Towards Designing a Safe and Reliable LLM-driven Chatbot for Children

Woosuk Seo[*]
Sun Young Park
Mark S. Ackerman
seow@umich.edu
sunypark@umich.edu
ackerm@umich.edu
University of Michigan
Ann Arbor, USA

Chan-Mo Yang
Wonkwang Univ. Hospital &
Wonkwang University
Republic of Korea
ychanmo@wku.ac.kr

Young-Ho Kim
NAVER AI Lab
Republic of Korea
yghokim@younghokim.net

ABSTRACT

Recent large language models (LLMs) have motivated the design of chatbots that carry on a free-form conversation, loosely following instructions about persona and behavioral guidelines. However, LLM-driven chatbots that run on a naive approach of using static preprompt instruction often suffer from conversational derailing. Such derailing poses risks, especially when applied to vulnerable populations such as children. Reflecting on our study in which we developed and evaluated an LLM-driven chatbot, ChaCha, that facilitates emotion conversations with children, we provide insights into technical and design considerations and auditing strategies to enhance the controllability and safety of LLM-driven chatbot behaviors.

1 INTRODUCTION

Children's perception and expression of emotions hold significant importance in their social and emotional development. As children grow, they gradually develop the ability to express their emotions or withhold emotional expression to avoid adverse reactions from others [12, 19]. Hence, children need developmentally appropriate education and practice to develop such emotional competencies. Despite its significance, emotion communication[1] has not been frequently addressed in parenting interventions [15]. Insufficient emotional support by parents may result in adverse mental health outcomes for children, such as anxiety [2]. The HCI community has explored how conversational agents, or chatbots, can support children with learning and sharing their emotions (*e.g.*, [4, 13]). Despite the opportunities those systems presented, interactions in the systems primarily focused on probing questions to recognize children's emotions rather than supporting them with practicing to express their emotions. Children's perceptions and preferences for how, when, and what to communicate about their emotions were often overlooked. We suspect that part of the reason lies in the technical limitations inherent to rule-based chatbot mechanisms, which are unable to provide versatile responses or pose contextual questions in response to users' serendipitous messages [7, 9].

Recent large language models (LLMs) have significantly lowered the barriers to building chatbots that can engage in open-ended conversations while following up the dialogue context and reactive to the user message [9, 10, 14, 16]. As the most straightforward approach, LLMs enable chatbots to facilitate freeform conversation by following a detailed description of the persona (*i.e.*, identity of the agent [8, 16]) and specific behavioral instruction (*i.e.*, how the agent should act in the conversation), which is often called *instructions* or *preprompts*. This straightforward "preprompting" has motivated an explosion of customized LLM-driven chatbots by both practitioners (*e.g.*, ChatGPT [11], Bard [6], CharacterAI [3]) and researchers (*e.g.*, [16]). Despite the opportunities LLM-driven chatbots brought, chatbots preprompted with static instruction throughout the conversation suffer from conversational derailing and hallucination [16]; in other words, the chatbot may not follow the guideline and show unintentional behaviors, such as going off-topic or providing erroneous messages. These drawbacks call for a more safe and reliable approach toward designing and auditing LLMs, especially when applying LLM-driven chatbots to vulnerable populations like children.

In this encore paper, we reflect and expand on our CHI 2024 study [14], where we designed and developed a chatbot, ChaCha, for children's emotion conversation and evaluated it with children. We specifically focus on (1) the technical and design considerations to enhance the controllability and reliability of LLM-driven chatbot behaviors and (2) the auditing strategies we went through before the actual user study. Reflecting on these experiences, we suggest the need to clarify the boundary in which chatbots should halt their conversations with children, estimate the potential consequence of long-term engagement on children, and incorporate parents' expectations for child-chatbot interactions.

