SocialRecNet: A Multimodal LLM-Based Framework for Assessing Social Reciprocity in Autism Spectrum Disorder

Xin-Yu Chen[1*], Yu-Ming Chen[2*], Chin-Po Chen[1], Bo-Hao Su[1], Susan Shur-Fen Gau[3], Chi-Chun Lee[1]
[1]*Department of Electrical Engineering, National Tsing Hua University, Taiwan*
[2]*Department of Psychiatry, Taipei Veterans General Hospital, Taiwan*
[3]*Department of Psychiatry, National Taiwan University Hospital and College of Medicine, Taiwan*
xychen@gapp.nthu.edu.tw, ymchen8@vghtpe.gov.tw, stu94116@gapp.nthu.edu.tw,
borrissu@gapp.nthu.edu.tw, gaushufe@gmail.com, cclee@ee.nthu.edu.tw

***Abstract*—Accurate assessment of social reciprocity is crucial for early diagnosis and intervention in Autism Spectrum Disorder (ASD). Traditional methods, often relying on unimodal data or lacking in cross-modal alignment, do not fully capture the complexity of social reciprocity. To address these limitations, we developed SocialRecNet, a novel Multimodal Large Language Model (MLLM) utilizing the Autism Diagnostic Observation Schedule (ADOS) dataset. SocialRecNet integrates conversational speech and text with the textual reasoning capabilities of LLMs to analyze social reciprocity across multiple dimensions. By effectively aligning speech and text, enhanced by properly designed prompts, SocialRecNet achieves an average Pearson correlation of 0.711 in predicting ADOS scores, marking a significant improvement of approximately 26.24% over the best-performing baseline method. This state of the art framework not only improves the prediction of social reciprocity scores but also provides deeper insights into ASD diagnosis and intervention strategies.**

***Index Terms*—Behavioral Signal Processing, Social Reciprocity, Autism Spectrum Disorder(ASD), Multimodal Large Language Model(MLLM), Autism Diagnostic Observation Schedule(ADOS)**

I. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by challenges in interpersonal interactions and the presence of restricted, repetitive behaviors. A key feature of ASD is impaired social reciprocity, which is the inability to dynamically engage in social exchanges using lexical choices, vocal patterns, and other nonverbal cues. This impairment significantly impacts the quality of life for children with ASD, highlighting the need for early identification and intervention. Current ASD assessments, however, are labor-intensive, time-consuming, and require highly trained examiners for consistent evaluations. An automated and objective tool that efficiently assesses social reciprocity can greatly enhance the quality of life for these children by streamlining the assessment process.

Despite the critical importance of assessing social reciprocity in ASD, accurately quantifying it is inherently challenging due to its multimodal nature. Traditionally, research has focused on characterizing social reciprocity through a prosodic-acoustic lens, exploring the phenomenon known as prosodic-acoustic entrainment. Various engineering methods such as correlation analysis, deep unsupervised learning, and attention mechanisms have been applied to study this phenomenon [1]–[5]. These methods have also been used to assess impairments in social reciprocity in ASD subjects from the prosodic-acoustic perspective [6]–[9]. However, this unimodal approach may not fully capture the intricate nature of social reciprocity. Recognizing this limitation, several research have shifted towards multimodal approaches that combine acoustics with linguistic embeddings, showing potential in analyzing deficits in social interactions characteristic of ASD [10]–[12]. Nevertheless, these methods often overlook the cross-alignment between speech and language, thus rendering the modeling of social reciprocity suboptimal.

To address these challenges, the integration of textual reasoning capabilities of Large Language Models (LLMs) with other modalities through cross-modal alignment in recent multimodal LLMs (MLLMs) offers a promising solution. However, existing audio language models, despite their capabilities in audio perception and reasoning, and response generation, are not directly suited for analyzing dyadic conversations [13]–[20]. In this work, we propose SocialRecNet, an MLLM specifically designed to assess the quality of social reciprocity by analyzing dyadic conversations in the Autism Diagnostic Observation Schedule (ADOS), a standardized tool used by clinical experts to evaluate social reciprocity and communication across multiple dimensions [21].

