it through ongoing presence and engagement, rather than treating empathy as a standalone goal to be achieved [20]. This requires continuous mutual understanding and empathy among stakeholders, but many practical challenges remain. As observed in Choi et al.'s study [35], the mental and emotional burden placed on participants in empathy-assistive systems can hinder empathic engagement. This issue calls for a system that lowers the barriers to empathy, ultimately fostering reciprocal communication and a deeper understanding of each other's experiences. Additionally, recent research by Lee et al. pointed out that emotional empathy alone is insufficient for fostering long-term awareness or behavioral change [86]. To guide users toward prosocial behavior, a combination of emotional changes and deeper cognitive understanding is required. However, there was a claim that studies based on immersive simulations—widely used in experiential methodologies—have limited effectiveness in promoting cognitive empathy, even though they excel at eliciting affective empathy [98].

We believe that EmoSync plays a role in addressing these challenges to some extent. First, as validated by prior research [74], using analogies rooted in familiar experiences naturally evokes interest and curiosity, encouraging active engagement with the target subject. Moreover, as supported by our qualitative results, empathy facilitated through analogy inherently involves both cognitive understanding and emotional resonance [17]. This suggests that empathy induced by EmoSync would have the potential to further guide users to prosocial behavior, which is evident in the improved "helping" scores observed in **Phase 3**.

Overall, EmoSync demonstrated that analogy-based empathy could effectively induce empathy for others' experiences in the context of *MA*s. By proposing a novel approach that aligns with the HCI community's goals for empathy-assistive systems while addressing existing challenges, we showcased a pathway forward. Still, there are challenges left to resolve, such as the risk of oversimplifying or bypassing the original experience. Future research should investigate the impacts of EmoSync through real-world applications and longitudinal studies to assess its effectiveness and refine its approach.

10 CONCLUSION

Our feelings upon the same experiences can vary due to individual differences. Previous works that aimed to foster empathy often immerse individuals in an experience identical to another's, overlooking the intricacies of personal differences and possibly limiting affective empathy. In this paper, we proposed a novel concept toward affective empathy by creating personalized analogies. Then, we embodied our concept as EmoSync, an LLM-based agent generating bespoke analogical vignettes in a context of *MA*. We designed and evaluated it through an extensive 3-phased study with 100+ individuals from diverse backgrounds. We reported multi-faceted findings and implications.

ACKNOWLEDGMENTS

We sincerely thank our research interns, Seungwoo Yoon and Jungin Jang, for their valuable assistance with the literature review and data annotation. This research was supported by the National Research Foundation of Korea(NRF) grants funded by the Korea government(MSIT) (No. 2021R1A2C200386612, RS-2024-00451947), and partly supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-ITRC(Information Technology Research Center) grants funded by the Korea government(MSIT) (IITP-2025-RS-2020-II201778, RS-2024-00437866).

REFERENCES

[1] 2021. Microaggressions. Retrieved Mar. 23, 2024 from https://www.microaggressions.com/

[2] 2023. Stable Video. Retrieved 2023-12-06 from https://stability.ai/stable-video

[3] 2024. Everyday Discrimination Scale. Retrieved Mar. 23, 2024 from https://scholar.harvard.edu/davidrwilliams/node/32397

[4] 2024. Llama-2-70b. Retrieved Mar. 28, 2024 from https://huggingface.co/meta-llama/Llama-2-70b

[5] 2024. Mixtral-8x7B-Instruct-v0.1. Retrieved Mar. 28, 2024 from https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

[6] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[7] Kupiri Ackerman-Barger and Negar Nicole Jacobs. 2020. The microaggressions triangle model: a humanistic approach to navigating microaggressions in health professions schools. *Academic Medicine* 95, 12S (2020), S28–S32.

[8] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*. PMLR, 337–371.

[9] Tanja Aitamurto, Shuo Zhou, Sukolsak Sakshuwong, Jorge Saldivar, Yasamin Sadeghi, and Amy Tran. 2018. Sense of Presence, Attitude Change, Perspective-Taking and Usability in First-Person Split-Sphere 360° Video. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3174119

[10] Omar Ali, Nancy Scheidt, Alexander Gegov, Ella Haig, Mo Adda, and Benjamin Aziz. 2020. Automated detection of racial microaggressions using machine learning. In *2020 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2477–2484.

[11] Anthropic. 2024. Claude 3.5 Sonnet. Retrieved Aug. 12, 2024 from https://www.anthropic.com/news/claude-3-5-sonnet

[12] Byung-Chull Bae, Su-ji Jang, Duck-Ki Ahn, and Gapyuel Seo. 2019. A vr interactive story using pov and flashback for empathy. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 844–845.

[13] Madeline Balaam, Rob Comber, Rachel E Clarke, Charles Windlin, Anna Ståhl, Kristina Höök, and Geraldine Fitzpatrick. 2019. Emotion work in experience-centered design. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[14] Claudio Baraldi. 2006. New forms of intercultural communication in a globalized world. *International Communication Gazette* 68, 1 (2006), 53–69.

[15] S Barber, PC Gronholm, S Ahuja, N Rüsch, and G Thornicroft. 2020. Microaggressions towards people affected by mental health problems: a scoping review. *Epidemiology and Psychiatric Sciences* 29 (2020), e82.

[16] Baptiste Barbot and James C Kaufman. 2020. What makes immersive virtual reality the ultimate empathy machine? Discerning the underlying mechanisms of change. *Computers in Human Behavior* 111 (2020), 106431.

[17] Allison Barnes and Paul Thagard. 1997. Empathy and analogy. *Dialogue: Canadian Philosophical Review/Revue canadienne de philosophie* 36, 4 (1997), 705–720.

[18] Tessa E Basford, Lynn R Offermann, and Tara S Behrend. 2014. Do you see what I see? Perceptions of gender microaggressions in the workplace. *Psychology of Women Quarterly* 38, 3 (2014), 340–349.

[19] Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-defined AI personas for on-demand feedback generation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[20] Cynthia L. Bennett and Daniela K. Rosner. 2019. The Promise of Empathy: Design, Disability, and Knowing the "Other". In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300528

[21] Chris Bevan, David Philip Green, Harry Farmer, Mandy Rose, Kirsten Cater, Danaë Stanton Fraser, and Helen Brown. 2019. Behind the Curtain of the "Ultimate Empathy Machine": On the Composition of Virtual Reality Nonfiction Experiences. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300736

[22] James Blair, Derek Mitchell, and Karina Blair. 2005. *The psychopath: Emotion and the brain.* Blackwell Publishing.

[23] Johanne Boisjoly, Greg J Duncan, Michael Kremer, Dan M Levy, and Jacque Eccles. 2006. Empathy or Antipathy? The Impact of Diversity. *American Economic Review* 96, 5 (Nov. 2006), 1890–1905. https://doi.org/10.1257/aer.96.5.1890

[24] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 1664–1674.

[25] Carolyn Calloway-Thomas. 2010. *Empathy in the global world: An intercultural perspective.* Sage.

[26] Turhan Canli, Zuo Zhao, John E. Desmond, Eunjoo Kang, James Gross, and John D. E. Gabrieli. 2001. An fMRI study of personality influences on brain reactivity to emotional stimuli. *Behavioral Neuroscience* 115, 1 (2001), 33–42. https://doi.org/10.1037/0735-7044.115.1.33

[27] Laura L Carstensen, Bulent Turan, Susanne Scheibe, Nilam Ram, Hal Ersner-Hershfield, Gregory R Samanez-Larkin, Kathryn P Brooks, and John R Nesselroade. 2011. Emotional experience improves with age: evidence based on over 10 years of experience sampling. *Psychology and aging* 26, 1 (2011), 21.

[28] Charles S. Carver, Steven K. Sutton, and Michael F. Scheier. 2000. Action, Emotion, and Personality: Emerging Conceptual Integration. *Personality and Social Psychology Bulletin* 26, 6 (Aug. 2000), 741–751. https://doi.org/10.1177/0146167200268008

[29] Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *Journal of medical systems* 47, 1 (2023), 33.

[30] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. 2020. Hello there! is now a good time to talk? opportune moments for proactive interactions with smart speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–28.

[31] Michael Chapman, Carolyn Zahn-Waxler, Geri Cooperman, and Ronald Iannotti. 1987. Empathy and responsibility in the motivation of children's helping. *Developmental Psychology* 23, 1 (Jan. 1987), 140–145. https://doi.org/10.1037/0012-1649.23.1.140

[32] Arnav Chavan, Raghav Magazine, Shubham Kushwaha, Mérouane Debbah, and Deepak Gupta. 2024. Faster and Lighter LLMs: A Survey on Current Challenges and Way Forward. *arXiv preprint arXiv:2402.01799* (2024).

[33] Zheng Chen. 2023. Palr: Personalization aware llms for recommendation. *arXiv preprint arXiv:2305.07622* (2023).

[34] Sungjae Cho, Yoonsu Kim, Jaewoong Jang, and Inseok Hwang. 2023. AI-to-Human Actuation: Boosting Unmodified AI's Robustness by Proactively Inducing Favorable Human Sensing Conditions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–32.

[35] Ryuhaerang Choi, Chanwoo Yun, Hyunsung Cho, Hwajung Hong, Uichin Lee, and Sung-Ju Lee. 2022. You Are Not Alone: How Trending Stress Topics Brought #Awareness and #Resonance on Campus. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–30. https://doi.org/10.1145/3555612

[36] Romit Roy Choudhury. 2021. Earable computing: A new area to think about. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*. 147–153.

[37] Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang. 2024. Comprehensive assessment of jailbreak attacks against llms. *arXiv preprint arXiv:2402.05668* (2024).

[38] Laura D Cosio, Oğuz "Oz" Buruk, Daniel Fernández Galeote, Isak De Villiers Bosman, and Juho Hamari. 2023. Virtual and Augmented Reality for Environmental Sustainability: A Systematic Review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–23. https://doi.org/10.1145/3544548.3581147

[39] Jennifer Crocker, Kristin Voelkl, Maria Testa, and Brenda Major. [n. d.]. Social Stigma: The Affective Consequences of Attributional Ambiguity. ([n. d.]).

[40] Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. 2016. Empathy: A review of the concept. *Emotion review* 8, 2 (2016), 144–153.

[41] Max T Curran, Jeremy Raboff Gordon, Lily Lin, Priyashri Kamlesh Sridhar, and John Chuang. 2019. Understanding digitally-mediated empathy: An exploration of visual, narrative, and biosensory informational cues. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[42] Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy. (1980).

