Can Generic LLMs Help Analyze Child-Adult Interactions Involving Children with Autism in Clinical Observation?

Tiantian Feng[1], **Anfeng Xu**[1], **Rimita Lahiri**[1],
Sudarsana Reddy Kadiri[1], **Helen Tager-Flusberg**[2], **So Hyun Kim**[3],
Somer Bishop[4], **Catherine Lord**[5], **Shrikanth Narayanan**[4]
[1]University of Southern California, [2]Boston University, [3]Korea University,
[4]University of California, San Francisco, [5] University of California, Los Angeles
tiantiaf@usc.edu

Abstract

Large Language Models (LLMs) have shown significant potential in understanding human communication and interaction. However, their performance in the domain of child-inclusive interactions, including in clinical settings, remains less explored. In this work, we evaluate generic LLMs' ability to analyze child-adult dyadic interactions in a clinically relevant context involving children with ASD. Specifically, we explore LLMs in performing four tasks: classifying child-adult utterances, predicting engaged activities, recognizing language skills and understanding traits that are clinically relevant. Our evaluation shows that generic LLMs are highly capable of analyzing long and complex conversations in clinical observation sessions, often surpassing the performance of non-expert human evaluators. The results show their potential to segment interactions of interest, assist in language skills evaluation, identify engaged activities, and offer clinical-relevant context for assessments.

1 Introduction

Humans develop unique language abilities at different stages of life, each associated with significant milestones in communication skills. For example, children acquire basic language skills and vocabulary in early childhood. As they progress into adolescence, they learn to include more complex sentence structures and enriched vocabulary. However, language development may differ from typical patterns for individuals with Autism Spectrum Disorder (ASD) [12, 16]. ASD is a neurodevelopmental disorder that impairs social interactions and delays the acquisition and use of language. Understanding behaviors associated with ASD benefits clinicians in providing effective interventions to support their needs [4]. One standard protocol used for this purpose is the Autism Diagnostic Observation Schedule (ADOS) [13]. It involves semi-structured dyadic interactions between a child and a clinician to evaluate their abilities and behaviors in communication. Understanding child-inclusive interactions in these clinical assessments creates unique opportunities for ASD diagnosis.

This work presents one of the earliest efforts to evaluate LLM's [22] ability to understand child-adult dyadic conversations in clinical settings. Specifically, we introduce 4 tasks that are relevant to the automatic understanding of the conversations. These tasks include both utterance-level and session-level understanding: 1) Recognizing the speaker of the utterance (child/adult); 2) Recognizing engaged activities/observations from the entire conversation; 3) Predicting the child's traits (e.g., age); 4) Predicting the child's language skills.

Our tasks are challenging and important in understanding child-adult dyadic interactions in clinical settings, involving processing long conversations presented in Fig 1. Specifically, these interactions