SocialRecNet integrates speeches, text transcripts, and a

[*] Equal contribution

specially designed prompt as inputs, employing a multiturn approach that aggregates several consecutive turns into a single contextual segment. The model enhances the characterization of social reciprocity within each segment through cross-attention mechanisms and leverages the intrinsic capabilities of LLMs with meticulous prompt design. Evaluated using the ADOS datasets, our framework demonstrates significant improvements in the accuracy of automated assessments across various dimensions of social reciprocity in ASD. This advancement facilitates more personalized interventions and precise monitoring, thereby improving ASD management.

II. METHODOLOGY

A. ADOS dataset

1) Dataset description: The dataset used in this study consists of speech data sample in 16K and manually transcribed text collected in collaboration with National Taiwan University Hospital[1] using the Autism Diagnostic Observation Schedule (ADOS) [21]. ADOS is a widely recognized diagnostic tool designed to assess autism severity through semi-structured, face-to-face observational interviews. Each ADOS session typically lasts between 40 to 60 minutes and consists of 14 distinct activities intended to elicit specific social and communicative behaviors. Our dataset comprises recordings from 105 children, all native Taiwanese Mandarin speakers, serving as the basis for our evaluation. Participants are categorized into subgroups of severe, moderate, and mild ASD based on the ADOS Calibrated Severity Scores (CSS) [22]. Detailed demographic information is provided in **Table 1**, with further details available in [23].

This research specifically focuses on the 'emotion' activity within the ADOS framework, which emphasizes spoken interactions. In this task, participants are encouraged to share and discuss personal experiences related to emotions, such as sadness and happiness. The duration of an emotion activity typically ranges from two to ten minutes, depending on the participant's engagement and the depth of their responses to the examiner's prompts. The interaction is conducted in a dialogic format, with the examiner guiding most exchanges, creating a dynamic back-and-forth conversation.

2) Score Selection: To concentrate on speech and language-based aspects of social reciprocity, we selected ADOS scores [21] that specifically pertain to conversational dynamics and verbal interaction. Items based on visual observations or related to restricted and repetitive behaviors were excluded. The selected scores are:

- **Conversation (CONV)**: Evaluates the ability to initiate and sustain a fluent conversation with the examiner.
- **Quality of Social Overtures (QSOV)**: Assesses the quality of actively expressing social intent towards the examiner.
- **Quality of Social Response (QSR)**: Summarizes the quality of social responses throughout the assessment.
- **Amount of Reciprocal Social Communication (ARSC)**: Measures the frequency of reciprocal exchanges, focusing on verbal communication.

[1]Approved by IRB: REC-10501HE002 and RINC-20140319.

TABLE I
DATABASE OVERVIEW: PARTICIPANT DEMOGRAPHICS

Participants	Age Mean (std)	Gender Male/Female	$ADOS_{social}$ Mean (std)
TD	13(4.14)	8/9	1.24(1.44)
Mild ASD	15.57(4.76)	11/3	4.29(1.48)
Moderate ASD	16.23(3.43)	34/5	7.13(1.42)
Severe ASD	17.4(3.94)	33/2	11.37(2.31)

- **Overall Quality of Rapport (OQR)**: Evaluates the extent to which the examiner must adjust their behavior to maintain effective interaction.

B. Problem definition

In this study, we aim to predict the ADOS scores($\hat{Y}$) by analyzing conversations between an examiner and a kid. To facilitate understanding, we first define the notations that will be consistently used throughout the manuscript. We denote $B_{E/K}$ as a block of signal, which includes both a speech turn and its corresponding transcript for either the examiner (E) or the kid (K). A complete emotion activity A comprises a sequence of such blocks with alternating speaker identities, represented as $A^i = (B^i_{E,0}, B^i_{K,0}, \ldots, B^i_{E,j}, B^i_{K,j})$, where i denotes the specific kid's identity, and j represents the turn index.