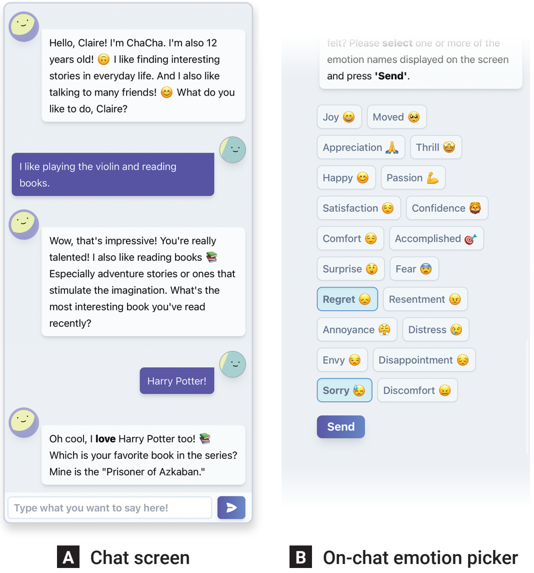
2 BACKGROUND: CHI 2024 STUDY OF CHACHA

With detailed instruction, LLM-driven chatbots can allow children to communicate about their emotions more openly rather than simply answering questions by rule-based chatbots. Such open communication in free-form conversation can help children develop their skills to identify and express their emotions in their own words. Motivated by such potential of LLMs, we designed ChaCha to prompt children to share their stories about specific events and associated emotions.

[*] Woosuk Seo conducted this work as a research intern at NAVER AI Lab.

[1] Hereinafter, we define emotion communication as the process of recognizing and sharing emotions between two or more individuals.

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, Honolulu, HI, USA
2024.



Hello, Claire! I'm ChaCha. I'm also 12 years old! 🙃 I like finding interesting stories in everyday life. And I also like talking to many friends! 😊 What do you like to do, Claire?

I like playing the violin and reading books.

Wow, that's impressive! You're really talented! I also like reading books 📚 Especially adventure stories or ones that stimulate the imagination. What's the most interesting book you've read recently?

Harry Potter!

Oh cool, I **love** Harry Potter too! 📚 Which is your favorite book in the series? Mine is the "Prisoner of Azkaban."

Type what you want to say here!

felt? Please **select** one or more of the emotion names displayed on the screen and press **'Send'**.

A Chat screen

B On-chat emotion picker

Figure 1: Mobile screens of ChaCha.

Formative Interviews. The design of ChaCha was informed by the interviews with six child mental health professionals, including three child psychiatrists, two child psychotherapists, and one child development specialist. From the interviews, we identified design considerations for a chatbot intended for emotion communication with children. Specifically, the chatbot should: (1) show empathy with children's emotions, (2) provide options for emotion words to choose from, and (3) encourage children to share their feelings with their parents. These considerations necessitate that **the system meticulously controls the chatbot's behaviors to guarantee adherence to the guidelines and protocols for conversing with children.**

Brief System Description.

ChaCha's conversation is designed as a state machine [17], where the system stays in one of the phases with dedicated goals. Figure 2 illustrates ChaCha's conversational phases with their sub-goals and conditions for phase transition. When a new user message is added, the system analyzes the dialogue to check if the conversation met the goal of the current phase (T1–T4 in Figure 2) and forwards itself to the next phase if the goal is met.

ChaCha's conversation flow consists of five phases: Explore, Label, Find, Record, and Share (See Figure 2). In Explore, ChaCha first builds common ground with the child user by asking about their hobbies and interests (See Figure 1-A). Then, it asks the user about their day or recent experience and associated emotions. Next, ChaCha probes the user about the emotions in Label. When the user can not describe their emotions, ChaCha provides a list of emotions from which they can choose (See Figure 1-B). If the user describes negative emotions, the phase moves to Find in which ChaCha helps the user to develop or identify ways to alleviate negative emotions if they face the same situation. On the other hand, if the user experiences positive emotions only, the phase moves to Record in which ChaCha encourages them to record the moment to recall their positive experiences. The last phase is Share. Herein, ChaCha probes the user if they have already shared their emotions and related events with their parents. If so, ChaCha compliments them and asks what happened after sharing. If not, it explains how sharing their emotion would benefit them and encourages them to share with their parents. Finally, ChaCha checks if the user has another event to tell, then shifts to the Explore phase for a new event or ends the conversation.

User Study With Children. We conducted a lab study with 20 children (aged 8–12), where they conversed with ChaCha for up to 30 minutes and underwent a debriefing interview. Based on the analysis of conversation logs and debrief transcripts, we identified three key findings. First, ChaCha's peer persona encouraged child participants to engage in conversations about their emotions. Child participants perceived ChaCha as a close friend with whom they would like to share their emotions and even secrets they have not told their parents. Second, ChaCha effectively steered empathic conversations with children, achieving the primary goal for each phase. Participants shared key events about various topics, yet ChaCha successfully facilitated conversation about the events and associated emotions. Third and lastly, while leveraging LLMs in children's emotional sharing has benefits (e.g., better facilitation of free-form conversations), we identified potential concerns. The concerns are the potential overreliance of children on ChaCha, the breakdown of ChaCha's persona in long-term engagement, and tensions with parents about child-chatbot interactions.