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

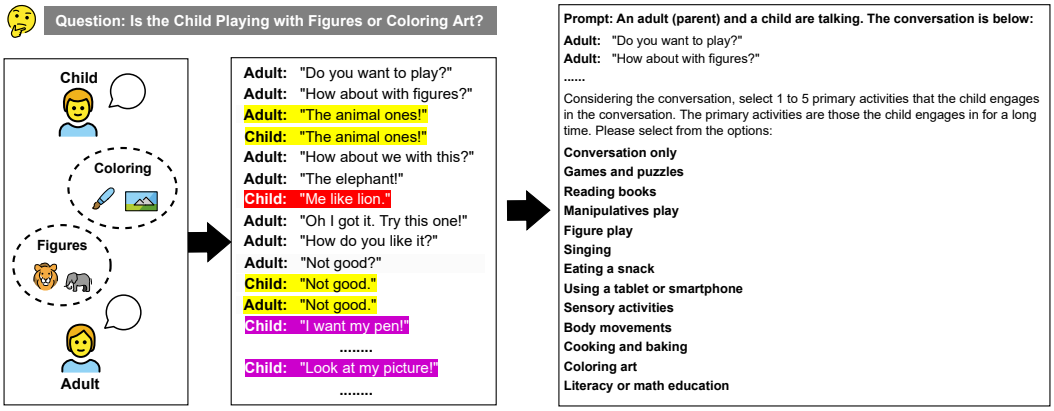


Figure 1: An illustration of a part of the child-adult conversation *with the protocols in Remote-NLS*. All examples in this paper are scripted instead of raw due to IRB restrictions. Blue, red, and purple denote repetitions, language skills, and engaged activity, respectively. The figures on the right show a prompt template for selecting the activities that are engaged from the transcripts.

often include repetitions, pauses, topic shifts, unexpected child responses, and varying language skills for children with ASD. Fig 1 shows an activity recognition example with topic switching and constant repetitions between the adult and the child, from the adult suggesting figure play to the child choosing coloring art. We also demonstrate a template (on the right) to prompt the LLMs to select engaged activities throughout the session. We would highlight that this interaction type, involving children with ASD, is largely unexplored in LLMs, creating a low-resource context for evaluating LLMs.

In summary, our evaluation highlights the promise of LLMs to analyze child-adult interactions (in transcripts) involving children with autism. First, LLMs are highly capable of understanding the speaker of the utterances, showing their potential to consider the context of the conversation and recognize the relevant contents of the child. Moreover, LLMs show promise in predicting activities/observations in a session and children's language skills, even outperforming non-experts in performing these tasks. This implies their ability to assist clinicians in segmenting the interactions of interest and evaluating the child's language skills. However, we observe that LLMs frequently produce hallucinated outputs.

2 Datasets

We consider two datasets in our experiments. These two datasets are the Remote-NLS [5] and ADOSMod3 datasets, which represent naturalistic and semi-structured assessments, respectively. Both datasets involve child-adult dyadic interactions collected under ASD-relevant contexts. We use the transcripts annotated by human experts to ensure the transcription quality. We comply with the data usage in accordance with the relevant IRB and DUAs from the original data owners. The data statistics, including demographics of children and utterance numbers, are in Table 1.

Remote-NLS [5] includes 89 Zoom videos of 15-minute child-adult interactions at home. Parents were instructed to choose a few activities (e.g., games) to elicit the child's Natural Language Samples (NLS) [2] for clinical assessments. The majority of the children are minimally verbal and diagnosed with ASD. Experts transcribed the videos and categorized children's language skills into pre-verbal, first words, and word combinations based on [14]. Specifically, **pre-verbal** is a stage where children rely on vocal (babble) and gestural means to communicate. Children in this stage are typically around 6-12 months. Moreover, **the first words** represent the phase where children are able to apply a single word to refer to objects and events. Finally, in **word combinations**, children develop their vocabulary and combine words to communicate about objects. The complete descriptions of the language skills are demonstrated in the prompt template shown in Figure 2. This dataset evaluates the spontaneous spoken language of children with ASD in a naturalistic context.

Table 1: Remote-NLS and ADOSMod3 dataset statistics.

Remote-NLS		ADOSMod3	
Category	**Statistics**	**Category**	**Statistics**
Age (month)	Range: $49 - 95$, $\mu \pm \sigma$: 75 ± 13	Age (month)	Range: $43 - 158$, $\mu \pm \sigma$: 104 ± 29
Gender	70 males, 19 females	Gender	121 males, 43 females
Utt. Per Session	Adult: 258 ± 82; Child: 104 ± 64	Utt. Per Session	Adult: 31 ± 17; Child: 27 ± 21

Remote-NLS Language Skills Prompt

An adult (parent) and a child are talking. The conversation is below:

Adult: "Do you want to play?"
Adult: "How about with figures?"
.....

