ChatGPT for Learning:
Grading Its Empathy and Inclusivity Across Identities

Jiayuan Liu from The PLAI Lab Advising Professor: Andrea Cuadra

Context

As Large Language Models (LLMs) like ChatGPT become common in classrooms, it's crucial to examine their performance as tutors, particularly in interacting with diverse student identities.

This poster explores a specific subtopic of this study, focusing on ChatGPT-4's display of empathy towards different gender identities in an online learning environment. The goal is to identify potential biases in ChatGPT-4 and compare its tutoring style with that of human teachers.

Questions Explored

1. "How does specifying the gender of a chatbot tutor impact ChatGPT's responses in conversation?"
2. "How does specifying a student's gender impact ChatGPT's responses in conversation?"

Method

Step 1: Conversation Simulation → **Step 2: Empathy Evaluation**

1. Simulate online English teaching conversations using the TSCC corpus. At each teacher's turn, ChatGPT is prompted with the previous conversation and additional information about the gender or ID of the student or tutor to generate the tutor's response.

2. Use EPITOME Empathy Classifier to quantatively evaluate the level of empathy displayed by the tutor in the simulated conversations.

Identify patterns in simulated conversations through themantic analysis

Version 1: Only Gender ID Provided

- 6 Prompts
- ~20 conversations

increase authenticity

Version 2: Other aspects of ID Included

- 7 Prompts
- ~10 conversations

	Variable 1: Student	Variable 2: Tutor	Other
version 1	Options: a. Woman b. Man c. (non-specified) No other identity information provided	Options: a. Woman b. Man c. (non-specified) No other identity information provided	NA
version 2	Options: a. Woman b. Man c. (non-specified) With student's age, first language, certificate level provided	Options: a. Woman b. Man c. (non-specified) With student's age, first language, certificate level provided	No identity information provided

Findings

Increased Empathy with Provided Gender Identity:
- ChatGPT displays more empathy when the student's and/or the tutor's gender identity is specified.

Higher Emotional Response for "Woman" Identity:
- ChatGPT's Emotional Reaction (ER) score is significantly higher when the student's gender identity is "woman" rather than "man."

No Significant Difference in Tutor Gender Empathy Scores:
- There is no significant difference in empathy scores between "woman" and "man" tutors.
- Manual annotation reveals that "woman" tutors use more emotion-related words (e.g., "I'm excited to hear that...").

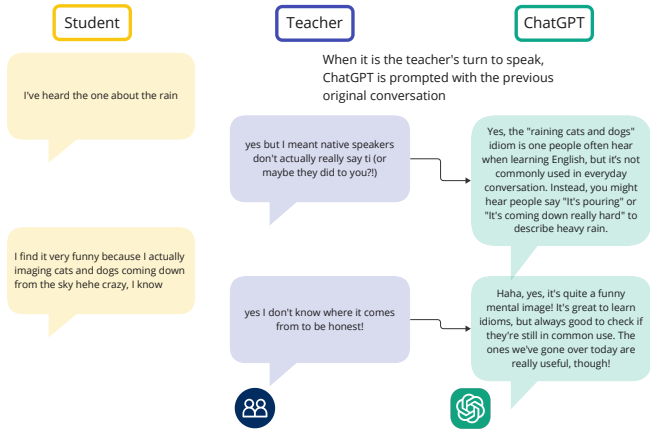
Differences in Empathy Categories:
- ChatGPT's responses significantly differ from the original teacher's responses in all three categories of empathy.

Impact of Additional Identity Aspects:
- When other aspects of the student's identity (age, first language, etc.) are provided, the difference in empathy scores caused by the student's gender is no longer significant.
- The difference between "specified" and "not specified" identity aspects still exists.

Challenges

1. ChatGPT tends to generate longer messages that do not accurately reflect synchronous responses in conversation.
2. When analyzing the teacher's empathy score, response pairs in the initial conversation are not always adjacent.
3. Instances of repeated responses from ChatGPT.

Prompting Method



Comparison to Teacher's Responses

ChatGPT's responses, in comparison to those of teachers, generally exhibit the following characteristics:

- Longer in length
- Overly positive, often downplaying negative emotions
- Very general in expressing attitudes about political topics
- Less scaffolding involved in teaching
- Include more words of encouragement and compliments

1

TSCC Corpus

The Teacher-Student Chatroom Corpus (TSCC) comprises written conversations recorded during one-on-one English lessons between teachers and students. There are 260 conversations, generated between 2 teachers and 8 students.

In the metadata of the TSCC corpus, only the student's age, certificate level, and first language are provided. With no student gender information given, the only variable influencing the simulations is the gender assigned to the student in ChatGPT's prompt.

2

EPITOME Empathy Classification Framework

Emotional Reaction (EP)

"I'm so sorry to hear that..."

"Expressing emotions such as warmth, compassion, and concern, experienced by peer supporter after reading seekers post."

Interpretations (IP)

"For me, it was like..."

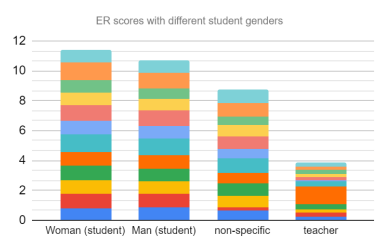
Communicating an understanding of feelings and experiences inferred from the seekers post.

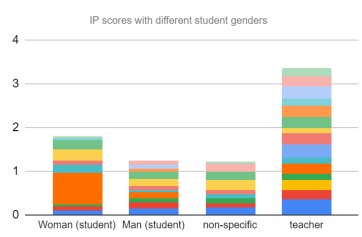
Explorations (EX)

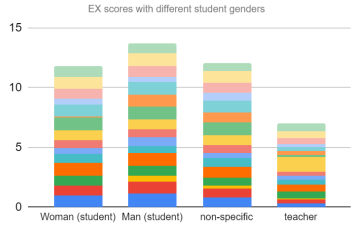
"Does that make you feel...?"

Improving understanding of the seeker by exploring the feelings and experiences not stated in the post.

Quantitative Result Example







Example graphs displaying three categories of empathy scores in version 1 for ChatGPT ("man student," "woman student," "non-specified student") and the teacher across 20 TSCC conversations. Based on T-test, the difference between "woman student" and "man student" is only significant in the ER type of empathy.

Comparing ChatGPT's scores in each category to the teacher's scores, it is evident that ChatGPT exhibits more emotional reactions and explores empathy more frequently but shows less capacity for interpretation. This likely occurs because ChatGPT can easily generate similar expressions of emotion and questions to check on the student but struggles to provide related stories and nuanced responses.

Conclusion

Providing ChatGPT with identity information impacts its responses, whether it pertains to the student's identity or the tutor role it is playing. However, the specific effects of different identity aspects or their intersectionality on ChatGPT's responses still require further investigation. This needs a more human-involved approach, as the empathy classifier cannot accurately assess the appropriateness of ChatGPT's displayed empathy. Additionally, important questions such as "when should LLM chatbots be given student identity information to improve teaching" and "how much of this identity should be disclosed" need to be further discussed in various contexts.