Building on previous definitions, we define a conversational turn as $U^i_l = (A^i_l, A^i_{l+1})$, where l is an element index within the sequence A^i. A contextual segment C^i_l of length n comprises n consecutive turn units, represented as $C^i_l = (U^i_l, U^i_{l+1}, \ldots, U^i_{l+n-1})$. Let C^i be the set of all contextual segments of kid i and Y^i be the label of A^i. The label of every $C \in C^i$ is assigned as Y^i.

Let $C^\dagger$ be all the set of all contextual segments of length n and Y be the label set of $C^\dagger$. The overall problem can be formulated as finding the optimized parameter θ^* of our proposed model such that

$$\theta^* = \arg\max_{\theta} P(\hat{Y} = Y|\theta; C^\dagger) \tag{1}$$

C. The SocialRecNet

The SocialRecNet framework employs a multimodal, multiturn approach with three key components: the Interlocutor Dynamics Extractor (IDE), the Modality Fusion Layer (MFL), and the Segment-Level Aggregation Layer (SLA). The IDE captures the interactions between two interlocutors by calculating the influence of each on the other, generating reciprocal embeddings. The MFL integrates turn-level embeddings from different modalities at each turn. Subsequently, the SLA aggregates these turn-level embeddings within a segment and maps them into the LLM's embedding space, resulting in segment-level embeddings that encapsulate the nuanced dynamics across the entire segment.

Therefore, the input of the LLM can be formulated as:

$$\text{Input}_{\text{LLM}} = \text{Prompt}_{\text{emb}} \oplus T_{\text{emb}} \oplus F_{\text{segment}} \tag{2}$$

where $\oplus$ denotes vector concatenation. $\text{Prompt}_{\text{emb}}$ represents prompt(as shown in **Fig. 2**) embeddings, encoded by the LLM encoder. T_{emb} corresponds to segment text tokens, which are kept unprojected to retain the raw information that is readable by the LLM. F_{segment} denotes segment-level features, and its computation is detailed in the subsequent subsections. Our

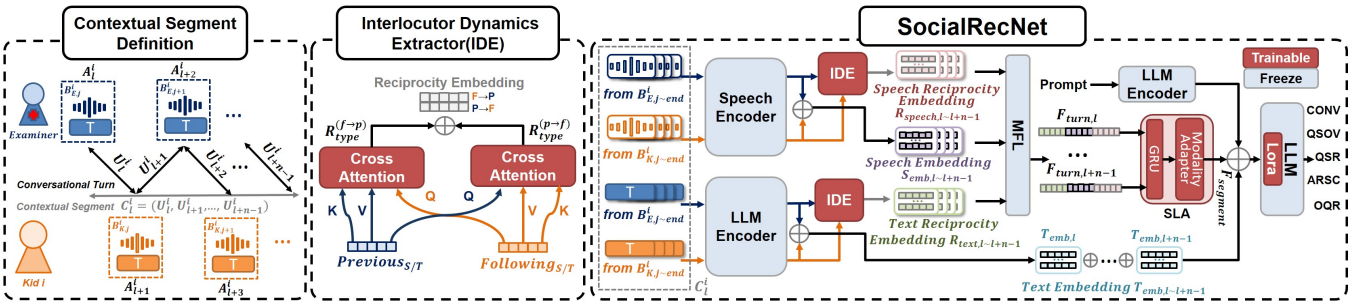

Fig. 1. Overview of SocialRecNet framework: The left section shows conversational turn units and contextual segments. The middle section highlights the Interlocutor Dynamics Extracto(IDE), which uses cross-attention to compute reciprocal embeddings for speech and text. The right section depicts how these embeddings are processed through Modality Fusion Layer(MFL), Segment-Level Aggregation Layer(SLA), and a LoRA-adapted LLM to predict ADOS scores.