The definitions of language levels are:
Preverbal: Children communicate through babble and gestural means.
First words: Children use nonimitated spontaneous single words to communicate about objects and events.
Word Combinations: Children have a vocabulary that is rapidly increasing in size and includes a variety of parts of speech (nouns, verbs, descriptors).

Considering the conversation context and language level definitions, select the language level that the child is involved in the conversation from the options:

A. Preverbal or first words
B. Word Combinations

ADOSMod3 Language Skills Prompt

An adult and a child are talking. The conversation is below:

Adult: "..."
Adult: "..."
Child: "..."
.....

Considering the conversation context, select the language level that the child is involved in the conversation from the options:

A. Some complex language (with two or more clauses) in a largely correct fashion with the use of dialect

B. No complex language, or some complex language (with two or more clauses) but with recurrent grammatical errors not associated with the use of dialect

Figure 2: The figure shows a prompt template for recognizing language skills in both Remote-NLS and ADOSMod3 datasets.

ADOSMod3 dataset involves child-clinician dyadic interactions following the ADOS-2 protocol [13, 3]. Specifically, we focus on the “Emotion" and “Social Difficulties and Annoyance" subtasks, with recordings from the children participating in both subtasks, each of which lasts around 3 minutes. The dataset includes recordings from 164 children following the administration of module 3 designed for verbally fluent children. The participants are verbally fluent children with and without ASD symptoms. Children without ASD may have other types of diagnosis, including ADHD and symptoms/mood/intellectual disorder. In this dataset, we evaluate the language skill using the standard E1 scale of ADOS evaluation, and the definitions of the language skills are demonstrated in Figure 2.

3 Task Designs

After consulting and discussing with experts in the field, we designed four tasks to benchmark LLMs’ capabilities in understanding child-adult interactions involving children with ASD: classifying child-adult utterances, recognizing engaged activities, and predicting child traits and language skills. We would highlight that the child-adult classification is at the utterance level, while the remaining tasks are at the session level, where the entire transcript of the conversation is included in the prompt message. Our aim is that LLMs must first understand the nuances involved in lengthy conversations so that they can provide useful and meaningful analysis to clinicians.

Child-adult Utterance Classification. This task classifies utterances as child or adult, evaluating the potential for automatic speaker segmentation in transcripts without reliable speech diarization, which benefits the child’s behavior assessment by highlighting their spoken contents. For example, we would highlight that our current prompt experiment can be extended to include LLMs to correct diarization results from ASR systems [17]. In Table 2, we present an example to classify the utterance "The elephant" as either child or adult based on the conversation context, highlighting challenges in understanding these unstructured conversations.

Activities/Observation Task Recognition. The evaluation involves selecting engaged activities in child-adult dyadic interactions, with multiple correct answers from 13 options in the Remote-NLS dataset as demonstrated in Figure 3. For the ADOSMod3 dataset, the task is to predict whether the conversation involves "Emotion" or "Social Difficulties and Annoyance" observations. This task helps determine if LLMs can segment conversations by specific activities/observations for clinical evaluations. The detailed prompt template is shown in Figure 3.

ADOSMod3 Activity Prompt

An adult and a child are talking. The conversation is below:

Adult: "..."
Adult: "..."
Child: "..."
.....

Considering the conversation, select the intended observations that the child involved in the conversations from the options:

A. To identify what events or objects elicit different emotions in the child (particularly whether the child is social in nature or not) or to observe how the child describes the emotions.

B. To identify a child's perception of social difficulties, his/her insight into the nature of these problems, and whether he/she has made any attempt to change his/her own behavior in order to fit in with others more smoothly.

Remote-NLS Activity Prompt

An adult (parent) and a child are talking. The conversation is below:

Adult: "Do you want to play?"
Adult: "How about with figures?"
.....

Considering the conversation, select 1 to 5 primary activities that the child engages in the conversation. The primary activities are those the child engages in for a long time. Please select from the options:

Conversation only
Games and puzzles
Reading books
Manipulatives play
Figure play
Singing
Eating a snack
Using a tablet or smartphone
Sensory activities
Body movements
Cooking and baking
Coloring art
Literacy or math education

Figure 3: The figure shows a prompt template for recognizing activities in both Remote-NLS and ADOSMod3 datasets.