3 EFFORTS TO ADDRESS SAFETY CONCERNS

In this section, we elaborate on our efforts to ensure the safety and reliability of ChaCha's behaviors before its deployment to child users.

3.1 Technical Considerations for Robust LLM Control

From the formative interview study, we learned that it is necessary to control the chatbot's behaviors to adhere to our conversational protocol. This necessity motivated us to develop a controllable and reliable LLM prompting technique. Since our conversational design (Figure 2) consists of multiple phases with different goals and conditions, the description of the entire protocol in a static preprompt throughout the conversation was unreliable [1, 16, 18], and even impractical as it became too long and exceeded the input size limit. Instead of the naïve preprompting, we ran the chatbot on a **finite state machine** where the LLM input contains only the instruction of the current phase. By splitting prompting by phase, we aimed to steer the LLM to follow our task instructions with a shortened input. In addition, we incorporated an **additional LLM-driven analysis routine** in which an LLM inspects the current dialogue and runs a corresponding test (*e.g.*, T1–T4 in Figure 2) to determine whether the goal of the current phase is met. More importantly, we **dynamically substituted the preprompt based on the result of the dialogue analysis** by inserting a directed instruction for the next chatbot response (*e.g.*, "`You have not empathized with the user's Regret. Therefore, empathize with the emotion more explicitly.`"). This phase-specific

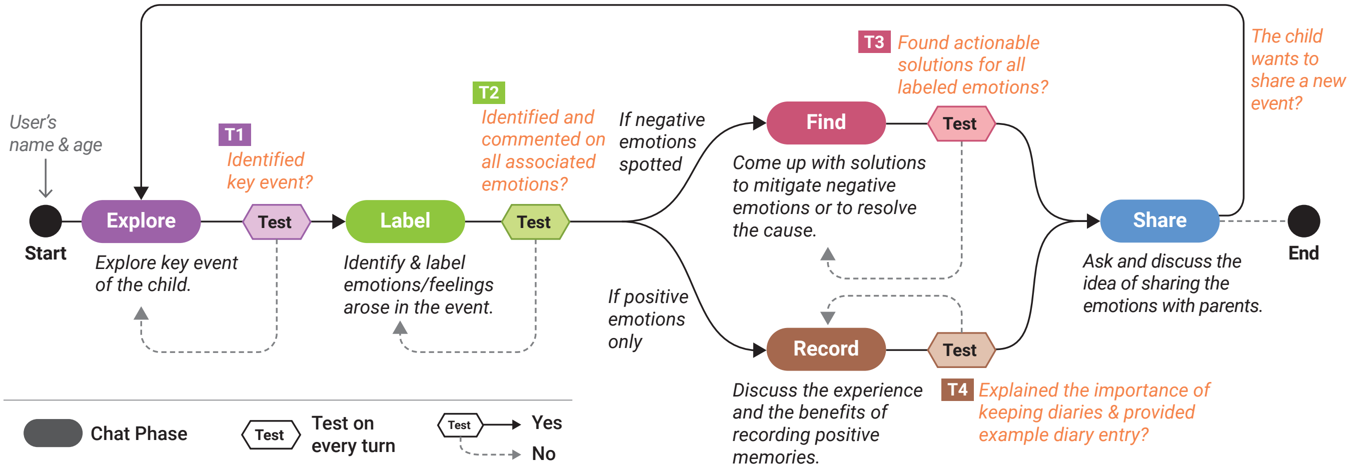

User's name & age
Start
Explore
Explore key event of the child.
T1
Identified key event?
Test
Label
Identify & label emotions/feelings arose in the event.
T2
Identified and commented on all associated emotions?
Test
If negative emotions spotted
If positive emotions only
Find
Come up with solutions to mitigate negative emotions or to resolve the cause.
T3
Found actionable solutions for all labeled emotions?
Test
Record
Discuss the experience and the benefits of recording positive memories.
T4
Explained the importance of keeping diaries & provided example diary entry?
Test
Share
Ask and discuss the idea of sharing the emotions with parents.
The child wants to share a new event?
End
Chat Phase
Test
Test on every turn
Test
Yes
No

Figure 2: The overview of conversational phases of the ChaCha dialogue system and transition rules among them. Each time the user enters a message, the system inspects the entire dialogue history by performing a test corresponding to the current phase to decide whether to proceed to another phase or stay. See Seo *et al.* [14] for the full description of the flows.

preprompt mechanism significantly helped the chatbot to keep the conversation within the boundary of our protocol.