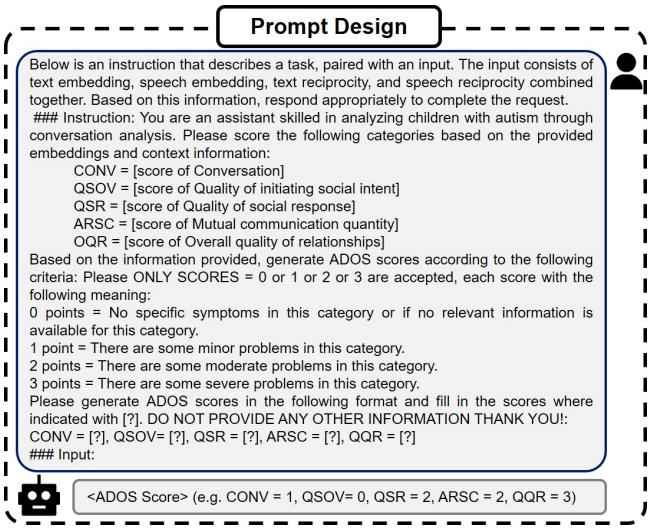

Fig. 2. Our Prompt Design

model is fine-tuned using the LoRA adapter [24], enabling the LLM to predict the diagnostic assessment of each segment:

$$\text{ADOS}_{\text{score}} = \text{LLM}(\text{Input}_{\text{LLM}}). \tag{3}$$

Following the scheme of clinicians, the overall diagnostic score for a kid is determined by averaging the segment scores. The overall structure and workflow of the model are illustrated in **Fig.1**. For full implementation details and the code, please refer to our GitHub repository[2].

1) Interlocutor Dynamics Extractor (IDE): The IDE employs multi-head cross-attention with 4 heads to capture the bidirectional relationships of speech (S) or text (T) between the previous and following utterances in a turn unit:

$$R^{(f\to p)}_{\{S,T\}} = \text{CrossAttention}(Previous_{\{S,T\}}, Following_{\{S,T\}}) \tag{4}$$

$$R^{(p\to f)}_{\{S,T\}} = \text{CrossAttention}(Following_{\{S,T\}}, Previous_{\{S,T\}}) \tag{5}$$

The outputs $R^{(f\to p)}_{\{S,T\}}$ and $R^{(p\to f)}_{\{S,T\}}$ are then concatenated to form reciprocal embeddings for speech and text, denoted as $R_{\text{text},l}$ and $R_{\text{speech},l}$, respectively.

2) Modality Fusion Layer (MFL): The MFL fuses turn-level information across modalities. It begins by averaging the embeddings along the temporal dimension to obtain compact representations. These are subsequently concatenated to form the turn-level feature representation:

$$F_{\text{fus},l} = \bar{S}_{\text{emb},l} \oplus \bar{R}_{\text{text},l} \oplus \bar{R}_{\text{speech},l} \tag{6}$$

[2]https://github.com/xinyu0308/SocialRecNet

where $S_{\text{emb},l}$ represents the turn-level speech embeddings obtained from the speech encoder, while $R_{\text{text},l}$ and $R_{\text{speech},l}$ are the reciprocal embeddings from the IDE for text and speech, respectively.

3) Segment-Level Aggregation Layer (SLA): The SLA employs a gated GRU [25] followed by a modality adapter to aggregate turn-level features $F_{\text{fus},l}$ across a segment. The GRU, with a hidden size of 512 and 2 layers, processes fused features to capture temporal dependencies and outputs the hidden states sequence. This sequence is then refined by a modality adapter based on BLSP [18], producing the final segment-level embedding:

$$F_{\text{segment}} = \text{SLA}(\text{F}_{\text{fus},l} \oplus \text{F}_{\text{fus},l+1} \oplus \ldots \oplus \text{F}_{\text{fus},l+n-1}) \tag{7}$$

III. Experiment

A. Experiment Setup

Our dataset is randomly split by participants into an 8:2 ratio for both ASD and TD groups, ensuring a speaker-independent approach. Specifically, 84 participants are used for training and 21 for testing. We use the Vicuna-7B-v1.5 model [27] as the LLM base, encoding text features with a hidden dimension of 4096. Speech features are extracted using self-supervised TERA embeddings [28], resulting in 768-dimensional vectors that capture speech patterns.