[43] Changming Duan and Clara E Hill. 1996. The current state of empathy research. *Journal of counseling psychology* 43, 3 (1996), 261.

[44] Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388* (2023).

[45] Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online

and Punta Cana, Dominican Republic, 2197–2214. https://doi.org/10.18653/v1/2021.emnlp-main.168

[46] Nancy Eisenberg and Paul A. Miller. 1987. The relation of empathy to prosocial and related behaviors. *Psychological Bulletin* 101, 1 (1987), 91–119. https://doi.org/10.1037/0033-2909.101.1.91

[47] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 14664–14690. https://doi.org/10.18653/v1/2024.acl-long.786

[48] Agneta H Fischer, Antony SR Manstead, and Ruud Zaalberg. 2003. Social influences on the emotion process. *European review of social psychology* 14, 1 (2003), 171–201.

[49] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–30.

[50] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs.CL] https://arxiv.org/abs/2312.10997

[51] Uğur Genç and Himanshu Verma. 2024. Situating Empathy in HCI/CSCW: A Scoping Review. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (Nov. 2024), 1–37. https://doi.org/10.1145/3687052

[52] Lewis R Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological assessment* 4, 1 (1992), 26.

[53] Lauren Gonzales, Kristin C Davidoff, Kevin L Nadal, and Philip T Yanos. 2015. Microaggressions experienced by persons with mental illnesses: An exploratory study. *Psychiatric rehabilitation journal* 38, 3 (2015), 234.

[54] Janelle R Goodwill, Robert Joseph Taylor, and Daphne C Watkins. 2021. Everyday discrimination, depressive symptoms, and suicide ideation among African American men. *Archives of suicide research* 25, 1 (2021), 74–93.

[55] Eric Gordon and Steven Schirra. 2011. Playing with empathy: digital role-playing games in public meetings. In *Proceedings of the 5th International Conference on Communities and Technologies*. 179–185.

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

[57] Katharine H Greenaway, Elise K Kalokerinos, and Lisa A Williams. 2018. Context is everything (in emotion research). *Social and Personality Psychology Compass* 12, 6 (2018), e12393.

[58] Sara A Heimpel, Joanne V Wood, Margaret A Marshall, and Jonathon D Brown. 2002. Do people with low self-esteem really want to feel better? Self-esteem differences in motivation to repair negative moods. *Journal of personality and social psychology* 82, 1 (2002), 128.

[59] Fernanda Herrera, Jeremy Bailenson, Erika Weisz, Elise Ogle, and Jamil Zaki. 2018. Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking. *PloS one* 13, 10 (2018), e0204494.

[60] Steven Hitlin, J. Scott Brown, and Glen H. Elder. 2006. Racial Self-Categorization in Adolescence: Multiracial Development and Social Pathways. *Child Development* 77, 5 (Sept. 2006), 1298–1308. https://doi.org/10.1111/j.1467-8624.2006.00935.x

[61] Inseok Hwang, Hyukjae Jang, Lama Nachman, and Junehwa Song. 2010. Exploring inter-child behavioral relativity in a shared social environment: a field study in a kindergarten. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 271–280.

[62] Inseok Hwang, Hyukjae Jang, Taiwoo Park, Aram Choi, Youngki Lee, Chanyou Hwang, Yanggui Choi, Lama Nachman, and Junehwa Song. 2012. Leveraging children's behavioral distribution and singularities in new interactive environments: Study in kindergarten field trips. In *Pervasive Computing: 10th International Conference, Pervasive 2012, Newcastle, UK, June 18-22, 2012. Proceedings 10*. Springer, 39–56.

[63] Inseok Hwang, Youngki Lee, Chungkuk Yoo, Chulhong Min, Dongsun Yim, and John Kim. 2019. Towards interpersonal assistants: next-generation conversational agents. *IEEE Pervasive Computing* 18, 2 (2019), 21–31.

[64] Inseok Hwang, Chungkuk Yoo, Chanyou Hwang, Dongsun Yim, Youngki Lee, Chulhong Min, John Kim, and Junehwa Song. 2014. TalkBetter: family-driven mobile intervention care for children with language delay. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1283–1296.

[65] Hyukjae Jang, Sungwon Peter Choe, Inseok Hwang, Chanyou Hwang, Lama Nachman, and Junehwa Song. 2012. RubberBand: augmenting teacher's awareness of spatially isolated children on kindergarten field trips. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 236–239.

[66] Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies* 28, 12 (2023), 15873–15892.

[67] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (Dec. 2023), 1–38. https://doi.org/10.1145/3571730

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

[69] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. arXiv:2401.04088 [cs.LG] https://arxiv.org/abs/2401.04088

[70] Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology* (1991).

[71] Kristen P Jones, Chad I Peddie, Veronica L Gilrane, Eden B King, and Alexis L Gray. 2016. Not so subtle: A meta-analytic investigation of the correlates of subtle and overt discrimination. *Journal of management* 42, 6 (2016), 1588–1613.

[72] Mahammed Kamruzzaman, Md Minul Islam Shovon, and Gene Louis Kim. 2023. Investigating subtler biases in llms: Ageism, beauty, institutional, and nationality bias in generative models. *arXiv preprint arXiv:2309.08902* (2023).

[73] Bumsoo Kang, Inseok Hwang, Jinho Lee, Seungchul Lee, Taegyeong Lee, Youngjae Chang, and Min Kyung Lee. 2018. My being to your place, your being to my place: Co-present robotic avatars create illusion of living together. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 54–67.

[74] Bumsoo Kang, Seungwoo Kang, and Inseok Hwang. 2021. MomentMeld: Ai-augmented mobile photographic memento towards mutually stimulatory intergenerational interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[75] Bumsoo Kang, Seungwoo Kang, and Inseok Hwang. 2023. AI-driven Family Interaction Over Melded Space and Time. *IEEE Pervasive Computing* 22, 1 (2023), 85–94.

[76] Bumsoo Kang, Sujin Lee, Alice Oh, Seungwoo Kang, Inseok Hwang, and Junehwa Song. 2015. Towards understanding relational orientation: attachment theory and facebook activities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 1404–1415.

[77] Shanna K Kattari. 2020. Ableist microaggressions and the mental health of disabled adults. *Community Mental Health Journal* 56, 6 (2020), 1170–1179.

[78] Jaechang Kim, Jinmin Goh, Inseok Hwang, Jaewoong Cho, and Jungseul Ok. 2024. Bridging the Gap between Expert and Language Models: Concept-guided Chess Commentary Generation and Evaluation. *arXiv preprint arXiv:2410.20811* (2024).

[79] Wonjung Kim, Seungchul Lee, Youngjae Chang, Taegyeong Lee, Inseok Hwang, and Junehwa Song. 2021. Hivemind: social control-and-use of IoT towards democratization of public spaces. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 467–482.

[80] Wonjung Kim, Seungchul Lee, Seonghoon Kim, Sungbin Jo, Chungkuk Yoo, Inseok Hwang, Seungwoo Kang, and Junehwa Song. 2020. Dyadic mirror: Everyday second-person live-view for empathetic reflection upon parent-child interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.

[81] Martijn J.L. Kors, Gabriele Ferri, Erik D. Van Der Spek, Cas Ketel, and Ben A.M. Schouten. 2016. A Breathtaking Journey. On the Design of an Empathy-Arousing Mixed-Reality Game. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. ACM, Austin Texas USA, 91–104. https://doi.org/10.1145/2967934.2968110

[82] Martijn J.L. Kors, Erik D. van der Spek, Julia A. Bopp, Karel Millenaar, Rutger L. van Teutem, Gabriele Ferri, and Ben A.M. Schouten. 2020. The Curious Case of the Transdiegetic Cow, or a Mission to Foster Other-Oriented Empathy Through Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376748

[83] Haechan Lee, Miri Moon, Taiwoo Park, Inseok Hwang, Uichin Lee, and Junehwa Song. 2013. Dungeons & swimmers: designing an interactive exergame for swimming. In *Proceedings of the 2013 ACM conference on Pervasive and Ubiquitous Computing adjunct publication*. 287–290.

[84] Jungeun Lee, Sungnam Kim, Minki Cheon, Hyojin Ju, JaeEun Lee, and Inseok Hwang. 2022. SleepGuru: Personalized Sleep Planning System for Real-life Actionability and Negotiability. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

[85] Jungeun Lee, Suwon Yoon, Kyoosik Lee, Eunae Jeong, Jae-Eun Cho, Wonjeong Park, Dongsun Yim, and Inseok Hwang. 2024. Open Sesame? Open Salami! Personalizing Vocabulary Assessment-Intervention for Children via Pervasive Profiling and Bespoke Storybook Generation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–32.

[86] Ken Jen Lee, Adrian Davila, Hanlin Cheng, Joslin Goh, Elizabeth Nilsen, and Edith Law. 2023. "We need to do more... I need to do more": Augmenting

Digital Media Consumption via Critical Reflection to Increase Compassion and Promote Prosocial Attitudes and Behaviors. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–20. https://doi.org/10.1145/3544548.3581355

[87] Yeoreum Lee, Youn-kyung Lim, and Hyeon-Jeong Suk. 2011. Altruistic interaction design: a new interaction design approach for making people care more about others. In *Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces*. ACM, Milano Italy, 1–4. https://doi.org/10.1145/2347504.2347514

[88] Brian Lennon. 2021. Foo, Bar, Baz...: The Metasyntactic Variable and the Programming Language Hierarchy. *Philosophy & Technology* 34, 1 (2021), 13–32.

[89] Colleen Lewis. 2020. New from csteachingtips. org: microaggressions: the game! *ACM SIGCSE Bulletin* 52, 1 (2020), 10–10.

[90] Jioni A Lewis, Ruby Mendenhall, Stacy A Harwood, and Margaret Browne Huntt. 2013. Coping with gendered racial microaggressions among Black women college students. *Journal of African American Studies* 17 (2013), 51–73.

[91] Alex Lindsey, Eden King, Michelle Hebl, and Noah Levine. 2015. The Impact of Method, Motivation, and Empathy on Diversity Training Effectiveness. *Journal of Business and Psychology* 30, 3 (Sept. 2015), 605–617. https://doi.org/10.1007/s10869-014-9384-3

[92] Fannie Liu, Laura Dabbish, and Geoff Kaufman. 2017. Can biosignals be expressive? How visualizations affect impression formation from shared brain activity. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–21.