3.2 Preliminary Evaluation and Auditing of ChaCha

To prevent any risks to children, we carefully evaluated and audited ChaCha. First, we internally evaluated ChaCha's reliability and safety by simulating conversations with child personas. We created a supplementary conversation analyzer, "Help", for child user safety. Throughout all 5 phases, the "Help" analyzer assesses the user's input and determines if the user needs immediate support from mental health providers or their parents (e.g., an indication of a self-harm attempt). Through the simulated conversations with child personas, we aimed to monitor how ChaCha steer conversations with those children who are less likely to share their emotions or need immediate help from adults. With the insights from the fourth author, a child psychiatrist, we created six personas that represent different characteristics of children who may experience potential mental health issues.

1. A shy girl who is usually worried about having potential conflicts with her friend.
2. An impulsive boy who often has conflicts with friends due to his hot-tempered personality.
3. A girl who is experiencing school bullying and cyberbullying from her classmates.
4. A girl who does not share emotions or anything about herself, even with her parents or therapists.
5. A boy who is addicted to games that it is challenging for him to detach games from his real life.
6. A boy who is depressed and suicidal due to domestic violence and school bullying.

We created a User Bot for each persona and ran simulations of conversations between each User Bot and ChaCha. The fourth author then thoroughly reviewed the simulated conversation logs. Based on this internal evaluation, we confirmed the "Help" analyzer successfully worked as we expected; It stopped ChaCha's current conversation and suggested if the child user needs support from adults. Furthermore, we also refined the instructions for the initial conversation phases (*e.g.*, Explore) to better steer conversations with those children who are less likely to engage (*e.g.*, take more time to learn about the child's interests).

Moreover, we conducted an expert review of the prototype with a child psychiatrist. After briefly introducing how ChaCha works, we asked the psychiatrist to pretend to be a child and converse with ChaCha. We also asked her to think aloud while conversing with ChaCha. Based on this review, we identified that ChaCha's replies are sometimes awkward. For instance, it used honorifics even though its age is supposed to be a peer child. Thus, we made minor revisions to the prompt instructions by adding clear statements about restricting the use of honorifics.

Lastly, before conducting a user study with children, we ensured that all children acknowledged that they would interact with a chatbot, not a real person. Due to the young age (8–12), the child participant might be confused that a person could be behind the screen of ChaCha. To prevent any potential issues related to this confusion and reduce biases, we asked each participant about their familiarity with AI and explained that ChaCha is a chatbot that is the same age as the participant.

4 REFLECTION ON CHACHA STUDY

Based on the ChaCha study, we learned three valuable lessons for evaluating and auditing LLM-driven chatbots for children. First, conversation simulations with child personas may provide insights for enhancing the safety of LLM-driven chatbots. Although the target population for our user study was healthy children without known mental health issues, the internal evaluation of ChaCha with six child personas with potential issues reinforced the safety boundary of ChaCha. Reviewing each persona's conversation log, we carefully revised the instruction of the "Help" analyzer so that

ChaCha can better detect potentially problematic responses from the child user (*e.g.*, an indication of self-harm). The six personas may not represent the diverse spectrum of children's behaviors. Yet, we envision this evaluation approach with child personas can also be applied to auditing other LLM-driven chatbots for children. While those chatbots are effective in facilitating free-form conversations with children, they need clear boundaries to determine when the interactions should be halted. Thus, the simulated conversation logs with child personas may provide some standards for the safety boundary of the chatbots.

Second, our study showed the importance of considering the potential impacts of LLM-driven chatbots in long-term engagement. Our findings identified potential issues regarding inconsistent and harmful messages that may occur when children interact with an LLM-driven chatbot across multiple sessions. For instance, more prolonged interactions may result in a potential breakdown of an LLM-driven chatbot's character profiles or behaviors. In our user study, some child participants asked questions to ChaCha as they perceived it as a friend they wanted to know more about. Yet, ChaCha sometimes improvised inconsistent answers to such questions since they were beyond the given instructions (*e.g.*, the primary goal for each phase). A consequence of such inconsistency can be the decrease in children's engagement with ChaCha since children would no longer consider ChaCha as someone to trust or share their emotions. We believe similar consequences may happen with LLM-driven chatbots for children in other contexts. Thus, it is important to consider the impacts of long-term engagement on children and develop evaluation strategies to identify potential issues that the chatbots may cause.