Training is conducted over 3 epochs with a batch size of 12, using the AdamW optimizer at a learning rate of 5e-5. The LoRA adapter has a rank of 8, an alpha of 16, and a dropout of 0.05. Training on an NVIDIA A100 GPU (80 GB) takes about 12 hours. Performance is evaluated using Pearson Correlation and Mean Absolute Error (MAE) to assess prediction accuracy and reliability.

B. Baseline Models

- **Llama 3.1**: We use Llama 3.1 8B [29] without fine-tuning, feeding it the entire raw session conversation as input to evaluate its performance on our task
- **Chen et al.**: Chen et al. [12] proposed a conversation-level multimodal approach utilizing BERT for feature extraction. We apply their model to our dataset to assess its effectiveness in our context.
- **Lahiri et al.**: Rimita Lahiri et al. [5] introduced the Contextual Entrainment Distance (CED) metric. We adapt their structure and train it on our dataset to compare its performance with our proposed method.

TABLE II
COMPARISON WITH OTHERS' WORK AND ABLATION STUDY
SCORES IN FORMAT: PEARSON SCORE (MAE), NOTE: **BOLD** DENOTES BEST SCORES.

Category	Model	CONV	QSOV	QSR	ARSC	OQR	Average
Baseline	Llama3.1	.254 (1.08)	.035 (.966)	.139 (.830)	.229 (1.682)	.360 (0.705)	.203 (1.053)
	Chen et al. [12]	.565 (.58)	.496 (.818)	.556 (.498)	.513 (.690)	.685 (.653)	.563 (.648)
	Lahiri et al. [5]	.040 (.426)	.040 (.553)	.021 (.275)	.019 (.574)	.033 (.511)	.031 (.468)
	Yang et al. [26]	.242 (.473)	.140 (.517)	.193 (.209)	.205 (.703)	.190 (.420)	.194 (.464)
Our Proposed	SocialRecNet	**.690** (.539)	**.628** (.607)	**.676** (.312)	**.807** (.525)	**.753** (.513)	**.711** (.499)
Ablation Study	Speech Only	.101 (.663)	.313 (.677)	.341 (.365)	.322 (.621)	-.048 (.687)	.206 (.603)
	Text Only	-.009 (.713)	.123 (.611)	.551 (.295)	-.074 (.655)	.137 (.641)	.145 (.583)
	Text Reciprocity Only	.403 (.559)	.439 (.626)	.429 (.316)	.491 (.623)	.553 (.571)	.463 (.539)
	Speech Reciprocity Only	.146 (.674)	.378 (.738)	.369 (.389)	.409 (.592)	-.035 (.701)	.253 (.619)
	Speech and Text	.541 (.583)	.394 (.657)	.645 (.312)	.587 (.566)	.363 (.620)	.506 (.548)
	Text with its Reciprocity	.575 (.563)	.606 (.648)	.537 (.335)	.534 (.594)	.724 (.510)	.595 (.530)
	Speech with its Reciprocity	.476 (.605)	.439 (.627)	.459 (.332)	.418 (.620)	-.036 (.672)	.351 (.571)
	Without Text	.502 (.594)	.470 (.666)	.565 (.312)	.494 (.599)	.457 (.584)	.498 (.551)

- **Yang et al.**: Yahan Yang et al. [26] developed various classifiers to predict children's conversation quality. We replicate their approach using our dataset to benchmark its predictive capability against our framework.

C. Results and Analysis

To comprehensively evaluate the effectiveness of our proposed model, **Table II** shows the correlation coefficients and MAE between the predicted and expert-labeled ADOS scores, including comparisons with baseline models and results from the ablation study.