[93] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387* (2021).

[94] Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2023. Check me if you can: Detecting ChatGPT-generated academic writing using CheckGPT. *arXiv preprint arXiv:2306.05524* (2023).

[95] P. Priscilla Lui, Shalanda R. Berkley, Savannah Pham, and Lauren Sanders. 2020. Is microaggression an oxymoron? A mixed methods study on attitudes toward racial microaggressions among United States university students. *PloS One* 15, 12 (2020), e0243058. https://doi.org/10.1371/journal.pone.0243058

[96] Zexin Ma. 2020. Effects of immersive stories on prosocial attitudes and willingness to help: testing psychological mechanisms. *Media Psychology* 23, 6 (Nov. 2020), 865–890. https://doi.org/10.1080/15213269.2019.1651655

[97] Brenda Major, Cheryl R. Kaiser, and Shannon K. McCoy. 2003. It's Not My Fault: When and Why Attributions to Prejudice Protect Self-Esteem. *Personality and Social Psychology Bulletin* 29, 6 (June 2003), 772–781. https://doi.org/10.1177/0146167203029006009

[98] Alison Jane Martingano, Fernanda Hererra, and Sara Konrath. 2021. Virtual reality improves emotional but not cognitive empathy: A meta-analysis. (2021).

[99] Albert Mehrabian and Norman Epstein. 1972. A measure of emotional empathy1. *Journal of Personality* 40, 4 (Dec. 1972), 525–543. https://doi.org/10.1111/j.1467-6494.1972.tb00078.x

[100] Abhinav Mehrotra, Fani Tsapeli, Robert Hendley, and Mirco Musolesi. 2017. MyTraces: Investigating Correlation and Causation between Users' Emotional States and Mobile Phone Interaction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 83 (sep 2017), 21 pages. https://doi.org/10.1145/3130948

[101] Batja Mesquita and Hazel Rose Markus. 2004. Culture and emotion: Models of agency as sources of cultural variation in emotion. In *Feelings and emotions: The Amsterdam symposium*. 341–358.

[102] Clara Moge, Katherine Wang, and Youngjun Cho. 2022. Shared user interfaces of physiological data: Systematic review of social biofeedback systems and contexts in hci. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16.

[103] Kristine M Molina, Tariana V Little, and Milagros C Rosal. 2016. Everyday discrimination, family context, and psychological distress among Latino adults in the United States. *Journal of Community Psychology* 44, 2 (2016), 145–165.

[104] Surya Monro. 2005. Beyond Male and Female: Poststructuralism and the Spectrum of Gender. *International Journal of Transgenderism* 8, 1 (March 2005), 3–22. https://doi.org/10.1300/J485v08n01_02

[105] Erica M Morales. 2014. Intersectional impact: Black students and race, gender and class microaggressions in higher education. *Race, Gender & Class* (2014), 48–66.

[106] Dawne M Mouzon, Robert Joseph Taylor, Amanda Toler Woodward, and Linda M Chatters. 2020. Everyday racial discrimination, everyday non-racial discrimination, and physical health among African-Americans. In *Microaggressions and Social Work Research, Practice and Education*. Routledge, 69–81.

[107] Daphne A Muller, Caro R Van Kessel, and Sam Janssen. 2017. Through Pink and Blue glasses: Designing a dispositional empathy game using gender stereotypes and Virtual Reality. In *Extended abstracts publication of the annual symposium on computer-human interaction in play*. 599–605.

[108] Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. *arXiv preprint arXiv:2406.09948* (2024).

[109] Kevin L Nadal, Marie-Anne Issa, Jayleen Leon, Vanessa Meterko, Michelle Wideman, and Yinglee Wong. 2011. Sexual orientation microaggressions:"Death by a thousand cuts" for lesbian, gay, and bisexual youth. *Journal of LGBT Youth* 8, 3 (2011), 234–259.

[110] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

[111] Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456* (2023).

[112] Suman Nath, Felix Xiaozhu Lin, Lenin Ravindranath, and Jitendra Padhye. 2013. SmartAds: bringing contextual ads to mobile apps. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. 111–124.

[113] Susan Nolen-Hoeksema and Amelia Aldao. 2011. Gender and age differences in emotion regulation strategies and their relationship to depressive symptoms. *Personality and individual differences* 51, 6 (2011), 704–708.

[114] Elizabeth A O'Neill, Kate Trout, and Virginia Ramseyer Winter. 2023. Relationships between experiencing anti-fat microaggressions, body appreciation, and perceived physical and mental health. *Journal of Health Psychology* 28, 2 (2023), 107–118.

[115] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[116] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[117] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 443–452.

[118] Delroy L Paulhus. 1984. Two-component models of socially desirable responding. *Journal of personality and social psychology* 46, 3 (1984), 598.

[119] Vyjeyanthi S Periyakoil, Linda Chaudron, Emorcia V Hill, Vincent Pellegrini, Eric Neri, and Helena C Kraemer. 2020. Common types of gender-based microaggressions in medicine. *Academic Medicine* 95, 3 (2020), 450–457.

[120] Chester Pierce. 1970. Offensive mechanisms. *The black seventies* (1970), 265–282.

[121] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV]

[122] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[123] Meera Radhakrishnan, Darshana Rathnayake, Ong Koon Han, Inseok Hwang, and Archan Misra. 2020. ERICA: enabling real-time mistake detection & corrective feedback for free-weights exercises. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 558–571.

[124] Ciaran Regan, Nanami Iwahashi, Shogo Tanaka, and Mizuki Oka. 2024. Can Generative Agents Predict Emotion? *arXiv preprint arXiv:2402.04232* (2024).

[125] Sonia Roccas and Marilynn B Brewer. 2002. Social identity complexity. *Personality and social psychology review* 6, 2 (2002), 88–106.

[126] Lilach Sagiv, Sonia Roccas, Jan Cieciuch, and Shalom H Schwartz. 2017. Personal values in human life. *Nature human behaviour* 1, 9 (2017), 630–639.

[127] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*. 1–10.

[128] Michael T. Schmitt and Nyla R. Branscombe. 2002. The Internal and External Causal Loci of Attributions to Prejudice. *Personality and Social Psychology Bulletin* 28, 5 (May 2002), 620–628. https://doi.org/10.1177/0146167202288006

[129] Kimber Shelton and Edward A Delgado-Romero. 2013. Sexual orientation microaggressions: the experience of lesbian, gay, bisexual, and queer clients in psychotherapy. (2013).

[130] Iain A Smith and Amanda Griffiths. 2022. Microaggressions, everyday discrimination, workplace incivilities, and other subtle slights at work: A meta-synthesis. *Human Resource Development Review* 21, 3 (2022), 275–299.

[131] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III, Jesse Dodge, Isabella Duan, et al. 2023. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949* (2023).

[132] Anselm Strauss and Juliet Corbin. 1998. Basics of qualitative research techniques. (1998).

[133] Derald Wing Sue, Sarah Alsaidi, Michael N Awad, Elizabeth Glaeser, Cassandra Z Calle, and Narolyn Mendez. 2019. Disarming racial microaggressions: Microintervention strategies for targets, White allies, and bystanders. *American Psychologist* 74, 1 (2019), 128.

[134] Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. 2007. Racial microaggressions in everyday life: implications for clinical practice. *American psychologist* 62, 4 (2007), 271.

[135] Derald Wing Sue and Lisa Spanierman. 2020. *Microaggressions in everyday life.* John Wiley & Sons.

[136] Zhaoxuan Tan and Meng Jiang. 2023. User Modeling in the Era of Large Language Models: Current Research and Future Directions. *arXiv preprint arXiv:2312.11518* (2023).

[137] Yilin Tang, Liuqing Chen, Ziyu Chen, Wenkai Chen, Yu Cai, Yao Du, Fan Yang, and Lingyun Sun. 2024. EmoEden: Applying Generative Artificial Intelligence to Emotional Learning for Children with High-Function Autism. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 1–20.

[138] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus* 3, 9 (2024), pgae346.

[139] Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N Bazarova. 2019. Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[140] Mistral AI team. 2023. Mixtral of Experts - Performance. Retrieved Dec. 3, 2024 from https://mistral.ai/news/mixtral-of-experts/

[141] Maria Testa and Brenda Major. 1990. The Impact of Social Comparisons After Failure: The Moderating Effects of Perceived Control. *Basic and Applied Social Psychology* 11, 2 (June 1990), 205–218. https://doi.org/10.1207/s15324834basp1102_7

[142] Alexandra To, Hillary Carey, Riya Shrivastava, Jessica Hammer, and Geoff Kaufman. 2022. Interactive fiction provotypes for coping with interpersonal racism. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–14.

[143] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[144] Greg Townley, Bret Kloos, Eric P. Green, and Margarita M. Franco. 2011. Reconcilable Differences? Human Diversity, Cultural Relativity, and Sense of Community. *American Journal of Community Psychology* 47, 1–2 (March 2011), 69–85. https://doi.org/10.1007/s10464-010-9379-9

[145] Andreia Valente, Daniel Simoes Lopes, Nuno Nunes, and Augusto Esteves. 2022. Empathic aurea: exploring the effects of an augmented reality cue for emotional sharing across three face-to-face tasks. In *2022 IEEE conference on virtual reality and 3D user interfaces (VR).* IEEE, 158–166.

[146] Kathleen Van Royen, Karolien Poels, Heidi Vandebosch, and Philippe Adam. 2017. "Thinking before posting?" Reducing cyber harassment on social networking sites through a reflective message. *Computers in human behavior* 66 (2017), 345–352.

[147] Philippe Verduyn and Karen Brans. 2012. The relationship between extraversion, neuroticism and aspects of trait affect. *Personality and Individual Differences* 52, 6 (2012), 664–669.

[148] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219* (2023).

[149] Jennifer Wang, Janxin Leu, and Yuichi Shoda. 2011. When the Seemingly Innocuous "Stings": Racial Microaggressions and Their Emotional Consequences. *Personality and Social Psychology Bulletin* 37, 12 (Dec. 2011), 1666–1678. https://doi.org/10.1177/0146167211416130

[150] Miaosen Wang, Sebastian Boring, and Saul Greenberg. 2012. Proxemic peddler: a public advertising display that captures and preserves the attention of a passerby. In *Proceedings of the 2012 international symposium on pervasive displays.* 1–6.