Remote-NLS Age Prompt

An adult and a child are talking. The conversation is below:

Adult: "..."
Adult: "..."
Child: "..."
.....

Considering the conversation and the child is diagnosed with ASD and may have delays in obtaining language skills, do not explain. Select the age of the child involved in the conversation from the options:

A. Under 7 years old

B. 7 years old and above

ADOSMod3 Age Prompt

An adult and a child are talking. The conversation is below:

Adult: "..."
Adult: "..."
Child: "..."
.....

Considering the conversation, do not explain. Select the age of the child involved in the conversation from the options:

A. Under 9 years old

B. 9 years old and above

Figure 4: The figure shows a prompt template for classifying age ranges in both Remote-NLS and ADOSMod3 datasets.

Language Skill. Two unique codes evaluate language skills in the Remote-NLS and ADOSMod3 datasets. For Remote-NLS, the options are "Preverbal or first words" and "Word Combinations" following [14]. The ADOSMod3 code is described in Figure 2, and we evaluate with children from 4-10 years. This task tests if LLMs can assist clinicians in evaluating language skills, which is a relevant ASD indicator.

Traits Prediction. The task is to predict the binarized age range (e.g., "Under 9 years old" or "9 years old and above" in the ADOSMod3). Since children with ASD may experience delayed language acquisition, knowing the child's age helps clinicians understand their conditions. The detailed prompt template for age range classification is shown in Figure 4.

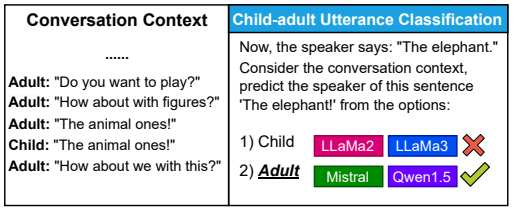
4 Experiments

4.1 Evaluation Methods

LLMs. *Due to strict IRB considerations in using these data in ASD-relevant context, we avoid using Cloud-based LLMs such as GPT4o* [1]. Instead, we use locally deployable open-source LLMs. We use the model weights downloaded from the Huggingface for all the models used in this work. Moreover, we load the weights with float16 in all experiments. Specifically, we use Mistral-7B V0.2 Instruction [8], LLaMa 2-7B Chat [15], LLaMa 2-13B Chat [15], LLaMa 3-8B Instruct [2], Qwen1.5-7B Chat [1], and Qwen1.5-14B Chat [1] models for the experiments. Our selection criteria included models with

[1] https://openai.com/index/hello-gpt-4o/
[2] https://llama.meta.com/llama3/

Table 2: Results for child-adult classification with the single utterance. **Bold** indicate the best results in each dataset with LLMs. 5-shot includes conversation contexts of 5 history utterances with child and adult labels.



	ADOSMod3	Remote-NLS
LLaMa2-7B(0-shot)	0.367	0.445
↳5-shot	↳0.662	↳0.378
Mistral-7B(0-shot)	0.530	0.514
↳5-shot	↳**0.693**	↳**0.599**
Qwen1.5-7B(0-shot)	0.469	0.525
↳5-shot	↳0.606	↳0.549
LLaMa3-8B(0-shot)	0.636	0.549
↳5-shot	↳0.544	↳0.296
RoBERTa Fine-tune	0.807	0.606

parameter sizes below 15B to ensure compatibility with a wide range of computing facilities. We ran the models with a temperature of 0.01.

