ASD-iLLM:An Intervention Large Language Model for Autistic Children based on Real Clinical Dialogue Intervention Dataset

Shuzhong Lai[1,2,3,5], Chenxi Li[4], Junhong Lai[1,2,3,6], Yucun Zhong[1,2,3,6], Chenyu Yan[4]
Xiang Li[2], Haifeng Li[4], Gang Pan[6,7], Lin Yao[1,2,3,6,7] *, Yueming Wang[2,6]
[1]MOE Frontiers Science Center for Brain and Brain-Machine Integration, Zhejiang University
[2]Nanhu Brain-Computer Interface Institute [3]Department of Neurobiology, Affiliated Mental Health Center and Hangzhou Seventh People's Hospital, Zhejiang University School of Medicine [4]Children's Hospital Zhejiang University School of Medicine
[5]Polytechnic Institute, Zhejiang University
[6]College of Computer Science and Technology, Zhejiang University
[7]State Key Laboratory of Brain-Machine Intelligence

Abstract

Currently, leveraging large language models (LLMs) for autism intervention is a significant yet challenging task, particularly when directly employing LLMs as an intervention doctor. Researchers have mainly focused on using prompt engineering for role play as an intervention doctor and integrating auxiliary elements such as visual stimuli to enhance the sensory experience of the intervention, while neglecting the challenge that LLMs' inherent dialogue style and intervention strategies do not meet the requirements of clinical dialogue interventions. To fill the gap, we propose a comprehensive framework for training LLMs to conduct dialogue interventions in accordance with the principles of Applied Behavior Analysis (ABA) which is commonly used by clinicians. Specifically, we collected clinical recordings of dialogue interventions for autistic children and constructed the topic dialogue dataset ASD-iLLM-8k. By incorporating the system prompt based on the ABA and ASD-iLLM-8k dataset, we fine-tuned LLMs to develop ASD-iLLM. We also proposed a role-play strategy in which LLMs act as autistic children to comprehensively evaluate the doctor model's capabilities at the dialogue level. Extensive experiments indicate that ASD-iLLM outperforms existing models in both automatic and human evaluation, with intervention strategies and dialogue style more closely resembling those of clinical intervention doctors. Our dataset, model, and code are available on https://github.com/Shuzhong-Lai/ASD-iLLM.

1 Introduction

Autism Spectrum Disorder (ASD) is one of the most common heterogeneous neurodevelopmental disorders in children, characterized by social interaction impairment and repetitive or stereotypical behavior patterns (Association et al., 2013). These manifestations create substantial challenges for them in social communication, severely affecting their educational and daily activities (Fuller and Kaiser, 2020). To alleviate symptoms of social impairment, clinicians train children to exhibit appropriate behaviors in different scenarios through dialogue, supplemented by visual stimuli or body actions. In practical interventions, clinicians adhere to Applied Behavior Analysis (ABA) (Cooper et al., 2007) principles commonly used for behavioral intervention to stimulate the development of various social skills for autistic children.

The global prevalence of autism is rising annually, reaching approximately 1% (Zeidan et al., 2022). Although timely diagnosis and treatment

Figure 1: An example from the test set which displays the responses of the human therapist, ASD-iLLM, and GPT-4.1 about the same dialogue history.

*Corresponding author: lin.yao@zju.edu.cn

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8058–8079
November 4-9, 2025

can significantly improve the core symptoms of individuals with autism (Estes et al., 2015), this often requires years of effort and a substantial financial burden. Moreover, due to uneven healthcare resources across regions, a considerable number of children do not receive effective interventions, leading to social disconnection and living independently (Liu et al., 2023; Lin et al., 2025). Therefore, there is an urgent need for the emergence of new forms of intervention.

With the rise of large language models (LLMs), researchers believe that LLMs' capabilities hold promise for application in autism therapy (Cho et al., 2023; Ciobanu et al., 2024; Jang et al., 2024). Current research primarily concentrates on two types. Firstly, augmenting various aspects of existing interventions with LLM capabilities. For example, providing professional advice to parents (Ren et al., 2023; Chu et al., 2024; Wang and Tang, 2024; He et al., 2024) or offering cues to assist children in communication (Jafri, 2024; Haroon and Dogar, 2024). However, these methods merely indirectly influence the treatment of autistic children and cannot serve as an intervention. Secondly, investigating whether LLMs can substitute clinicians in delivering interventions (Ren et al., 2023; Tang et al., 2024; Deng et al., 2024). But these methods only constrain LLMs to act as intervention doctors through prompt engineering, and their conversational style and intervention strategies still differ significantly from real clinical intervention dialogues, as shown in Figure 1.

To address this, we propose **A**utism **S**pectrum **D**isorder **i**ntervention **L**arge **L**anguage **M**odel (ASD-iLLM), which is designed for topic dialogue intervention for autistic children. We collected real clinical intervention dialogue recordings of autistic children and constructed a multi-turn dialogue dataset, ASD-iLLM-8k. Following ABA principles, we designed the system prompt and used ASD-iLLM-8k for fine-tuning on LLMs to learn the conversational style and intervention strategies used by doctors. Results from both automatic and human evaluations indicate that ASD-iLLM outperforms existing state-of-the-art (SOTA) LLMs in topic dialogue intervention tasks for autistic children. The whole framework for ASD-iLLM is shown in Figure 2.