1) Comparison with Baseline: Our proposed model shows a 26.24% improvement in average Pearson correlation over the best baseline, proposed by Chen et al. [12], mainly due to the integration of LLM. This advantage allows the model to effectively capture the complexities of verbal content and nuanced language patterns, outperforming baselines in key areas. For instance, in ARSC, the model demonstrates a 57.40% improvement, highlighting its ability to analyze reciprocal exchanges with greater precision. In QSR, it achieves a 21.56% improvement, reflecting its strength in assessing response adequacy by leveraging both the content and context of interactions.

Compared to purely text-based models like Llama3.1, the limitations of relying solely on language are evident. Although one might assume that text-centric LLMs could effectively track responses for metrics such as QSR and ARSC, they overlook critical paralinguistic cues such as intonation, articulation, pause, and prosody, which are essential for clinicians to accurately assess the adequacy of responses.

2) Ablation Study: The results provide valuable insights into how different feature sets influence model performance across social reciprocity metrics. For the CONV score, focusing on turn-taking in dialogues, the "Text with its Reciprocity" configuration ranks highly, slightly below the "Speech and Text" but outperforming the "Speech with its Reciprocity." This highlights that incorporating reciprocity in text features significantly enhances the model's ability to capture conversational dynamics, likely due to the richer representation of reciprocal interactions in text. Notably, removing text features results in a marked drop in performance, underscoring the critical role of text in capturing the subtleties of conversational reciprocity.

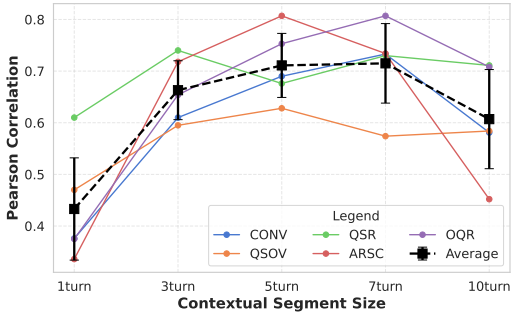

Fig. 3. Comparison of Contextual Segment Sizes in Pearson Correlation

For QSOV and QSR scores, which evaluate social overtures and responses, our model achieves the best results, underscoring the importance of a holistic integration of features. This approach effectively captures both the proactive and reactive dimensions of social communication. The importance of reciprocity features is particularly pronounced in the ARSC score, where our model significantly outperforms the second-best "Speech and Text" model by 37.5%. These findings reinforce the necessity of combining speech, text, and their reciprocal elements to accurately model the complexities of social reciprocity across various dimensions.

3) Analysis of Contextual Segment Size: In analyzing the relationship between model performance and contextual segment size, **Fig. 3** illustrates the trends in Pearson correlation scores across various metrics with respect to different contextual segment sizes, showcasing the model's adaptability to conversational contexts. For CONV, performance peaks at 7 turns, suggesting that moderate turn lengths optimize conversational fluency. QSOV performs best with shorter turns, which are appropriate for assessing social intent. Both QSR and ARSC show significant improvement with 3 to 5 turns, making this range ideal for evaluating response quality and reciprocal interaction. Similarly, OQR peaks at 5 turns, which is also optimal for assessing rapport. Performance typically plateaus or declines beyond 5 turns, indicating that this length provides the best balance between offering sufficient context for nuanced assessments and avoiding information overload.

IV. CONCLUSION AND FUTURE WORK

This study introduces SocialRecNet, a multimodal LLM specifically designed to multifacetedly analyze social reciprocity by integrating conversational speeches and dialogue transcripts across multiple conversational turns. Utilizing specially designed prompts and cross-attention mechanisms, this model offers comprehensive insights into social dynamics, facilitating nuanced ASD assessments, personalized interventions, and precise monitoring. Future work will aim to enhance SocialRecNet's ability to capture complex interactions by incorporating visual modalities and broadening its application to various domains of social interaction.

REFERENCES

[1] R. Levitan and J. B. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," 2011.

[2] C.-C. Lee, M. Black, A. Katsamanis, A. C. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[3] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, 2012, pp. 11–19.

[4] M. Nasir, B. Baucom, C. Bryan, S. Narayanan, and P. Georgiou, "Modeling vocal entrainment in conversational speech using deep unsupervised learning," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1651–1663, 2020.