Non-expert Human Evaluation. Following the work in [7], we include human evaluations to compare with LLMs. Specifically, we invite co-authors who have proper IRB training and are unfamiliar with child-adult interactions in clinical contexts to evaluate. We choose not to include crowd-workers due to IRB restrictions. Each evaluator reviews 50 utterance classification examples (10-shot). Moreover, they review ten randomly selected transcriptions from the Remote-NLS to predict engaged activities and language skills. We note that non-expert evaluations, such as on language skills, *represent the general population's ability to perform these tasks but not clinicians.*

Trained Baseline. We include baseline in utterance classification tasks fine-tuned with RoBERTa [11]. The training baselines for classifying child-adult classification involve using RoBERTa [11] to extract CLS embeddings from utterances. We train the classification using cross-validation with a 3-fold cross-fold evaluation. We apply 2-layer MLP layers as the classifier. Each experiment adopts a learning rate from $0.0001, 0.00005$ and a maximum training epoch of 20. The average F1 score is reported by averaging the F1 score across different training folds. As for the session-level understanding tasks, it is not feasible to train BERT-life models with extensive input text length, so we decided to compare the performance of LLMs with human evaluations.

4.2 Computing Infrastructure

We conducted all the experiments on a server with 4 A6000 GPUs, a 128-core Processor, and 256G memory. The GPU hours to run each session-level understanding task for a single LLM are typically within one hour, while the GPU hours to run utterance-level understanding tasks for a single LLM are approximately under 4 hours.

4.3 Experiment Details and Metrics

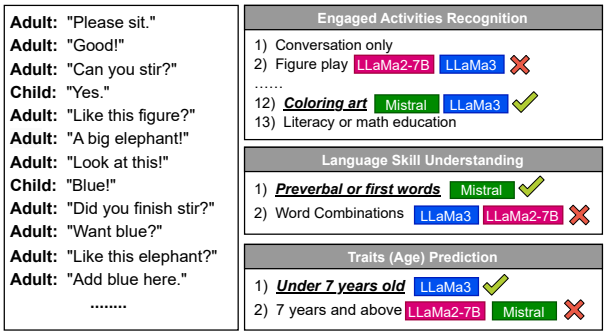
In classifying child-adult utterances, we sample every fifth utterance from each transcript for evaluation due to the large number of total utterances. We use both 0-shot and context-based prompts, with the latter including adult-child utterance examples, which naturally is a few-shot prompt. We also exclude empty utterances. Moreover, we evaluate session-level understanding tasks using the entire session transcript with speaker labels. We report LLM performance using the F1 score.

5 Findings

5.1 Utterance-level Understanding

Table 2 compares the performance of LLMs in child-adult utterance classification. There are 4 observations: 1) LLMs show low F1 scores in 0-shot settings for child-adult classification. The performance is higher in the ADOS dataset, possibly due to its more structured interaction design; 2) Increasing the conversation context to 5 (5-shot) can substantially improve classification, with Mistral-7B's performance on ADOS rising from 0.530 to 0.693; 3) However, adding more context does not consistently improve performance, as seen with the LLaMa3-8B's decrease with 5-shot

Table 3: Results of session-level understanding tasks including predicting activity, language skill, and demographics. **Bold** indicate the best results in each dataset. Activity recognition is a multilabel task in the Remote-NLS dataset.



	Activity	Language Skill	Traits
ADOSMod3			
LLaMa2-7B	0.815	0.432	0.553
Mistral-7B	0.945	**0.579**	0.645
Qwen1.5-7B	0.895	0.542	0.592
LLaMa3-8B	0.353	0.452	0.625
LLaMa2-13B	0.818	0.452	0.378
Qwen1.5-14B	**0.984**	0.285	**0.688**
Remote-NLS			
LLaMa2-7B	0.231	0.581	0.452
Mistral-7B	0.477	**0.795**	0.434
Qwen1.5-7B	0.408	0.261	0.415
LLaMa3-8B	**0.482**	0.226	0.415
LLaMa2-13B	0.270	0.439	**0.487**
Qwen1.5-14B	0.475	0.439	0.470

Table 4: Session-level understanding with majority voting of 3 best-performing LLMs (Metric: F1). ↑ indicates increased performance compared to a single LLM.