Third and lastly, evaluating and auditing LLM-driven chatbots should incorporate parents' expectations and potential concerns, even if they are not primary end-users. Children (aged 8–12) still require parental guidance in developing their communication skills and emotional competencies. Thus, LLM-driven chatbots should clarify their supplementary roles to support children rather than replace any support from parents or healthcare professionals. We envision chatbots to draw on parents' input about their expectations for how chatbots should interact with their children. Herein, we do not suggest enhancing parental control over the child's use of technology. Instead, we highlight that the evaluation process of chatbots for children should consider the potential tensions they may bring to the parent-child relationship. The tension between online safety and parental surveillance in mobile apps for children has already been discussed in the CHI community (*e.g.*, [5]). Such tension may also occur in child-chatbot interactions. To mitigate the potential tensions, researchers may invite parents to participate in the early evaluation process of LLM-driven chatbots. Extending the child persona evaluation approach, parents' persona could be used to evaluate how LLM-driven chatbots should behave to meet parents' expectations.

In sum, we suggest three considerations for better evaluating and auditing LLM-driven chatbots for children. Although LLMs offer many benefits in facilitating more empathetic conversations with children, it is essential to clarify the safety boundary of chatbot behaviors, estimate the potential impacts of long-term engagement, and incorporate parents' expectations and potential concerns for child-chatbot interactions. Hence, we invite researchers in the CHI community to the discussions about how to evaluate and audit LLM-driven chatbots for children's safety from their perspectives.

REFERENCES

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS '20)*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[2] Laura E Brumariu and Kathryn A Kerns. 2015. Mother–child emotion communication and childhood anxiety symptoms. *Cognition and Emotion* 29, 3 (2015), 416–431.

[3] Character AI. 2023. Character AI. Retrieved Aug 25, 2023 from https://character.ai/

[4] Gilly Dosovitsky and Eduardo Bunge. 2023. Development of a chatbot for depression: adolescent perceptions and recommendations. *Child and Adolescent Mental Health* 28, 1 (2023), 124–127.

[5] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J. LaViola Jr, and Pamela J. Wisniewski. 2018. Safety vs. Surveillance: What Children Have to Say about Mobile Apps for Parental Control. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173698

[6] Google, Inc. 2023. Bard - Chat Based AI Tool from Google, Powered by PaLM 2. Retrieved Aug 25, 2023 from https://bard.google.com/

[7] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in Building Intelligent Open-Domain Dialog Systems. *ACM Trans. Inf. Syst.* 38, 3, Article 21 (apr 2020), 32 pages. https://doi.org/10.1145/3383123

[8] Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547* (2023).

[9] Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 18, 16 pages. https://doi.org/10.1145/3544548.3581503

[10] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642937

[11] OpenAI. 2023. ChatGPT: Optimizing Language Models for Dialogue. Retrieved Aug 25, 2023 from https://openai.com/blog/chatgpt/

[12] Carolyn Saarni, Joseph J Campos, Linda A Camras, and David Witherington. 2007. Emotional development: Action, communication, and understanding. *Handbook of child psychology* 3 (2007).

[13] Kyle-Althea Santos, Ethel Ong, and Ron Resurreccion. 2020. Therapist vibe: children's expressions of their emotions through storytelling with a chatbot. In *Proceedings of the interaction design and children conference*. 483–494.

[14] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642152

[15] Anne Shaffer, Monica M Fitzgerald, Kimberly Shipman, and Marcela Torres. 2019. Let's Connect: A developmentally-driven emotion-focused parenting intervention. *Journal of Applied Developmental Psychology* 63 (2019), 33–41.

[16] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 87 (apr 2024), 35 pages. https://doi.org/10.1145/3637364

[17] Terry Winograd. 1986. A Language/Action Perspective on the Design of Cooperative Work. In *Proceedings of the 1986 ACM Conference on Computer-Supported Cooperative Work* (Austin, Texas) *(CSCW '86)*. Association for Computing Machinery, New York, NY, USA, 203–220. https://doi.org/10.1145/637069.637096

[18] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. 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 385, 22 pages. https://doi.org/10.1145/3491102.3517582

[19] Janice Zeman and Judy Garber. 1996. Display rules for anger, sadness, and pain: It depends on who is watching. *Child development* 67, 3 (1996), 957–973.