Our contributions are as follows:

- To the best of our knowledge, we are the first team to construct a complete framework for training LLMs for autism intervention and build an available Chinese clinical autism intervention dialogue dataset, ASD-iLLM-8k.
- We propose ASD-iLLM for dialogue intervention with autistic children, which closely emulates the conversational style of intervention doctors and engages with children following ABA principles.
- We introduced a role-play strategy, employing a method of randomly selecting response intents, to enable LLMs to simulate autistic children. Comprehensive experiments indicate that ASD-iLLM significantly outperforms existing models in topic dialogue intervention tasks for autistic children.

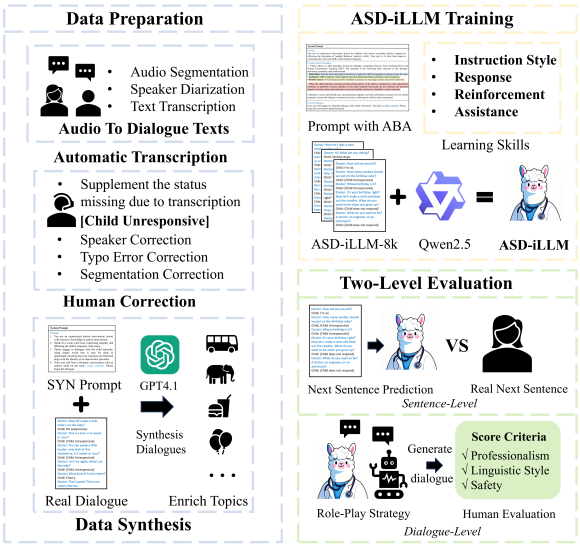

Figure 2: The whole framework for ASD-iLLM.

2 Related Work

With the substantial increase in the scale and capabilities of LLMs, more researchers are exploring the application of LLMs in autism treatment (Ciobanu et al., 2024). LLM's conversational and instruction-following abilities allow it to adapt to various dialogue scenarios. For adults with autism, the researchers (Li et al., 2024, 2025) employ virtual reality (VR) to create scenarios and utilize LLMs for dialogue generation, aiding them in practicing communication skills for job seeking. Additionally, (Mishra and Conn Welch, 2024; Mishra et al., 2024) constructed interactive scenarios for intervention using the NAO robot equipped with GPT-2 (Radford et al., 2019), and assessed its effec-

tiveness through conversations with experts. Echo-Teddy (Lee et al., 2025) developed a social dialogue robot shaped as a toy teddy bear, embedded with the LLM for conversation, showcasing the potential in supporting children. In addressing social impairments, ChatASD Therapist (Ren et al., 2023) utilizes GPT-4 (Achiam et al., 2023) and facial video generation technology for intervention, while ASD-Chat (Deng et al., 2024) employs a design paradigm integrating Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP) (Sundberg, 2008) and ChatGPT for topic dialogue interventions. (Pai et al., 2024) created a dataset of user-consultant interactions for fine-tuning Falcon7B (Almazrouei et al., 2023) and integrated Yolov5 (Redmon et al., 2016) and MobileNet (Howard et al., 2017) to enhance visual ability. However, the paper lacks detailed information about the dataset and rigorous quantitative analysis. For emotional support, EmoEden (Tang et al., 2024) utilizes GPT-4 for dialogue and employs Midjourney to generate conversational scenarios, assisting children in emotion recognition and expression training.

3 Dataset

Our goal is to modify the conversational style and intervention strategy of LLM to more closely resemble real clinical dialogue intervention scenarios. However, there is currently no publicly available dataset for autism dialogue interventions. Therefore, we have created a multi-turn dialogue dataset for intervention between doctors and autistic children, named **ASD-iLLM-8k**.

3.1 Data Collection

To ensure the authenticity and quality of the data, we collaborated with six treatment centers for autistic children after obtaining ethical approval, involving a total of 20 experienced clinicians. With the full informed consent of both parents and children, we used the recording device (H1-Pro, iFlytek Inc., China) to collect audio recordings during topic dialogue interventions. For clearer audio reception, the voice recorder was placed in the chest pocket of the doctor's coat.

The standards for data collection are as follows: Firstly, autistic children often suffer from language development delays. Their chronological age does not necessarily reflect their language abilities. Therefore, we included children whose actual language development age was greater than 24 months. Secondly, the researchers (Dekker et al., 2019; Hanrahan et al., 2020; van der Wilt et al., 2022) have indicated that topic dialogue interventions can alleviate social impairment for children. Therefore, we collected recordings in the form of topic dialogues, with each record focusing on a specific topic. Lastly, the recording sampling rate is 16,000 Hz, and the files are stored in WAV format.

Category	Doctor	Child
Participants Num	20	74
Turns per dialogue	13.55	10.17
Char. per sentence	18.94	4.40
Distinct-2	$76.74_{\pm 8.43}$	$69.03_{\pm 17.97}$
Distinct-3	$91.12_{\pm 7.26}$	$66.25_{\pm 22.93}$

Table 1: Data statistics of ASD-iLLM-8k dataset.

3.2 Data Processing

We employed a two-stage processing approach to transcribe the original audio recordings into multi-dialogue texts.

Automatic Transcription First, we utilized existing automated transcription tools to convert the original recordings into multi-turn dialogues. DeepFMSN (Zhang et al., 2018) was employed for audio segmentation, Cam++ (Wang et al., 2023a) was used to identify whether the segmented audio belonged to the child or the doctor. Paraformer (Gao et al., 2022) was utilized for the audio-to-text transcription. CT-Transformer (Chen et al., 2020) was applied to predict the punctuation.

Manual Transcription We aimed to enhance the quality of the multi-turn dialogue text through manual transcription.