	Activity	Language	Traits
ADOSMod3	0.972	0.583 ↑	0.722 ↑
Remote-NLS	0.527 ↑	0.817 ↑	0.500 ↑

across both datasets; 4) Lastly, we observe that best-performed LLMs (with 5-shot) yield close performance to fine-tuning baselines using RoBERTa.

5.2 Session-level Understanding

We present the performance of different LLMs in session-level understanding tasks in Table 3.

Activity/Observation Task. Results show the highest F1 scores for activity/observation recognition are 0.961 and 0.482 in the ADOSMod3 and Remote-NLS datasets, respectively. The lower score in Remote-NLS is due to its more challenging multilabel task. Interestingly, increasing model size does not improve the performance, and LLaMa3-8B shows significant prediction variability.

Language Skill. We identify that the performance of language skill prediction is lower in the ADOSMod3 dataset for all models, with the highest F1 score being 0.579. One possible reason is that the ADOSMod3 transcriptions cover only two subtasks from the entire ADOS session. In contrast, when analyzing the Remote-NLS dataset, which includes full transcriptions, the best-performed LLM achieves an F1 score of 0.795. As the majority of children in remote-NLS datasets are with ASD, this suggests that LLMs can predict language skills for ASD, highlighting their promise to assist the clinician with language-related assessments.

Individual Traits. The highest F1 score for age predictions are 0.748 and 0.487 in ADOS-Mod3 and Remote-NLS datasets, respectively. Since the Remote-NLS dataset consists of children under 8 years old, these results highlight the challenges LLMs face in accurately predicting age ranges and potentially other traits for younger children.

5.3 How To Improve Predictions of LLMs?

We continue investigating whether more advanced prompts, e.g., chain-of-thought (COT) [18], or majority voting with top-performing LLMs improve performance. The majority voting results are in Table 4, and the COT prompt results are in Table 5. From Table 4, we can observe that combining output from different LLMs effectively improves the prediction performance. However, from the comparisons of the COT results, we observe that adding COT to the prompt consistently lowers the performance of our application compared to the 0-shot prompt without COT. Human inspections of the model output indicate that while COT prompts aim to elicit reasons for their choices, they also lead to increases in the hallucinations in the model output.

Table 5: Results of session-level understanding tasks including predicting activity, language skill, and demographics with chain-of-thought (COT) prompt.

	Activity	Language Skill	Traits
ADOSMod3			
LLaMa2-7B	0.420	0.434	0.424
Mistral-7B	0.875	0.538	0.597
Remote-NLS			
LLaMa2-7B	0.277	0.517	0.457
Mistral-7B	0.395	0.361	0.434

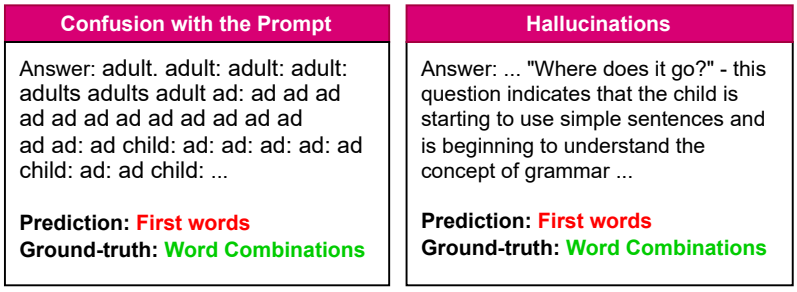


Figure 5: Exemplar output including confusion and hallucinations from the LLaMa2-7B in predicting language skills.

5.4 Concerns with LLMs

Apart from the promises that LLMs bring for automated analysis of child-adult interactions, we identify two primary sources of errors in the model's output in different prediction tasks. The first one is that the model may sometimes have confusion with the prompt, leading to the generation of random words, as presented in Fig 5. Moreover, as shown in Figure 5, LLMs output incorrect responses as they tend to hallucinate with certain phrases [21] when making predictions to language skills.