[151] Wenxiao Wang, Wei Chen, Yicong Luo, Yongliu Long, Zhengkai Lin, Liye Zhang, Binbin Lin, Deng Cai, and Xiaofei He. 2024. Model compression and efficient inference for large language models: A survey. *arXiv preprint arXiv:2402.09748* (2024).

[152] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology* 17 (2023), 18344909231213958.

[153] Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. 2023. Recmind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296* (2023).

[154] Chezare A. Warren. 2014. Towards a Pedagogy for the Application of Empathy in Culturally Diverse Classrooms. *The Urban Review* 46, 3 (Sept. 2014), 395–419. https://doi.org/10.1007/s11256-013-0262-5

[155] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems* 36 (2024).

[156] Yumou Wei, Paulo F Carvalho, and John Stamper. 2024. Uncovering Name-Based Biases in Large Language Models Through Simulated Trust Game. *arXiv preprint arXiv:2404.14682* (2024).

[157] Daniel J Wheeler, Josué Zapata, Denise Davis, and Calvin Chou. 2019. Twelve tips for responding to microaggressions and overt discrimination: when the patient offends the learner. *Medical teacher* 41, 10 (2019), 1112–1117.

[158] David R Williams, Yan Yu, James S Jackson, and Norman B Anderson. 1997. Racial differences in physical and mental health: Socio-economic status, stress and discrimination. *Journal of health psychology* 2, 3 (1997), 335–351.

[159] Kelly G Wilson, Emily K Sandoz, Jennifer Kitchens, and Miguel Roberts. 2010. The Valued Living Questionnaire: Defining and measuring valued action within a behavioral framework. *The Psychological Record* 60 (2010), 249–272.

[160] Shengyao Xiao, Xiaoyu Cui, Yuanqin Fan, Boyuan Lu, Haiyun Wu, Michael Christel, Shirley Saldamarco, and Geoff Kaufman. 2021. Playing through Microaggressions on a College Campus with "Blindspot". In *2021 IEEE Conference on Games (CoG).* IEEE, 1–4.

[161] Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674* (2023).

[162] Zihan Yan, Yufei Wu, Yang Zhang, and Xiang 'Anthony' Chen. 2022. EmoGlass: an End-to-End AI-Enabled Wearable Platform for Enhancing Self-Awareness of Emotional Health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 13, 19 pages. https://doi.org/10.1145/3491102.3501925

[163] Tina Yao, Soojeong Yoo, and Callum Parker. 2021. Evaluating Virtual Reality as a Tool for Empathic Modelling of Vision Impairment: Insights from a simulated public interactive display experience. In *33rd Australian Conference on Human-Computer Interaction.* ACM, Melbourne VIC Australia, 190–197. https://doi.org/10.1145/3520495.3520519

[164] Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. *arXiv preprint arXiv:2205.12247* (2022).

[165] Chungkuk Yoo, Inseok Hwang, Seungwoo Kang, Myung-Chul Kim, Seonghoon Kim, Daeyoung Won, Yu Gu, and Junehwa Song. 2017. Card-stunt as a service: Empowering a massively packed crowd for instant collective expressiveness. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services.* 121–135.

[166] Chungkuk Yoo, Inseok Hwang, Eric Rozner, Yu Gu, and Robert F Dickerson. 2016. Symmetrisense: Enabling near-surface interactivity on glossy surfaces using a single commodity smartphone. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* 5126–5137.

[167] Chungkuk Yoo, Seungwoo Kang, Inseok Hwang, Chulhong Min, Seonghoon Kim, Wonjung Kim, and Junehwa Song. 2019. Mom, I see You Angry at Me! Designing a Mobile Service for Parent-child Conflicts by In-situ Emotional Empathy. In *Proceedings of the 5th ACM Workshop on Mobile Systems for Computational Social Science.* 21–26.

[168] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment Analysis in the Era of Large Language Models: A Reality Check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 3881–3906. https://doi.org/10.18653/v1/2024.findings-naacl.246

[169] Haozhe Zhou, Mayank Goel, and Yuvraj Agarwal. 2024. Bring Privacy To The Table: Interactive Negotiation for Privacy Settings of Shared Sensing Devices. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 1–22.

[170] Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D Hwang, Swabha Swayamdipta, and Maarten Sap. [n. d.]. COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements. ([n. d.]).

[171] Marvin Zuckerman, Bernard Lubin, Lawrence Vogel, and Elizabeth Valerius. 1960. Multiple Affect Adjective Check List. *Journal of Personality and Social Psychology* (1960).

A APPENDIX

A.1 Themes and examples of *MA*s

Table 9: Themes and examples of *MA*s [135]

Theme	*MA* Examples	Implication
Alien in own land	"Where are you from?" "You speak good English." A person asking an Asian American to teach them words in their native language.	You are not American.
Ascription of intelligence	"You are a credit to your race." Asking an Asian person to help with a math or science problem.	People of color are generally not as intelligent as Whites.
Color blindness	"When I look at you, I don't see color."	Denying a person of color's racial/ethnic experiences.
Criminality/assumption of criminal status	A store owner following a customer of color around the store.	You are going to steal.
Denial of individual racism	"I'm not racist. I have several Black friends."	I am immune to racism because I have friends of color.
Myth of meritocracy	"I believe the most qualified person should get the job."	People of color are given extra unfair benefits because of their race.
Pathologizing cultural values/communication styles	Asking a Black person: "Why do you have to be so loud/animated? Just calm down."	Assimilate to dominant culture.
Second-class citizen	Person of color mistaken for a service worker.	People of color are servants to Whites. They couldn't possibly occupy high-status positions.
Environmental microaggressions	A college or university with buildings that are all named after White heterosexual upper class males.	You don't belong here. There is only so far you can go.

A.2 Our Choice of LLM Model: Analysis on Cost and Performance

Since prompt engineering requires many iterations, using a commercial LLM (e.g., `GPT-4`) was expected to incur a huge expense. A single iteration of inference requires 246 API calls (= 41 participants × 6 combinations). Given a single call consuming 8k tokens including input and output, 1.968M tokens per iteration (= $8k \times 246$), and the `GPT-4` API charging $30 per 1M input tokens (or $60 per 1M output tokens), a single iteration is estimated to cost about $100. As prompt engineering often involves tens to hundreds of iterations, the total estimated cost was impractical. For alternatives, we searched for an open-source LLM that fit in our local GPU server (AMD EPYC 7513 CPU, 512 GB main memory, 4× GPUs of Nvidia 6000 Ada with 48 GB memory each), where `Llama2-13B` [4] and `Mixtral-8x7B-Instruct-v0.1` [5] fit well. Unfortunately, Llama2 refused the task with *MA*s due to its moderation policy (*"It's not appropriate to make assumptions about someone's identity based on their race..."*). As Mixtral accepted and performed our task well, we compared its performance with GPT-4. Table 10 lists the parameters and $mean\ AE_{item}$ from one of our earlier prompts, showing Mixtral is quite comparable to GPT-4 in inference accuracy. We used the Mixtral model throughout our whole study.

Table 10: LLMs performance and parameter setting

LLM	$mean\ AE_{item}$	hyperparams.
GPT-4	1.227	temperature=0.2
Mixtral-8x7B	1.249	top_p=1

A.3 Participants Information

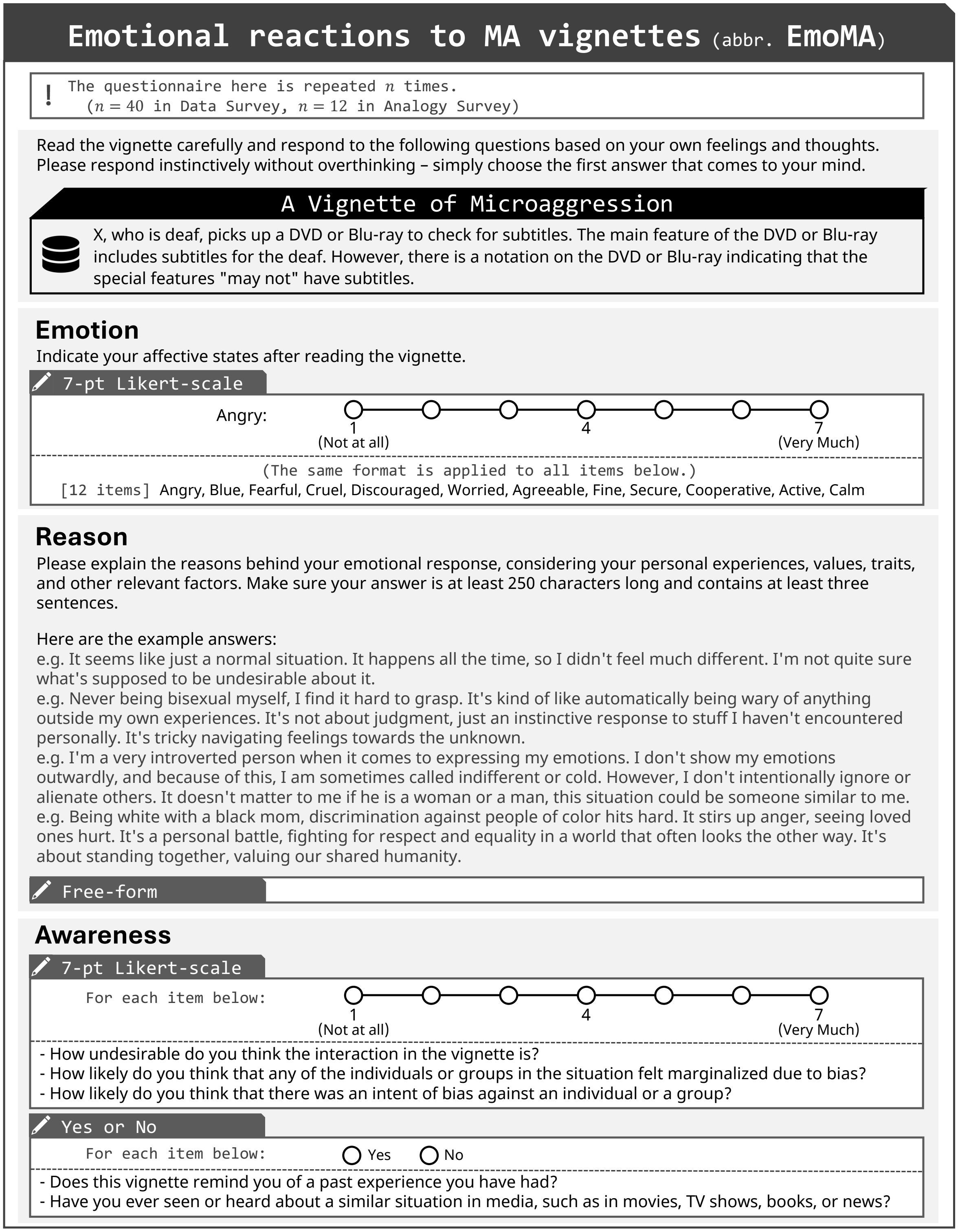
Table 11: The distributions of the participants' demographics & attributes in Phases 1 and 2 (Design)