It is noteworthy that during conversations between the doctor and the child, the child may become unresponsive due to loss of attention or lack of sentence understanding. In such cases, the doctor will take appropriate actions, such as repeating questions or simplifying the inquiries, to maintain the dialogue and encourage the child to engage. These are the intervention strategies that we expect the model to learn. However, this state is lost in the audio-to-text transcription process, as automated transcription cannot identify the moments when the child is silent. Therefore, we require the assistance of specialized intervention doctors to help reconstruct the child's unresponsive state and enhance the quality of the multi-turn dialogue text.

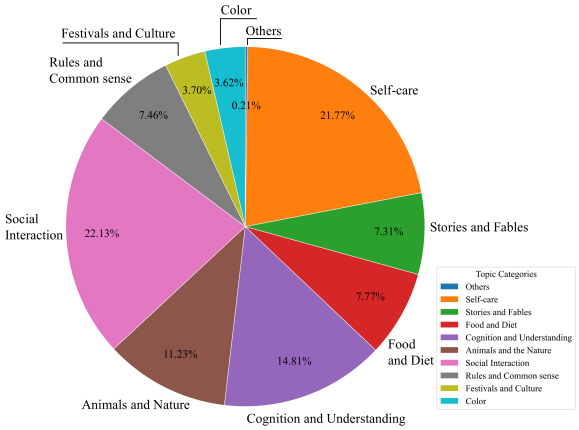

Figure 3: Topic distribution of ASD-iLLM-8k dataset.

The annotating doctors inserted soft labels **[Child Unresponsive]** into the dialogue to supplement the child's unresponsive state. Meanwhile, they correct errors present in the automatic transcription, such as inaccuracies in audio segmentation, speaker identification mistakes, and typographical errors. Additionally, to protect the privacy of participants, we de-identified the data by replacing specific names or places with aliases.

3.3 Data Augmentation

To diversify the topics and dialogue scenarios of the dataset, we utilized GPT-4.1 to synthesize multi-turn dialogues. We used cleaned real dialogues mentioned above as references to have GPT-4.1 mimic their style to generate new dialogues on different topics, thus enriching the dataset. We derived 27 subtopics from 10 main topics clustering from real clinical dialogues for data synthesis. For each instance of real multi-turn dialogue, we used it as a reference dialogue to generate 27 different subtopic synthetic dialogues to enrich the dataset.

3.4 ASD-iLLM-8K Dataset

We collected 64.2 hours of audio data and transcribed it into 751 instances of topic multi-turn dialogues. After cleaning, we obtained 287 high-quality real multi-turn dialogues. We then used GPT-4.1 for data synthesis, resulting in a total of 8,035 instances of topic multi-turn dialogues. We released this dataset as the **ASD-iLLM-8k**.

The statistical information of ASD-iLLM-8k dataset is shown in Table 1. On average, each topic dialogue lasts 13.55 rounds, with the child exhibiting an average of 3.18 unresponsiveness states per dialogue. Furthermore, during the intervention, both the doctor and the child use relatively few characters per sentence, with doctors averaging 18.94 characters and children just 4.4 characters. The doctor needs to use concise and easily understandable sentences to encourage the child to engage, while the child's language development delays and social impairments significantly reduce their frequency of responses and word count. We used Distinct-n (Li et al., 2016) metrics to assess the diversity of the dataset, revealing that the linguistic richness of doctors is relatively high. Also, the larger standard deviations for autistic children suggest that their articulation is less stable. More details for data cleaning, topics description, synthesis prompt, and how to assess the synthesis data quality are provided in Appendix B. The topic distribution of ASD-iLLM-8k is illustrated in Figure 3, showing a balanced distribution of topics.

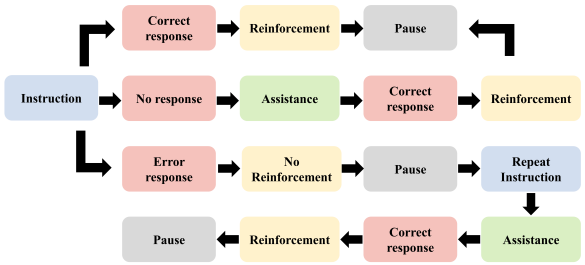

Figure 4: The workflow of DTT from ABA.

4 Methodology

4.1 Prompt Design with ABA

ABA is a structured approach commonly used as a behavioral therapy in treating autism (Foxx, 2008; Roane et al., 2016). Specifically, doctors integrate Discrete Trial Teaching (DTT) and Natural Environment Teaching (NET) methods from ABA to intervene with autistic children.

DTT consists of five fundamental elements: instruction, response, reinforcement, assistance, and pause. The basic flow is illustrated in Figure 4.

Instructions are issued by the doctor, who ensures they are concise and easy for the child to comprehend. Through these instructions, the doctor guides the child in understanding language and learning social skills.

Response refers to the child's reaction to the instruction. In topic dialogue intervention scenarios, the response is the child's verbal expression.

Reinforcement involves providing stimuli when a child responds to an instruction. The purpose of reinforcement is to encourage the continued occurrence of appropriate behaviors, while inappropriate

System Prompt

Role
You are an experienced intervention doctor for children with autism, providing effective support by following the principles of Applied Behavior Analysis (ABA). Your goal is to help them improve communication and social skills within thematic dialogues.

Intervention Principles
1. Please adhere to ABA principles during the dialogue, integrating Discrete Trial Teaching (DTT) and Natural Environment Teaching (NET). Pay attention to the following three elements in the dialogue: instruction, assistance, and reinforcement.
• **Instruction**: Provide clear and simple instructions to guide the child in engaging in dialogue about the topic.
• **Assistance**: Offer moderate verbal support when the child needs help, promoting correct responses.
• **Reinforcement**: Provide timely positive feedback and praise to encourage correct and positive behavior.