[5] R. Lahiri, M. Nasir, C. Lord, S. H. Kim, and S. Narayanan, "A context-aware computational approach for measuring vocal entrainment in dyadic conversations," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] K. Ochi, N. Ono, K. Owada, M. Kuroda, S. Sagayama, and H. Yamasue, "Entrainment analysis for assessment of autistic speech prosody using bottleneck features of deep neural network," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8492–8496.

[7] S. Z. Asghari, S. Farashi, S. Bashirian, and E. Jenabi, "Distinctive prosodic features of people with autism spectrum disorder: a systematic review and meta-analysis study," *Scientific reports*, vol. 11, no. 1, p. 23093, 2021.

[8] S. P. Patel, K. Nayar, G. E. Martin, K. Franich, S. Crawford, J. J. Diehl, and M. Losh, "An acoustic characterization of prosodic differences in autism spectrum disorder and first-degree relatives," *Journal of Autism and Developmental Disorders*, vol. 50, pp. 3032–3045, 2020.

[9] C. J. Wynn, S. A. Borrie, and T. P. Sellers, "Speech rate entrainment in children and adults with and without autism spectrum disorder," *American Journal of Speech-Language Pathology*, vol. 27, no. 3, pp. 965–974, 2018.

[10] S. Cho, M. Liberman, N. Ryant, M. Cola, R. T. Schultz, and J. Parish-Morris, "Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations." in *Interspeech*, 2019, pp. 2513–2517.

[11] H. MacFarlane, A. C. Salem, L. Chen, M. Asgari, and E. Fombonne, "Combining voice and language features improves automated autism detection," *Autism Research*, vol. 15, no. 7, pp. 1288–1300, 2022.

[12] C.-P. Chen, S. S.-F. Gau, and C.-C. Lee, "Learning converse-level multimodal embedding to assess social deficit severity for autism spectrum disorder," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[13] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, "Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities," *arXiv preprint arXiv:2402.01831*, 2024.

[14] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," *arXiv preprint arXiv:2305.10790*, 2023.

[15] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 090–18 108, 2023.

[16] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," *arXiv preprint arXiv:2305.11000*, 2023.

[17] D. Zhang, X. Zhang, J. Zhan, S. Li, Y. Zhou, and X. Qiu, "Speechgpt-gen: Scaling chain-of-information speech generation," *arXiv preprint arXiv:2401.13527*, 2024.

[18] C. Wang, M. Liao, Z. Huang, J. Lu, J. Wu, Y. Liu, C. Zong, and J. Zhang, "Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing," *arXiv preprint arXiv:2309.00916*, 2023.

[19] C. Wang, M. Liao, Z. Huang, J. Wu, C. Zong, and J. Zhang, "Blsp-emo: Towards empathetic large speech-language models," *arXiv preprint arXiv:2406.03872*, 2024.

[20] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Extending large language models for speech and audio captioning," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 236–11 240.

[21] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. 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*, vol. 30, pp. 205–223, 2000.

[22] K. Gotham, A. Pickles, and C. Lord, "Standardizing ados scores for a measure of severity in autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 39, pp. 693–705, 2009.

[23] C.-P. Chen, S. S.-F. Gau, and C.-C. Lee, "Toward differential diagnosis of autism spectrum disorder using multimodal behavior descriptors and executive functions," *Computer Speech & Language*, vol. 56, pp. 17–35, 2019.

[24] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[25] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (gru) neural networks," in *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*. IEEE, 2017, pp. 1597–1600.

[26] Y. Yang, S. Cho, M. Covello, A. Knox, O. Bastani, J. Weimer, E. Dobriban, R. Schultz, I. Lee, and J. Parish-Morris, "Automatically predicting perceived conversation quality in a pediatric sample enriched for autism," in *Interspeech*, vol. 2023. NIH Public Access, 2023, p. 4603.

[27] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, vol. 2, no. 3, p. 6, 2023.

[28] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[29] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.