Demo	Gender, *n*		Physical attributes, mean ± SD		**Big5**, mean ± SD	
	Female	21	Height (cm)	170.34 ± 10.86	Extroversion	2.84 ± 1.06
	Male	20	Weight (kg)	79.35 ± 29.25	Agreeableness	3.94 ± 0.77
	Non-binary	0	Highest level of education, *n*		Conscientiousness	3.95 ± 0.77
	Age, *n*		High school or equivalent	6	Neuroticism	2.81 ± 0.97
	18-24	5	Some college or vocational training	6	Openness	3.74 ± 0.77
	25-34	18	Bachelor's degree	18	**VLQ**, mean ± SD	
	35-44	10	Master's degree	9		7.25 ± 1.56
	45-54	4	Doctoral or professional degree	2	**EES**, mean ± SD	
	55-64	3	Type of disability, *n*			39.37 ± 30.38
	65 or older	1	No disability	29	**EmoMA**, mean ± SD	
	Sexual orientation, *n*		Cognitive disability	0	*Negativity* score	3.98 ± 1.03
	Asexual	2	Physical disability	0	*Hostility* score	4.26 ± 0.96
	Bisexual	6	Psychological/mental health disability	8	*Anxiety* score	3.74 ± 1.11
	Heterosexual	32	Sensory disability	0	*Depression* score	3.95 ± 1.10
	Homosexual	0	Other	4		
	Pansexual	1	Race/Ethnicity, *n*			
	Other	0	American Indian or Alaskan Native	0		
	Annual household income, *n*		Asian	7		
	Under $25,000	3	Black or African American	10		
	$25,000 - $49,999	10	Hispanic or Latino or Spanish Origin of any race	4		
	$50,000 - $74,999	9	Multi-racial/mixed race	9		
	$75,000 - $99,999	8	Native Hawaiian or Other Pacific Islander	0		
	$100,000 - $149,999	9	White	11		
	$150,000 or more	2	Other	0		

Table 12: The distributions of the participants' demographics & attributes in Phase 3 (Evaluation)

Demo	Gender, *n*		Physical attributes, mean ± SD		**Big5**, mean ± SD	
	Female	30	Height (cm)	171.4 ± 11.11	Extroversion	2.89 ± 0.96
	Male	29	Weight (kg)	80.24 ± 24.85	Agreeableness	3.97 ± 0.59
	Non-binary	1	Highest level of education, *n*		Conscientiousness	3.94 ± 0.78
	Age, *n*		High school or equivalent	4	Neuroticism	2.73 ± 1.04
	18-24	5	Some college or vocational training	20	Openness	3.65 ± 0.69
	25-34	19	Bachelor's degree	22	**VLQ**, mean ± SD	
	35-44	16	Master's degree	11		7.29 ± 1.51
	45-54	10	Doctoral or professional degree	3	**EES**, mean ± SD	
	55-64	8	Type of disability, *n*			34.32 ± 29.16
	65 or older	2	No disability	46	**EmoMA**, mean ± SD	
	Sexual orientation, *n*		Cognitive disability	0	*Negativity* score	4.04 ± 0.82
	Asexual	2	Physical disability	2	*Hostility* score	4.38 ± 0.82
	Bisexual	6	Psychological/mental health disability	8	*Anxiety* score	3.69 ± 0.98
	Heterosexual	49	Sensory disability	1	*Depression* score	4.05 ± 0.88
	Homosexual	2	Other	3		
	Pansexual	0	Race/Ethnicity, *n*			
	Other	1	American Indian or Alaskan Native	2		
	Annual household income, *n*		Asian	13		
	Under $25,000	10	Black or African American	9		
	$25,000 - $49,999	16	Hispanic or Latino or Spanish Origin of any race	7		
	$50,000 - $74,999	9	Multi-racial/mixed race	9		
	$75,000 - $99,999	10	Native Hawaiian or Other Pacific Islander	0		
	$100,000 - $149,999	9	White	19		
	$150,000 or more	6	Other	1		

A.4 Questionnaires



Emotional reactions to MA vignettes (abbr. **EmoMA**)

! The questionnaire here is repeated n times.
($n = 40$ in Data Survey, $n = 12$ in Analogy Survey)

Read the vignette carefully and respond to the following questions based on your own feelings and thoughts. Please respond instinctively without overthinking – simply choose the first answer that comes to your mind.

A Vignette of Microaggression

X, who is deaf, picks up a DVD or Blu-ray to check for subtitles. The main feature of the DVD or Blu-ray includes subtitles for the deaf. However, there is a notation on the DVD or Blu-ray indicating that the special features "may not" have subtitles.

Emotion

Indicate your affective states after reading the vignette.

7-pt Likert-scale

Angry: ○ ○ ○ ○ ○ ○ ○
1 (Not at all) 4 7 (Very Much)

(The same format is applied to all items below.)
[12 items] Angry, Blue, Fearful, Cruel, Discouraged, Worried, Agreeable, Fine, Secure, Cooperative, Active, Calm

Reason

Please explain the reasons behind your emotional response, considering your personal experiences, values, traits, and other relevant factors. Make sure your answer is at least 250 characters long and contains at least three sentences.

Here are the example answers:
e.g. It seems like just a normal situation. It happens all the time, so I didn't feel much different. I'm not quite sure what's supposed to be undesirable about it.
e.g. Never being bisexual myself, I find it hard to grasp. It's kind of like automatically being wary of anything outside my own experiences. It's not about judgment, just an instinctive response to stuff I haven't encountered personally. It's tricky navigating feelings towards the unknown.
e.g. I'm a very introverted person when it comes to expressing my emotions. I don't show my emotions outwardly, and because of this, I am sometimes called indifferent or cold. However, I don't intentionally ignore or alienate others. It doesn't matter to me if he is a woman or a man, this situation could be someone similar to me.
e.g. Being white with a black mom, discrimination against people of color hits hard. It stirs up anger, seeing loved ones hurt. It's a personal battle, fighting for respect and equality in a world that often looks the other way. It's about standing together, valuing our shared humanity.

Free-form

Awareness

7-pt Likert-scale

For each item below: ○ ○ ○ ○ ○ ○ ○
1 (Not at all) 4 7 (Very Much)

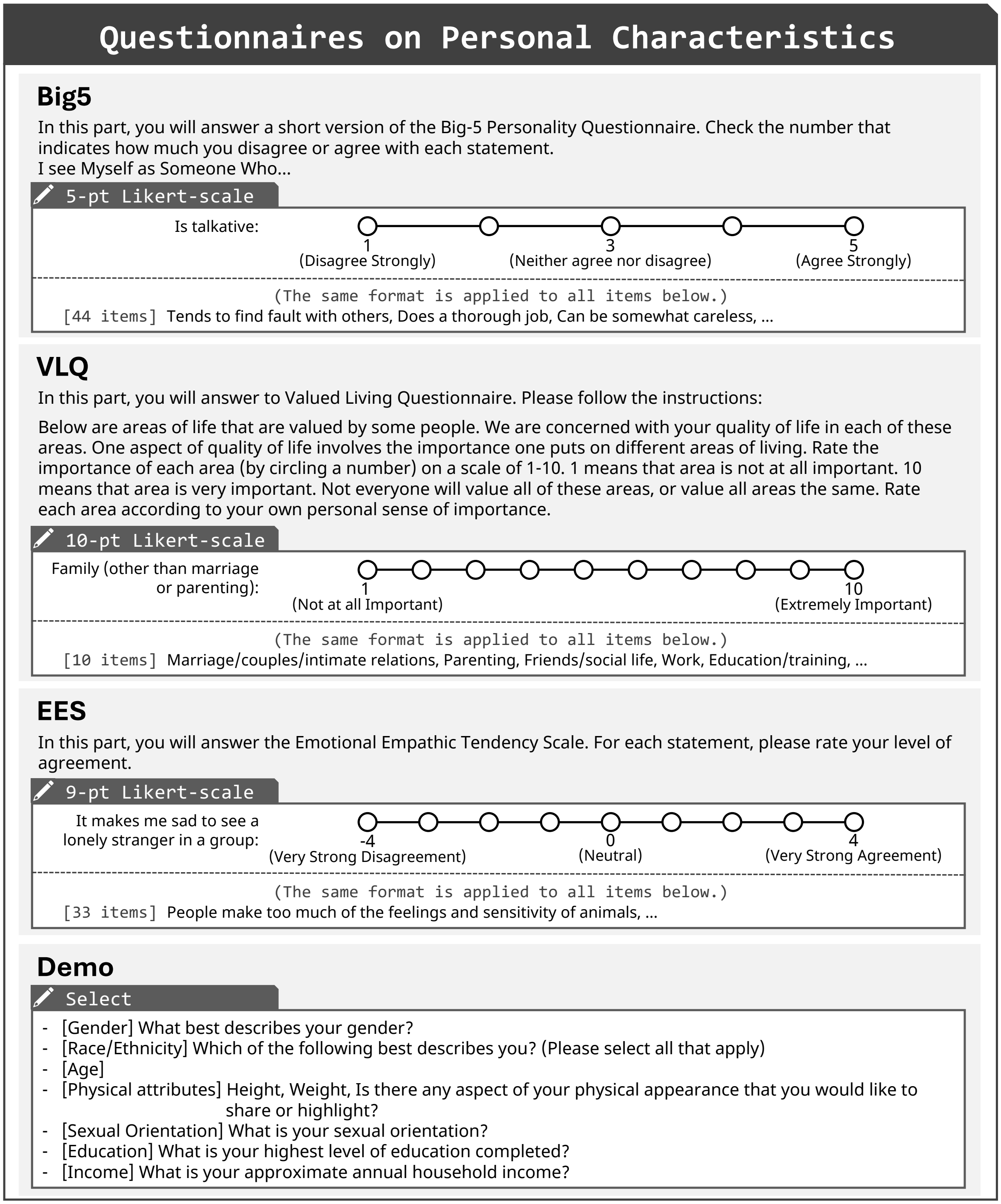
- How undesirable do you think the interaction in the vignette is?
- How likely do you think that any of the individuals or groups in the situation felt marginalized due to bias?
- How likely do you think that there was an intent of bias against an individual or a group?