2.When the child responds correctly, provide reinforcement; if the child is unresponsive, offer appropriate assistance to promote a correct response; if the child responds incorrectly, do not reinforce the incorrect response. Instead, reissue the instruction or provide suitable assistance to facilitate a correct response.

3.Maintain a warm and friendly tone, demonstrating empathy, and offer affirmation and praise for the child's responses. Ensure the dialogue is natural and concise, allowing the child to easily understand.

Start Dialogue
Now, you will engage in a thematic dialogue with a child with autism. The topic is: ***{topic_content}***. Please begin the conversation about this theme.

Figure 5: System prompt with ABA principles translated from Chinese.

behaviors diminish or disappear due to a lack of reinforcement. Reinforcement can be physiological, such as favorite foods or toys, or social, such as praise. In social dialogue interventions, we emphasize social reinforcement, enhancing the child's socialization through verbal praise and empathy.

Assistance refers to the support provided by the doctor when an autistic child struggles to respond. This support can take the form of physical, visual, or verbal aid. It must be timely and appropriate to prevent causing feelings of failure or creating dependence on assistance. In topic dialogue interventions, assistance typically takes the form of verbal aids, such as rephrasing questions, breaking down problems, or prompting answers.

Pause refers to the brief interval between each trial, allowing the child time to reflect on and internalize their response and the reinforcement.

Training detached from real-life scenarios is meaningless. Therefore, during DTT, doctors integrate interventions with NET methods. NET involves simulating daily or social scenarios through role-playing to conduct dialogue, helping children better handle daily social situations.

According to the aforementioned ABA principles, we use prompt engineering to guide the LLM to follow the ABA principles during topic dialogue interventions. The system prompt we designed for training and evaluation is shown in Figure 5.

4.2 Instruction Tuning

We trained LLMs using instruction tuning (Wang et al., 2023b), enabling it to adapt its dialogue style close to an intervention doctor and learn corresponding intervention strategies to follow ABA. We employed the LoRA (Hu et al., 2022) method for fine-tuning on the ASD-iLLM-8k dataset.

Algorithm 1: Dialogue Generation via Role Play

Input : System prompt with topic S, number of turns N
Output: Generated dialogue G

f_{child}: LLM designed to role-play as a child with autism
f_{doctor}: LLM designed to role-play as a doctor for intervention
GenerateChildResponse(): Function to generate child's response
$IntentSet$: The set composed of four types of child response intents
$i \leftarrow 1$
Function `GenerateDialogue`(S, N, f_{child}, f_{doctor})**:**
  for $i \neq N$ **do**
    $X_i \leftarrow f_{doctor}(S)$;
    `// Randomly select one intent for the current response`
    $I \leftarrow random.choice(IntentSet)$;
    `// Simulate the fleeting attention span of a child`
    $Y_i \leftarrow GenerateChildResponse(I, [X_i, [X_{i-1}, Y_{i-1}]], f_{child})$;
    $S \leftarrow S.append(X_i)$;
    $S \leftarrow S.append(Y_i)$;
    $i \leftarrow i + 1$;
  end
  `// Return all dialogues except system prompt`
  return S[1:];

Figure 6: Pseudocode for generating dialogues through role-playing strategy.

For the $t+1$ round of dialogue, the LLM generates the doctor's response in the form of the following conditional probability:

$$\mathcal{P}(a_{t+1}|S, a_{1:t}, u_{1:t}; \theta) \tag{1}$$

Where $a_{1:t}$ and $u_{1:t}$ represent the previous dialogue history of t rounds between the doctor and the child, and a_{t+1} is the doctor's response to be predicted. θ is denoted as the original parameters of LLMs. The loss function for training can be defined as:

$$\mathcal{L}(\theta') = -\sum_{t=1}^{T} log\mathcal{P}(a_{t+1}|S, a_{1:t}, u_{1:t}; \theta') \tag{2}$$

θ' is denoted as the update parameters of LLMs. LoRA employs the concept of low-rank approximation by introducing auxiliary matrices A and B with smaller intrinsic dimensions for parameter updates, thereby adapting to downstream tasks:

$$\triangle W = A \times B \tag{3}$$

$\triangle W$ denotes the updated weights of LLMs, which can be decomposed into the updates of low-rank matrices $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times d}$, $r \ll d$. r is the selected LoRA rank.

4.3 Evaluation Design

To comprehensively evaluate the model's performance in the autistic children's topic dialogue intervention task, we propose an innovative role-playing

Dimension	Category	Explanation
Professionalism	Principle	Dialogues adhere to the DTT method or NET approach outlined.
	Assistance	Doctor provides timely and appropriate assistance to the child.
	Reinforcement	Doctor's feedback is positive and effectively reinforces the child's correct responses or positive behaviors.
	Personalization	Doctor makes personalized adjustments based on the child's needs and responses.
Linguistic	Relevance	Dialogue contents must focused on the topic.
	Style	Linguistic style aligned with the clinical intervention style, ensuring responses are simple and easily understandable.
	Fluency	Dialogue is natural and fluent, avoiding complex sentences that may be difficult for children to comprehend.
Safety	Guidance	The content include suitable guidance or suggestions, avoiding any potential misdirection.
	Privacy	The Child's privacy is strictly protected during the dialogue.
	Content	Dialogues avoid harmful content for children.