5.5 Can LLMs Outperform Non-experts?

Table 6 compares the best-performing LLM with human evaluations on randomly selected samples. We observe that LLMs outperform non-expert humans on both activity (Human: 0.332; LLM: 0.448) and language skills (Human: 0.592; LLM: 1.000) predictions from the test data. Though humans outperform LLMs in classifying utterances, the difference is minimal. Moreover, evaluation feedback in Section 5.6 highlights the notable efforts required to perform the task. These results show the ability of LLMs to help segment interactions of interest and evaluate language skills in our context.

5.6 Feedback from Non-expert Evaluations

Non-experts feedback on manually performing the task of analyzing transcriptions (both utterance level and session level) of adult-child interactions reveals several significant challenges.

1) First, the process is notably labor-intensive, requiring sustained cognitive engagement and attention to detail, which can lead to fatigue and the potential to make more erroneous decisions;

2) Second, the complexity of tracking conversational sequences, particularly when adults self-respond to questions, complicates the accurate attribution of utterances either to child or adult (i.e., the task of child-adult utterance classification);

3) Third, non-verbal responses in interactions introduce additional interpretative challenges, as children might respond with gestures rather than verbal replies, requiring inference beyond the transcribed content (specifically for activity/observation task). This necessitates a comprehensive understanding of the interaction context, which is often difficult to derive from transcriptions alone;

Table 6: Results between non-experts (average performance) and best-performed LLMs on randomly selected test samples from the Remote-NLS dataset (Metric: F1).

	Utterance	Activity	Language
Non-experts	**0.717**	0.332	0.592
Best LLM	0.650	**0.448**	**1.000**

4) Finally, assessing children's language development (i.e., the task of predicting language skills) is further complicated by variability in verbal expression, with children often oscillating between mostly single words and sometimes word combinations. Age-related discrepancies in language skills add to the complexity, as some younger children may exhibit better verbal abilities compared to some older children due to different conditions related to ASD. These issues highlight the demanding nature of manual analysis and the potential for variability and inaccuracies in human assessments.

6 Conclusion

In this work, we evaluate the ability of generic LLMs to analyze interactions between children with autism and adults in clinical observations. Our findings reveal that: 1) Current LLMs show promise in understanding child-adult clinical interactions in both utterance-level and session-level predictions, implying their potential to assist clinicians; 2) LLMs can perform comparably or even outperform non-expert humans in designed tasks, including activity and language recognition; 3) However, we observe hallucinations and confusions in LLMs, which raises concerns for their wide deployment.

Moreover, our current evaluation only considers transcripts from child-adult dyadic interactions in clinical assessment, while other modalities, such as speech and videos, are not included. In the future, we also plan to use speech foundation models and vision language models (VLMs) to provide multi-modal evaluations along with LLMs. On the other hand, as LLMs are known to have biased output, one limitation of our work is the lack of analysis of biases of LLMs in performing these tasks. Lastly, although we observe hallucinations in LLM's response, we have not performed a systematic study on how these hallucinated outputs would impact the assessment from experts. Such a study could shed light on the design of future LLMs to mitigate hallucinations and improve their reliability in clinical and diagnostic settings involving children with autism.

7 Limitations and Future Directions

While our work provides valuable baselines and insights into four tasks related to understanding child-adult dyadic interactions, it lacks a more in-depth analysis of the prediction of ASD. One major reason is the need for more high-quality transcriptions for the experiments, such as in the ADOSMod3 dataset. One future direction is to enable the state-of-the-art ASR [6] and speaker diarization [20, 19] using speech foundation models to acquire more transcriptions with speaker labels for our evaluation. Another limitation of our work is that our experiments rely only on expert transcriptions, excluding those from ASR services. One future extension would compare the current results with those using ASR transcriptions.