Yes or No

For each item below: ○ Yes ○ No

- Does this vignette remind you of a past experience you have had?
- Have you ever seen or heard about a similar situation in media, such as in movies, TV shows, books, or news?

Figure 10: EmoMA questionnaire



Questionnaires on Personal Characteristics

Big5

In this part, you will answer a short version of the Big-5 Personality Questionnaire. Check the number that indicates how much you disagree or agree with each statement.
I see Myself as Someone Who...

5-pt Likert-scale

Is talkative: ○ ○ ○ ○ ○
1 (Disagree Strongly) 3 (Neither agree nor disagree) 5 (Agree Strongly)

(The same format is applied to all items below.)
[44 items] Tends to find fault with others, Does a thorough job, Can be somewhat careless, ...

VLQ

In this part, you will answer to Valued Living Questionnaire. Please follow the instructions:

Below are areas of life that are valued by some people. We are concerned with your quality of life in each of these areas. One aspect of quality of life involves the importance one puts on different areas of living. Rate the importance of each area (by circling a number) on a scale of 1-10. 1 means that area is not at all important. 10 means that area is very important. Not everyone will value all of these areas, or value all areas the same. Rate each area according to your own personal sense of importance.

10-pt Likert-scale

Family (other than marriage or parenting): ○ ○ ○ ○ ○ ○ ○ ○ ○ ○
1 (Not at all Important) 10 (Extremely Important)

(The same format is applied to all items below.)
[10 items] Marriage/couples/intimate relations, Parenting, Friends/social life, Work, Education/training, ...

EES

In this part, you will answer the Emotional Empathic Tendency Scale. For each statement, please rate your level of agreement.

9-pt Likert-scale

It makes me sad to see a lonely stranger in a group: ○ ○ ○ ○ ○ ○ ○ ○ ○
-4 (Very Strong Disagreement) 0 (Neutral) 4 (Very Strong Agreement)

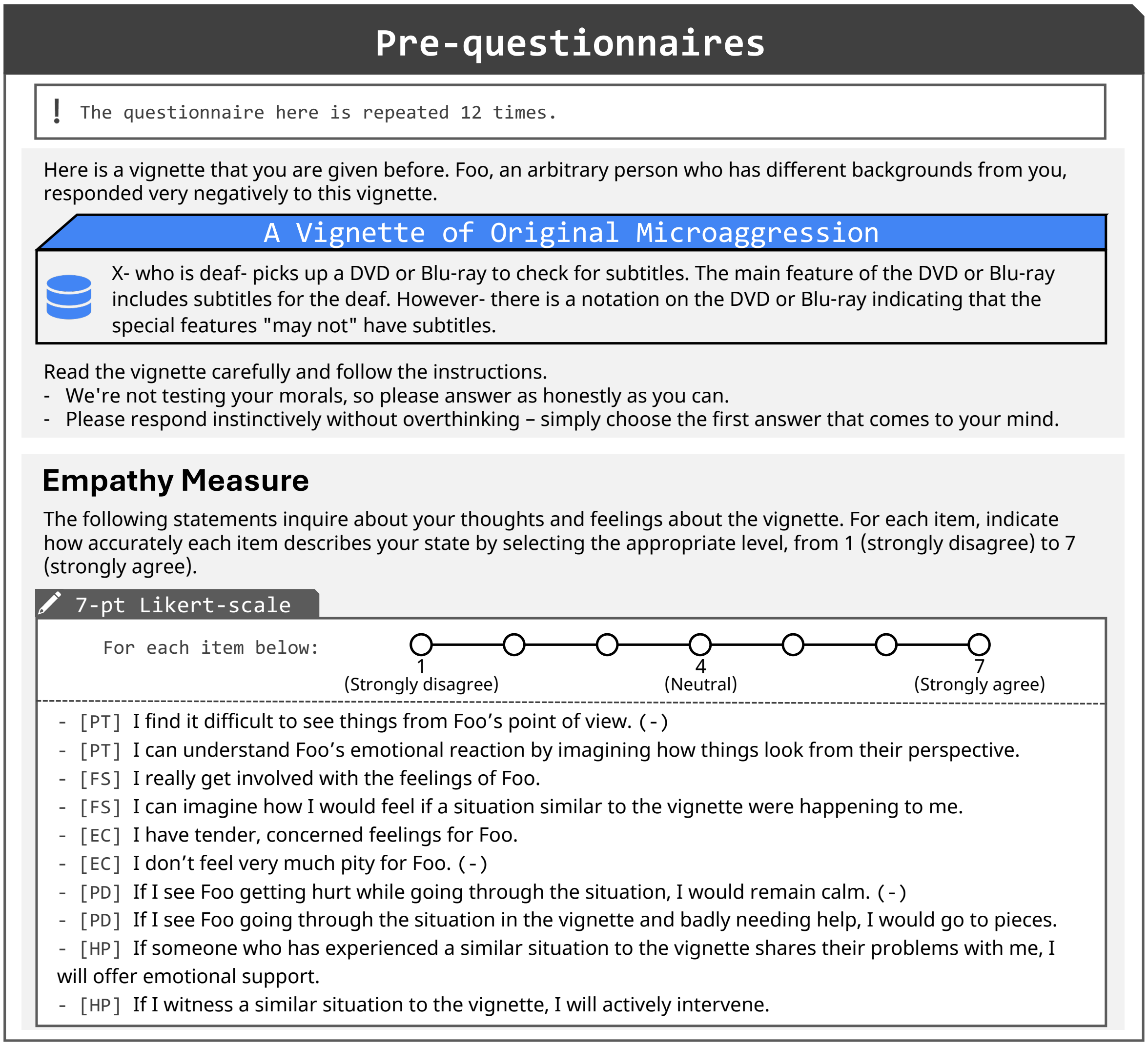
(The same format is applied to all items below.)
[33 items] People make too much of the feelings and sensitivity of animals, ...

Demo

Select

- [Gender] What best describes your gender?
- [Race/Ethnicity] Which of the following best describes you? (Please select all that apply)
- [Age]
- [Physical attributes] Height, Weight, Is there any aspect of your physical appearance that you would like to share or highlight?
- [Sexual Orientation] What is your sexual orientation?
- [Education] What is your highest level of education completed?
- [Income] What is your approximate annual household income?

Figure 11: Questionnaires on personal characteristics



Pre-questionnaires

! The questionnaire here is repeated 12 times.

Here is a vignette that you are given before. Foo, an arbitrary person who has different backgrounds from you, responded very negatively to this vignette.

A Vignette of Original Microaggression

X- who is deaf- picks up a DVD or Blu-ray to check for subtitles. The main feature of the DVD or Blu-ray includes subtitles for the deaf. However- there is a notation on the DVD or Blu-ray indicating that the special features "may not" have subtitles.

Read the vignette carefully and follow the instructions.

- We're not testing your morals, so please answer as honestly as you can.
- Please respond instinctively without overthinking – simply choose the first answer that comes to your mind.

Empathy Measure

The following statements inquire about your thoughts and feelings about the vignette. For each item, indicate how accurately each item describes your state by selecting the appropriate level, from 1 (strongly disagree) to 7 (strongly agree).

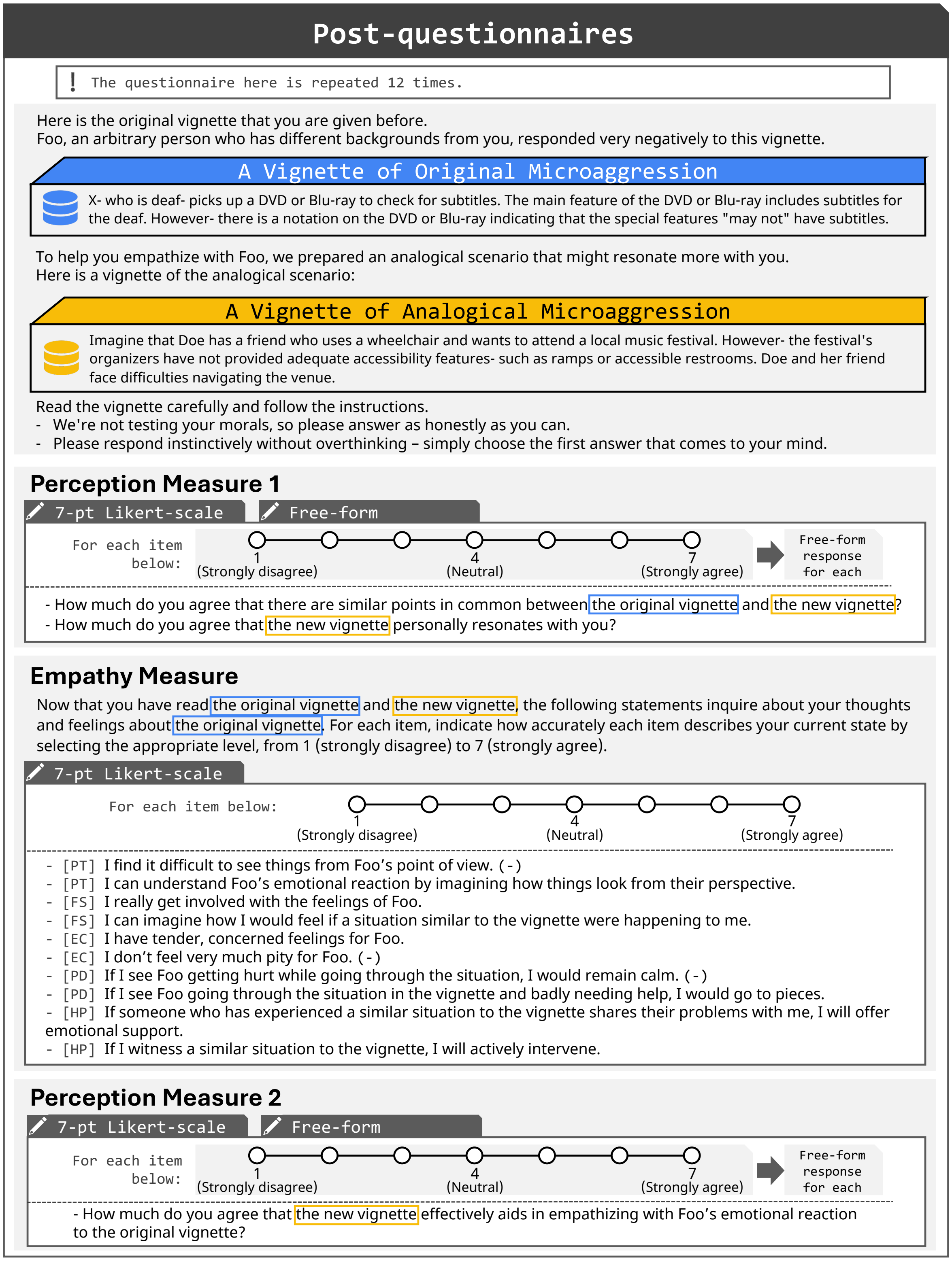
7-pt Likert-scale

For each item below: ○ ○ ○ ○ ○ ○ ○

1 (Strongly disagree) 4 (Neutral) 7 (Strongly agree)