Table 2: The evaluation criteria for LLM capability dimensions, which are divided into 3 dimensions and ten categories with their explanations. Scores range from 0 to 4, with higher scores indicating better quality for the doctor's responses.

strategy to evaluate the model's performance at the dialogue level, where GPT-4o (Hurst et al., 2024) acts as an autistic child for dialogue interactions.

To simulate the chaotic and socially impaired state of an autistic child, we categorized the child's responses into four states based on ABA principles: *unresponsive*, *repeat*, *correct response*, and *incorrect response*. Each time a child's response is needed, we randomly sample an intent from the four intents to generate the child's response Y_i. When the intent is unresponsive, return *[Child Unresponsive]* directly. When the intent is *repeat*, to simulate a real-life scenario where a child instinctively echoes the last few words spoken by the doctor, we use jieba[1] package for word segmentation of the doctor's current instruction, excluding stop words, and returning the last word as the child's response. For intents indicating *correct response* or *incorrect response*, we use GPT-4o with different prompts to simulate the child's response.

Furthermore, to simulate the short-term attention of a child, we will concatenate the current doctor's instructions X_i with the previous conversation content $[X_{i-1}, Y_{i-1}]$ to create a dialogue history for the LLM to generate a response.

We assume that LLMs designed for autism intervention can cope with the potentially disordered expressions of autistic children while still employing their intervention strategies and sustaining the topic dialogue effectively.

By providing a topic and dialogue rounds, we can simulate conversations between the doctor model and the autistic child model, allowing us to test the LLMs' ability to intervene at the dialogue level. The pseudocode for the dialogue generation workflow is shown in Figure 6. For more details, please refer to the Appendix C.

5 Experiment

5.1 Baseline

When selecting baseline models, we believed that explicit dialogue proficiency does not require specialized medical domain knowledge but rather focuses on dialogue style and intrinsic intervention logic. So we chose current popular LLMs with strong conversational capabilities as baselines, such as GPT-4.1, GPT-4o-mini, GPT-4.1-mini, Gemini2.0-flash (Team et al., 2023), Deepseek-v3 (Liu et al., 2024). We selected the current popular open-source LLMs in the Chinese domain as the backbone for fine-tuning, such as Qwen2.5-7B-Instruct (Yang et al., 2024a), Llama-3-Chinese-8B-Instruct (Cui et al., 2023), Yi-1.5-9B-Chat-16K (Young et al., 2024), InternLM3-8B-Instruct (Cai et al., 2024), Baichuan2-7B-Chat (Yang et al., 2023a), and GLM-4-9B-Chat (GLM et al., 2024).

[1] https://github.com/fxsjy/jieba

Model_Name	BLEU	GLEU	R-1.	R-2.	R-L.	MET.	BS.	BGE.
GPT-4o-mini	13.62	17.89	28.74	7.38	22.95	24.75	66.68	65.69
GPT-4.1	11.35	14.64	25.48	6.49	19.96	24.33	65.14	65.02
GPT-4.1-mini	13.24	17.09	28.02	7.16	22.28	25.04	66.24	65.59
Gemini-2.0-flash	11.91	15.14	27.30	7.41	20.90	25.62	66.13	65.17
Deepseek-v3	14.02	19.08	28.83	9.02	24.00	23.47	66.81	64.69
Baichuan2-7B	13.57	19.25	28.87	7.91	24.00	22.11	66.58	64.84
Llama3-chinese-8B	14.90	21.36	30.81	10.28	26.69	21.97	63.64	65.11
Internlm3-8B	13.65	18.66	28.92	9.79	24.36	22.27	62.48	64.97
Yi-1.5-9B	14.85	19.21	32.52	10.26	24.98	25.80	67.02	65.15
GLM4-9B	11.59	15.43	26.11	5.93	20.47	22.76	64.64	63.83
Qwen2.5-7B	12.72	16.66	27.30	7.27	21.62	24.13	65.83	65.03
Baichuan2-7B-SFT	16.78	24.20	34.71	12.84	30.60	24.11	69.20	66.78
Llama-3-chinese-8B-SFT	16.10	23.02	33.42	12.48	29.55	23.43	65.06	66.24
Internlm3-8B-SFT	18.03	24.77	35.68	**14.56**	31.34	25.27	65.79	66.87
Yi-1.5-9B-SFT	18.46	25.39	**36.76**	14.13	32.14	26.56	70.24	**67.61**
GLM4-9B-SFT	17.86	25.07	36.23	13.70	31.72	25.59	70.03	67.31
Qwen2.5-7B-SFT (ASD-iLLM)	**18.68**	**25.87**	36.60	14.30	**32.69**	**26.75**	**70.47**	66.64

Table 3: Evaluation results of the automatic metrics on the test set. Models ending with SFT signify those models fine-tuned on the ASD-iLLM-8k training set. MET refers to METEOR metric. BS refers to BertScore metric. In the experiments, all models fine-tuned on the ASD-iLLM-8k dataset outperformed their base versions across all metrics and surpassed existing SOTA general LLMs on most metrics.

5.2 Experiment Detail

We selected 100 instances from the real multi-turn dialogue part of the ASD-iLLM-8k dataset as the test set, while the remaining data was used as the training set. This approach was aimed at evaluating the LLMs' intervention capabilities in real intervention scenarios after training. To ensure the richness of the test set, we randomly sampled 10 instances from each of the ten topics to compose the test set. All prompts and experimental settings in this paper are within the Chinese context.