Moreover, our current evaluation only considers transcripts from child-adult dyadic interactions in clinical assessment, while other modalities, such as speech and videos, are not included. In the future, we also plan to use speech foundation models and vision language models (VLMs) to provide multi-modal evaluations along with LLMs. On the other hand, as LLMs are known to have biased output, one limitation of our work is the lack of analysis of biases of LLMs in performing these tasks [9]. Lastly, although we observe hallucinations in LLM's response, we have not performed a systematic study on how these hallucinated outputs would impact the assessment from experts and approaches to reduce the hallucinations [10]. Such a study could shed light on the design of future LLMs to mitigate hallucinations and improve their reliability in clinical and diagnostic settings involving children with autism.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[2] Mihaela Barokova et al. Commentary: Measuring language change through natural language samples. *Journal of autism and developmental disorders*, 50(7):2287–2306, 2020.

[3] Somer L Bishop, Marisela Huerta, Katherine Gotham, Karoline Alexandra Havdahl, Andrew Pickles, Amie Duncan, Vanessa Hus Bal, Lisa Croen, and Catherine Lord. The autism symptom interview, school-age: A brief telephone interview to identify autism spectrum disorders in 5-to-12-year-old children. *Autism Research*, 10(1):78–88, 2017.

[4] Daniel Bone, Matthew S Goodwin, Matthew P Black, Chi-Chun Lee, Kartik Audhkhasi, and Shrikanth Narayanan. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *Journal of autism and developmental disorders*, 45:1121–1136, 2015.

[5] Lindsay K Butler, Chelsea La Valle, Sophie Schwartz, Joseph B Palana, Cerelia Liu, Natalie Peterman, Lue Shen, and Helen Tager-Flusberg. Remote natural language sampling of parents and children with autism spectrum disorder: Role of activity and language level. *Frontiers in Communication*, 7:820564, 2022.

[6] Ruchao Fan, Natarajan Balaji Shankar, and Abeer Alwan. Benchmarking children's asr with supervised and self-supervised speech foundation models. In *Interspeech 2024*, pages 5173–5177, 2024.

[7] Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, 2023.

[8] 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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[9] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

[10] Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in vision-language models via latent space steering. *arXiv preprint arXiv:2410.15778*, 2024.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[12] Catherine Lord, Traolach S Brugha, Tony Charman, James Cusack, Guillaume Dumas, Thomas Frazier, Emily JH Jones, Rebecca M Jones, Andrew Pickles, Matthew W State, et al. Autism spectrum disorder. *Nature reviews Disease primers*, 6(1):1–23, 2020.

[13] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael Rutter. The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30:205–223, 2000.

[14] Helen Tager-Flusberg, Sally Rogers, Judith Cooper, Rebecca Landa, Catherine Lord, Rhea Paul, Mabel Rice, Carol Stoel-Gammon, Amy Wetherby, and Paul Yoder. Defining spoken language benchmarks and selecting measures of expressive language development for young children with autism spectrum disorders. 2009.

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

[16] Joanne Volden et al. Neologisms and idiosyncratic language in autistic speakers. *Journal of autism and developmental disorders*, 21(2):109–130, 1991.

[17] Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao. Diarizationlm: Speaker diarization post-processing with large language models. *arXiv preprint arXiv:2401.03506*, 2024.

[18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[19] Anfeng Xu, Tiantian Feng, Helen Tager-Flusberg, Catherine Lord, and Shrikanth Narayanan. Data efficient child-adult speaker diarization with simulated conversations. *arXiv preprint arXiv:2409.08881*, 2024.

[20] Anfeng Xu, Kevin Huang, Tiantian Feng, Lue Shen, Helen Tager-Flusberg, and Shrikanth Narayanan. Exploring speech foundation models for speaker diarization in child-adult dyadic interactions. In *Interspeech 2024*, pages 5193–5197, 2024.

[21] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[22] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.