- [PT] I find it difficult to see things from Foo's point of view. (-)
- [PT] I can understand Foo's emotional reaction by imagining how things look from their perspective.
- [FS] I really get involved with the feelings of Foo.
- [FS] I can imagine how I would feel if a situation similar to the vignette were happening to me.
- [EC] I have tender, concerned feelings for Foo.
- [EC] I don't feel very much pity for Foo. (-)
- [PD] If I see Foo getting hurt while going through the situation, I would remain calm. (-)
- [PD] If I see Foo going through the situation in the vignette and badly needing help, I would go to pieces.
- [HP] If someone who has experienced a similar situation to the vignette shares their problems with me, I will offer emotional support.
- [HP] If I witness a similar situation to the vignette, I will actively intervene.

Figure 12: Pre-questionnaires



Post-questionnaires

! The questionnaire here is repeated 12 times.

Here is the original vignette that you are given before.
Foo, an arbitrary person who has different backgrounds from you, responded very negatively to this vignette.

A Vignette of Original Microaggression

X- who is deaf- picks up a DVD or Blu-ray to check for subtitles. The main feature of the DVD or Blu-ray includes subtitles for the deaf. However- there is a notation on the DVD or Blu-ray indicating that the special features "may not" have subtitles.

To help you empathize with Foo, we prepared an analogical scenario that might resonate more with you.
Here is a vignette of the analogical scenario:

A Vignette of Analogical Microaggression

Imagine that Doe has a friend who uses a wheelchair and wants to attend a local music festival. However- the festival's organizers have not provided adequate accessibility features- such as ramps or accessible restrooms. Doe and her friend face difficulties navigating the venue.

Read the vignette carefully and follow the instructions.
- We're not testing your morals, so please answer as honestly as you can.
- Please respond instinctively without overthinking – simply choose the first answer that comes to your mind.

Perception Measure 1

7-pt Likert-scale | Free-form

For each item below: ○ ○ ○ ○ ○ ○ ○ 1 (Strongly disagree) 4 (Neutral) 7 (Strongly agree) → Free-form response for each

- How much do you agree that there are similar points in common between the original vignette and the new vignette?
- How much do you agree that the new vignette personally resonates with you?

Empathy Measure

Now that you have read the original vignette and the new vignette, the following statements inquire about your thoughts and feelings about the original vignette. For each item, indicate how accurately each item describes your current state by selecting the appropriate level, from 1 (strongly disagree) to 7 (strongly agree).

7-pt Likert-scale

For each item below: ○ ○ ○ ○ ○ ○ ○ 1 (Strongly disagree) 4 (Neutral) 7 (Strongly agree)

- [PT] I find it difficult to see things from Foo's point of view. (-)
- [PT] I can understand Foo's emotional reaction by imagining how things look from their perspective.
- [FS] I really get involved with the feelings of Foo.
- [FS] I can imagine how I would feel if a situation similar to the vignette were happening to me.
- [EC] I have tender, concerned feelings for Foo.
- [EC] I don't feel very much pity for Foo. (-)
- [PD] If I see Foo getting hurt while going through the situation, I would remain calm. (-)
- [PD] If I see Foo going through the situation in the vignette and badly needing help, I would go to pieces.
- [HP] If someone who has experienced a similar situation to the vignette shares their problems with me, I will offer emotional support.
- [HP] If I witness a similar situation to the vignette, I will actively intervene.

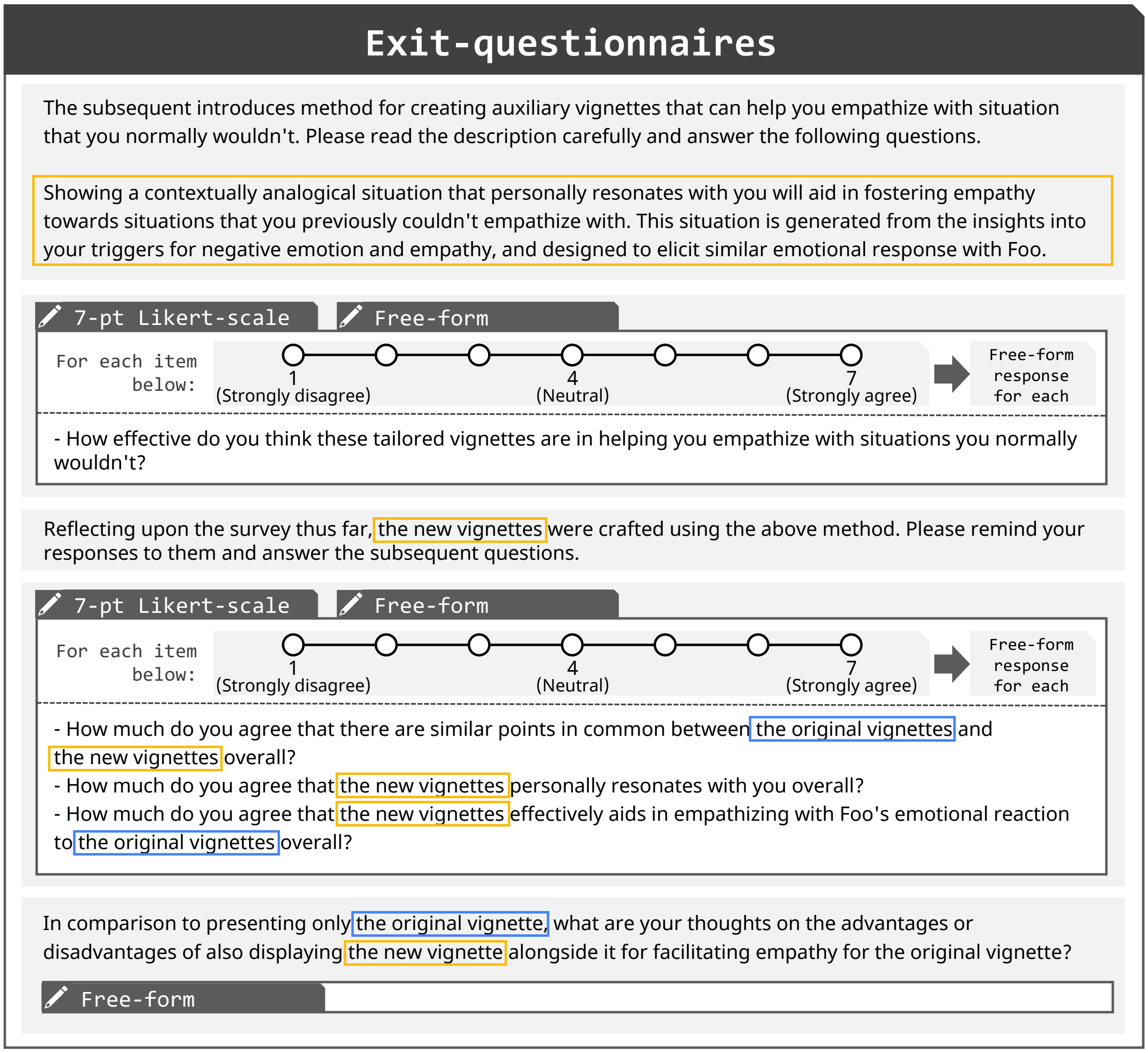
Perception Measure 2

7-pt Likert-scale | Free-form

For each item below: ○ ○ ○ ○ ○ ○ ○ 1 (Strongly disagree) 4 (Neutral) 7 (Strongly agree) → Free-form response for each

- How much do you agree that the new vignette effectively aids in empathizing with Foo's emotional reaction to the original vignette?

Figure 13: Post-questionnaires



Exit-questionnaires

The subsequent introduces method for creating auxiliary vignettes that can help you empathize with situation that you normally wouldn't. Please read the description carefully and answer the following questions.

Showing a contextually analogical situation that personally resonates with you will aid in fostering empathy towards situations that you previously couldn't empathize with. This situation is generated from the insights into your triggers for negative emotion and empathy, and designed to elicit similar emotional response with Foo.

7-pt Likert-scale | Free-form

For each item below: ○ ○ ○ ○ ○ ○ ○ 1 (Strongly disagree) 4 (Neutral) 7 (Strongly agree) → Free-form response for each

- How effective do you think these tailored vignettes are in helping you empathize with situations you normally wouldn't?

Reflecting upon the survey thus far, the new vignettes were crafted using the above method. Please remind your responses to them and answer the subsequent questions.