We used the fine-tuning framework ms-swift (Zhao et al., 2025) for training LLMs on ASD-iLLM-8k dataset via LoRA method, utilizing 8 GTX 4090 GPUs. For hyperparameters, we set the epoch to 5, seed to 42, and learning rate to 1e-4, with LoRA rank at 8 and LoRA alpha at 32.

5.3 Evaluation Metrics

Automatic Evaluation We used common automated evaluation metrics to assess the differences between the predicted and reference sentences at the sentence level. In the Chinese context, we believe that style similarity is reflected in two aspects: the choice of words and the semantics of sentences. Firstly, in terms of word choice, different contexts require different words. For instance, casual social situations tend to be more colloquial, while communication with autistic children should be as concise and understandable as possible. Therefore, we used certain word overlap metrics such as BLEU (Papineni et al., 2002), GLEU (Wu et al., 2016), ROUGE (Lin, 2004), METEOR (Lavie and Agarwal, 2007) to assess the matching at the word level. Secondly, at the semantic and sentence level, we aim for the model's outputs to be semantically similar to real dialogues to achieve intervention effects similar to those of clinicians, so we chose BertScore (Zhang et al., 2020) and BGE-M3 embedding similarity (Chen et al., 2024) to measure the semantic similarity of the model's output.

Human Evaluation For the dialogue level, after discussions with the autism intervention doctors and inspired by the references (Yang et al., 2023b, 2024b; Zhang et al., 2024; Na, 2024), we conducted a comprehensive scoring evaluation of the doctors' parts in the multi-turn topic intervention dialogues generated by the role-play strategy. This evaluation was based on 3 aspects within 10 dimensions, as detailed in Table 2. Each dimension is rated on a scale from 0 to 4, with higher scores indicating better quality of the doctors' outputs. We invited three experienced clinical intervention doctors specializing in autism to score. More information for

Model	Professionalism				Linguistic			Safety		
	Prin.	Assi.	Rein.	Pers.	Rele.	Style	Fluency	Guid.	Priv.	Cont.
Doctor*	**3.55**	**3.49**	**3.15**	**3.44**	**3.71**	**3.77**	**3.68**	**3.77**	3.83	**4.00**
GPT4.1	1.62	1.57	1.76	1.07	1.93	0.62	0.81	2.70	**3.99**	**4.00**
GPT4o-mini	1.96	1.82	1.85	1.42	2.22	1.23	1.28	2.81	3.98	**4.00**
Qwen2.5-7b	1.71	1.60	1.62	1.17	1.82	0.73	0.87	2.75	3.96	**4.00**
ASD-iLLM	2.49	2.35	2.11	1.88	3.05	2.87	2.65	3.21	3.95	**4.00**

Table 4: Results of the human evaluation scoring in the dialogue level. Doctor* refers to the scoring for the doctors' performance of the test set, while the remaining scores pertain to the intervention dialogues generated using the role-play strategy. Higher scores indicate stronger capabilities in that category. The results indicate that, for the safe privacy aspects, the ASD-iLLM exhibits slight differences from the other models, while exceeding the performance of the other three models on the left majority of metrics. It is also the closest in performance to that of clinicians.

them are described in Appendix D.1.

6 Result And Analysis

6.1 Automatic Evaluation

The 100 multi-turn dialogues from the ASD-iLLM-8k test set are used to create the 1890 single-sentence prediction tasks for sentence-level evaluation. The evaluation results are shown in Table 3. We observed a significant improvement in various metrics for the 7b model fine-tuned on the ASD-iLLM-8k dataset. All fine-tuned models performed better than existing models across most metrics. This indicates that the model's conversational style and intervention strategy at the sentence level closely resemble authentic clinical dialogue interventions. For Qwen2.5-7b-Instruct, compared with its original version, the BLEU-4 metric increases by 5.96%, GLEU by 9.21%, Rouge-1 by 9.3%, Rouge-2 by 7.03%, Rouge-L by 11.07%, METEOR by 2.62%, BertScore by 4.64%, BGE by 1.61%.

Regarding the changes in various metrics, we believe that the differences in performance improvements between models stem from variations in their pre-training corpora, the differences in n-gram and vocabulary distributions lead to the observed variations in effectiveness after fine-tuning. For example, llama3-Chinese-8b is an improved version derived from llama3, but since llama3's pre-training corpus is primarily English, its performance improvement after fine-tuning on Chinese corpora is the least substantial, at only 1.42 points.

To validate the contribution of the designed system prompt that integrates ABA principles and the different parts of ASD-iLLM-8k dataset, we conduct more comparison experiments. Meanwhile, we take the Qwen-2.5 series as an example to conduct scale and hyperparameter ablation experiments. More details for these experiments are shown in Appendix E.

Overall, the fine-tuned Qwen2.5-7b model outperforms most metrics of existing SOTA LLMs. Therefore, we will adopt this model to represent the ASD-iLLM series for further evaluation.

6.2 Human Evaluation

For dialogue-level evaluation, we used the topics from the test set for the role-play strategy to generate multi-turn dialogues. Meanwhile, the multi-turn dialogues of the test set were scored as a baseline to assess their authority through comparison. The results are shown in Table 4, indicating that the various capabilities of ASD-iLLM closely align with those of clinical intervention doctors.

In terms of *Professionalism*, the four metrics closely resemble ABA principles, and ASD-iLLM shows significant improvement across all four indicators. This suggests that through fine-tuning on our dataset, ASD-iLLM has effectively learned the intervention strategies of doctors and can apply them appropriately in different turns. Regarding *Linguistics*, ASD-iLLM is the closest to clinicians, with the most significant improvement seen in *Style*, increasing by over 1.6 scores compared to the highest score among the other three models, which is gpt-4o-mini. The improvements in these metrics align with our fine-tuning objective: to enable the model to learn the strategies of clinical intervention dialogue and match its conversational style, demonstrating the feasibility and effectiveness of our proposed framework.

6.3 Case Study

Figure 7 illustrates a case of dialogue intervention about the season conducted by ASD-iLLM. We can

Figure 7: Case study for ASD-iLLM. Doctor's parts are outputs of ASD-iLLM. Blue indicates instructions from ABA, green denotes assistance, yellow signifies reinforcement, and red represents the child's responses.

observe that ASD-iLLM's dialogue style closely resembles clinical interventions, with expressions that are concise and easy for children to understand. Moreover, it follows ABA principles by providing appropriate instruction repetition, assistance, and reinforcement to different types of child responses. This fully demonstrates ASD-iLLM's potential to effectively replace intervention doctors in future dialogue interventions.

7 Conclusion

In this study, we developed a comprehensive framework for constructing LLMs in dialogue intervention for autistic children, encompassing data collection, model training, and evaluation. We developed ASD-iLLM-8k, the first available Chinese dataset for dialogue intervention in autism, to advance future research. Our model, ASD-iLLM, is compact yet powerful, addressing the shortcomings of existing models whose dialogue styles do not closely align with clinical interventions, and intervention strategies do not adhere to ABA principles. Extensive experiments demonstrate that our proposed framework is effective, the dataset is of high quality, and the model exhibits excellent capabilities in performing various aspects of dialogue intervention tasks for autistic children.

Limitations

The limitations of our work are twofold. First, we did not perform automatic or human evaluations across a broader range of topics. Evaluations were conducted solely on the topics derived from the ASD-iLLM-8k and discussions with intervention doctors, which may pose risks regarding the model's dialogue capabilities beyond these topics. Secondly, ASD-iLLM has not been tested in real clinical dialogue interventions for autistic children to assess its practical potential and intervention effectiveness.

Ethics Statement

Data Privacy Throughout the entire process of dataset construction, we implemented strict privacy protection measures. We perform both automatic and manual cleaning to replace or remove any potential privacy or sensitive information, such as names and addresses, from the raw data, ensuring that the dataset contains no sensitive or privacy-related content. We release the dataset publicly for further research purposes only.

Considerations of using ASD-iLLM Despite extensive experiments demonstrating the model's superiority in dialogue intervention tasks for autistic children, the current ASD-iLLM cannot be directly applied to formal clinical interventions due to the absence of preliminary clinical trials. More consideration and experiments regarding ethics, safety, and efficacy are required.

Acknowledgements

We thank all volunteers for their participation in the study. This work was supported in part by STI 2030—Major Projects under Grant 2021ZD0200400, in part by the National Natural Science Foundation of China under Grant 62336007, in part by the Key Research and Development Program of Zhejiang under Grant 2023C03003, in part by the Key R&D Program of Zhejiang (2024SSYS0016), in part by the Starry Night Science Fund of the Zhejiang University Shanghai Institute for Advanced Study under Grant SN-ZJU-SIAS-002, in part by the Fundamental Research Funds for the Central Universities, in part by ZJU-GENSCI CHILDREN'S HEALTH RESEARCH & DEVELOPMENT CENTER (ZJU-GENSCI2024YB003), in part by the Project for Hangzhou Medical Disciplines of Excellence, and in part by the Key Project for Hangzhou Medical Disciplines.

References

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

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Qian Chen, Mengzhe Chen, Bo Li, and Wen Wang. 2020. Controllable time-delay transformer for real-time punctuation prediction and disfluency detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8069–8073. IEEE.

Yujin Cho, Mingeon Kim, Seojin Kim, Oyun Kwon, Ryan Donghan Kwon, Yoonha Lee, and Dohyun Lim. 2023. Evaluating the efficacy of interactive language therapy based on llm for high-functioning autistic adolescent psychological counseling. *arXiv preprint arXiv:2311.09243*.

Lei Chu, Hongyan Wu, and Yi Pan. 2024. Chatasd: A dialogue framework for llms enhanced by autism knowledge graph retrieval. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–8.

Madalina G Ciobanu, Cesare Tucci, and Fausto Fasano. 2024. Llms for autism treatment: Current trends and emerging strategies. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 6797–6804. IEEE.

John O Cooper, Timothy E Heron, William L Heward, et al. 2007. Applied behavior analysis.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Vera Dekker, Maaike H Nauta, Marieke E Timmerman, Erik J Mulder, Lianne van der Veen-Mulders, Barbara J van den Hoofdakker, Sjoukje van Warners, Leonieke JJ Vet, Pieter J Hoekstra, and Annelies de Bildt. 2019. Social skills group training in children with autism spectrum disorder: a randomized controlled trial. *European child & adolescent psychiatry*, 28:415–424.

Chengyun Deng, Shuzhong Lai, Chi Zhou, Mengyi Bao, Jingwen Yan, Haifeng Li, Lin Yao, and Yueming Wang. 2024. Asd-chat: An innovative dialogue intervention system for children with autism based on llm and vb-mapp. *arXiv preprint arXiv:2409.01867*.

Annette Estes, Jeffrey Munson, Sally J Rogers, Jessica Greenson, Jamie Winter, and Geraldine Dawson. 2015. Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 54(7):580–587.

Richard M Foxx. 2008. Applied behavior analysis treatment of autism: The state of the art. *Child and adolescent psychiatric clinics of North America*, 17(4):821–834.