7-pt Likert-scale | Free-form

For each item below: ○ ○ ○ ○ ○ ○ ○ 1 (Strongly disagree) 4 (Neutral) 7 (Strongly agree) → Free-form response for each

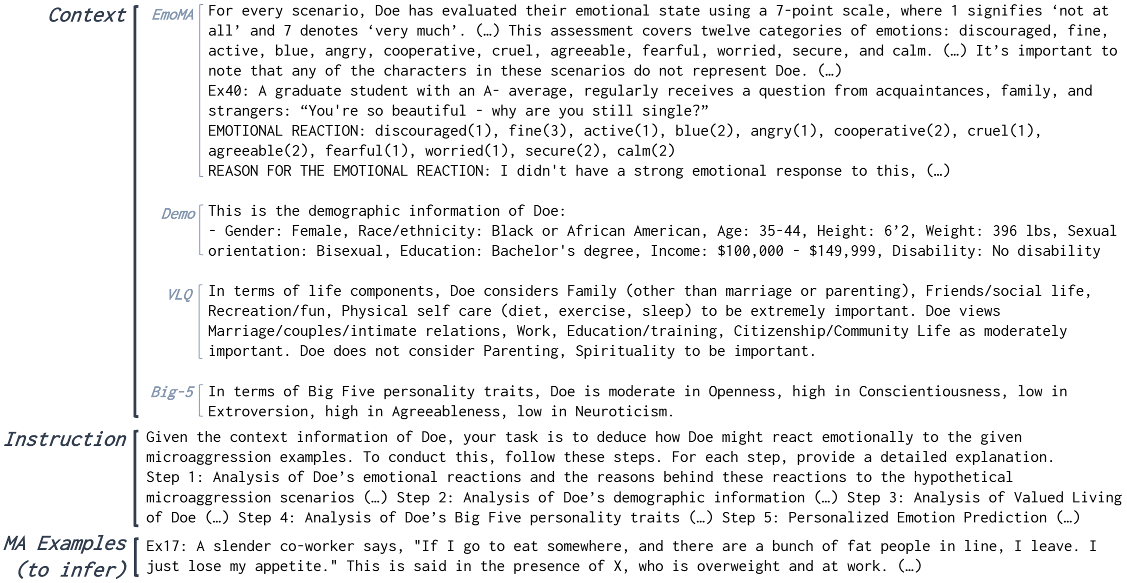
- How much do you agree that there are similar points in common between the original vignettes and the new vignettes overall?
- How much do you agree that the new vignettes personally resonates with you overall?
- How much do you agree that the new vignettes effectively aids in empathizing with Foo's emotional reaction to the original vignettes overall?

In comparison to presenting only the original vignette, what are your thoughts on the advantages or disadvantages of also displaying the new vignette alongside it for facilitating empathy for the original vignette?

Free-form

Figure 14: Exit-questionnaires

A.5 Example Prompts



Context

EmoMA

For every scenario, Doe has evaluated their emotional state using a 7-point scale, where 1 signifies ‘not at all’ and 7 denotes ‘very much’. (…) This assessment covers twelve categories of emotions: discouraged, fine, active, blue, angry, cooperative, cruel, agreeable, fearful, worried, secure, and calm. (…) It’s important to note that any of the characters in these scenarios do not represent Doe. (…)
Ex40: A graduate student with an A- average, regularly receives a question from acquaintances, family, and strangers: “You're so beautiful - why are you still single?”
EMOTIONAL REACTION: discouraged(1), fine(3), active(1), blue(2), angry(1), cooperative(2), cruel(1), agreeable(2), fearful(1), worried(1), secure(2), calm(2)
REASON FOR THE EMOTIONAL REACTION: I didn't have a strong emotional response to this, (…)

Demo

This is the demographic information of Doe:
- Gender: Female, Race/ethnicity: Black or African American, Age: 35-44, Height: 6’2, Weight: 396 lbs, Sexual orientation: Bisexual, Education: Bachelor's degree, Income: $100,000 - $149,999, Disability: No disability

VLQ

In terms of life components, Doe considers Family (other than marriage or parenting), Friends/social life, Recreation/fun, Physical self care (diet, exercise, sleep) to be extremely important. Doe views Marriage/couples/intimate relations, Work, Education/training, Citizenship/Community Life as moderately important. Doe does not consider Parenting, Spirituality to be important.

Big-5

In terms of Big Five personality traits, Doe is moderate in Openness, high in Conscientiousness, low in Extroversion, high in Agreeableness, low in Neuroticism.

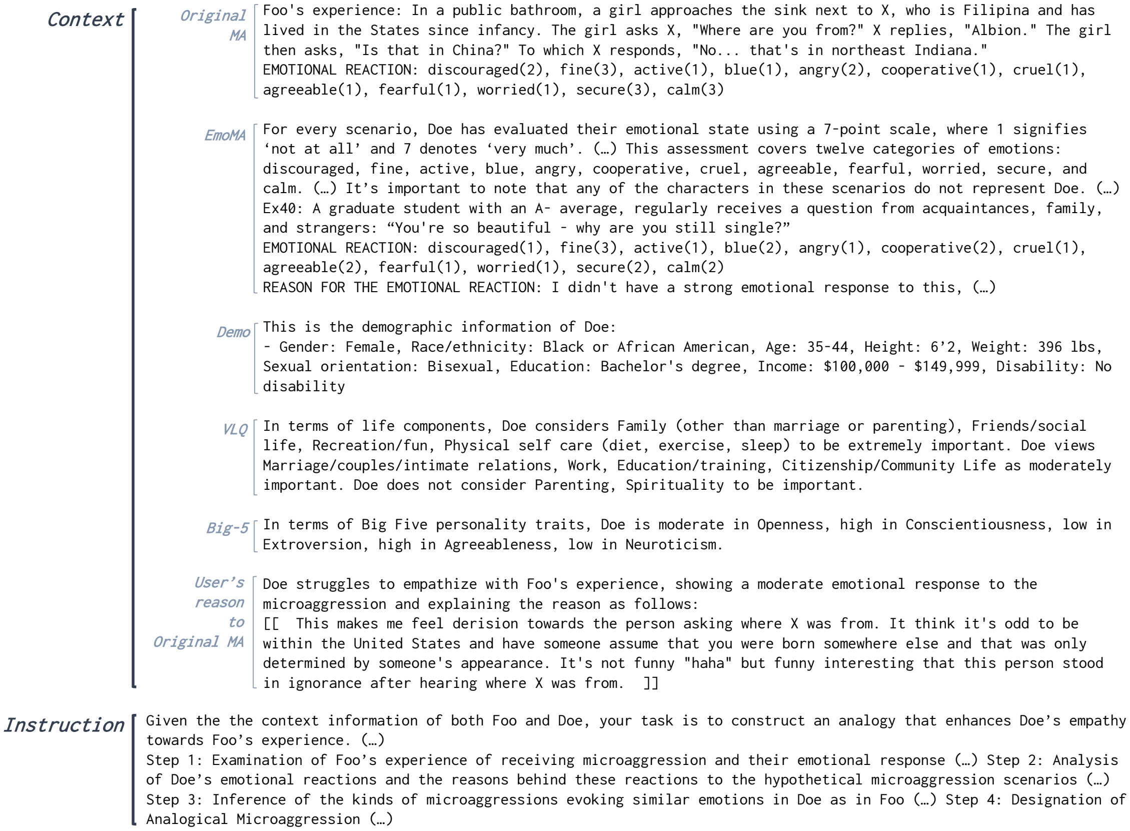
Instruction

Given the context information of Doe, your task is to deduce how Doe might react emotionally to the given microaggression examples. To conduct this, follow these steps. For each step, provide a detailed explanation.
Step 1: Analysis of Doe’s emotional reactions and the reasons behind these reactions to the hypothetical microaggression scenarios (…) Step 2: Analysis of Doe’s demographic information (…) Step 3: Analysis of Valued Living of Doe (…) Step 4: Analysis of Doe’s Big Five personality traits (…) Step 5: Personalized Emotion Prediction (…)

MA Examples (to infer)

Ex17: A slender co-worker says, "If I go to eat somewhere, and there are a bunch of fat people in line, I leave. I just lose my appetite." This is said in the presence of X, who is overweight and at work. (…)

Figure 15: Example of `Base-prompt`



Context

Original MA

Foo's experience: In a public bathroom, a girl approaches the sink next to X, who is Filipina and has lived in the States since infancy. The girl asks X, "Where are you from?" X replies, "Albion." The girl then asks, "Is that in China?" To which X responds, "No... that's in northeast Indiana."
EMOTIONAL REACTION: discouraged(2), fine(3), active(1), blue(1), angry(2), cooperative(1), cruel(1), agreeable(1), fearful(1), worried(1), secure(3), calm(3)

EmoMA

For every scenario, Doe has evaluated their emotional state using a 7-point scale, where 1 signifies ‘not at all’ and 7 denotes ‘very much’. (…) This assessment covers twelve categories of emotions: discouraged, fine, active, blue, angry, cooperative, cruel, agreeable, fearful, worried, secure, and calm. (…) It’s important to note that any of the characters in these scenarios do not represent Doe. (…)
Ex40: A graduate student with an A- average, regularly receives a question from acquaintances, family, and strangers: “You're so beautiful - why are you still single?”
EMOTIONAL REACTION: discouraged(1), fine(3), active(1), blue(2), angry(1), cooperative(2), cruel(1), agreeable(2), fearful(1), worried(1), secure(2), calm(2)
REASON FOR THE EMOTIONAL REACTION: I didn't have a strong emotional response to this, (…)

Demo

This is the demographic information of Doe:
- Gender: Female, Race/ethnicity: Black or African American, Age: 35-44, Height: 6’2, Weight: 396 lbs, Sexual orientation: Bisexual, Education: Bachelor's degree, Income: $100,000 - $149,999, Disability: No disability

VLQ

In terms of life components, Doe considers Family (other than marriage or parenting), Friends/social life, Recreation/fun, Physical self care (diet, exercise, sleep) to be extremely important. Doe views Marriage/couples/intimate relations, Work, Education/training, Citizenship/Community Life as moderately important. Doe does not consider Parenting, Spirituality to be important.

Big-5

In terms of Big Five personality traits, Doe is moderate in Openness, high in Conscientiousness, low in Extroversion, high in Agreeableness, low in Neuroticism.

User’s reason to Original MA

Doe struggles to empathize with Foo's experience, showing a moderate emotional response to the microaggression and explaining the reason as follows:
[[This makes me feel derision towards the person asking where X was from. It think it's odd to be within the United States and have someone assume that you were born somewhere else and that was only determined by someone's appearance. It's not funny "haha" but funny interesting that this person stood in ignorance after hearing where X was from.]]

Instruction

Given the the context information of both Foo and Doe, your task is to construct an analogy that enhances Doe’s empathy towards Foo’s experience. (…)
Step 1: Examination of Foo’s experience of receiving microaggression and their emotional response (…) Step 2: Analysis of Doe’s emotional reactions and the reasons behind these reactions to the hypothetical microaggression scenarios (…)
Step 3: Inference of the kinds of microaggressions evoking similar emotions in Doe as in Foo (…) Step 4: Designation of Analogical Microaggression (…)

Figure 16: Example of `Final-prompt`