Elizabeth A Fuller and Ann P Kaiser. 2020. The effects of early intervention on social communication outcomes for children with autism spectrum disorder: A meta-analysis. *Journal of autism and developmental disorders*, 50(5):1683–1700.

Zhifu Gao, ShiLiang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Interspeech 2022*, pages 2063–2067.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

R Hanrahan, E Smith, H Johnson, A Constantin, and M Brosnan. 2020. A pilot randomised control trial of digitally-mediated social stories for children on the autism spectrum. *Journal of autism and developmental disorders*, 50:4243–4257.

Rukhshan Haroon and Fahad Dogar. 2024. Twips: A large language model powered texting application to simplify conversational nuances for autistic users. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–18.

Wenjie He, Wenyan Zhang, Ya Jin, Qiang Zhou, Huadan Zhang, and Qing Xia. 2024. Physician versus large language model chatbot responses to web-based questions from autistic patients in chinese: cross-sectional comparative analysis. *Journal of Medical Internet Research*, 26:e54706.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yue Huang, Siyuan Wu, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Chaowei Xiao, Jianfeng Gao, Lichao Sun, et al. 2024. Datagen: Unified synthetic dataset generation via large language models. In *The Thirteenth International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Rabia Jafri. 2024. A social communication support application for autistic children using computer vision and large language models. In *International Conference on Computers Helping People with Special Needs*, pages 217–223. Springer.

JiWoong Jang, Sanika Moharana, Patrick Carrington, and Andrew Begel. 2024. "it's the only thing i can trust": Envisioning large language model use by autistic workers for communication assistance. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Unggi Lee, Hansung Kim, Juhong Eom, Hyeonseo Jeong, Seungyeon Lee, Gyuri Byun, Yunseo Lee, Minji Kang, Gospel Kim, Jihoi Na, et al. 2025. Echoteddy: Preliminary design and development of large language model-based social robot for autistic students. *arXiv preprint arXiv:2502.04029*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Ziming Li, Pinaki Prasanna Babar, Mike Barry, and Roshan L Peiris. 2024. Exploring the use of large language model-driven chatbots in virtual reality to train autistic individuals in job communication skills. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Ziming Li, Pinaki Prasanna Babar, and R Peiris. 2025. Generative role-play communication training in virtual reality for autistic individuals: A study on job coach experiences in vocational training programs. In *CHI'25*. Association for Computing Machinery.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yingying Lin, Guozhi Chen, Huaxiang Lu, Rongfei Qin, Jinsheng Jiang, Weiwei Tan, Caibin Luo, Ming Chen, Qin Huang, Liangliang Huang, et al. 2025. Inequality and heterogeneity in medical resources for children with autism spectrum disorders: a study in the ethnic minority region of southern china. *BMC Public Health*, 25(1):1–17.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Bennett M Liu, Kelley Paskov, Jack Kent, Maya McNealis, Soren Sutaria, Olivia Dods, Christopher Harjadi, Nate Stockham, Andrey Ostrovsky, and Dennis P Wall. 2023. Racial and ethnic disparities in geographic access to autism resources across the us. *JAMA Network Open*, 6(1):e2251182–e2251182.

Ruchik Mishra and Karla Conn Welch. 2024. Towards scalable robotic intervention of children with autism spectrum disorder using llms. *arXiv e-prints*, pages arXiv–2402.

Ruchik Mishra, Karla Conn Welch, and Dan O Popa. 2024. Human-mediated large language models for robotic intervention in children with autism spectrum disorders. *arXiv preprint arXiv:2402.00260*.

Hongbin Na. 2024. Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2930–2940.

Krishna Pai, Vidhita Jagwani, Shivalik Pandita, and Dhananjay Kalbande. 2024. Multimodal integration, fine tuning of large language model for autism support. In *2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, pages 630–634. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Tianyu Ren, Hui Wang, and Karen Rafferty. 2025. Enhancing question generation through diversity-seeking reinforcement learning with bilevel policy decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25083–25091.

Xiaoyu Ren, Yuanchen Bai, Huiyu Duan, Lei Fan, Erkang Fei, Geer Wu, Pradeep Ray, Menghan Hu, Chenyuan Yan, and Guangtao Zhai. 2023. Chatasd: Llm-based ai therapist for asd. In *International Forum on Digital TV and Wireless Multimedia Communications*, pages 312–324. Springer.

Henry S Roane, Wayne W Fisher, and James E Carr. 2016. Applied behavior analysis as treatment for autism spectrum disorder. *The Journal of pediatrics*, 175:27–32.

Mark L Sundberg. 2008. *VB-MAPP Verbal Behavior Milestones Assessment and Placement Program: a language and social skills assessment program for children with autism or other developmental disabilities: guide*. Mark Sundberg.

Yilin Tang, Liuqing Chen, Ziyu Chen, Wenkai Chen, Yu Cai, Yao Du, Fan Yang, and Lingyun Sun. 2024. Emoeden: Applying generative artificial intelligence to emotional learning for children with high-function autism. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Femke van der Wilt, Renske Bouwer, and Chiel van der Veen. 2022. Dialogic classroom talk in early childhood education: The effect on language skills and social competence. *Learning and Instruction*, 77:101522.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023a. Cam++: A fast and efficient network for speaker verification using context-aware masking. In *Interspeech 2023*, pages 5301–5305.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508.

Yongfu Wang and Tiffany Y Tang. 2024. Position paper: A personalized large language model (llm)-based chat companion for autistic children early intervention. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 697–700.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023b. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19368–19376.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Jinan Zeidan, Eric Fombonne, Julie Scorah, Alaa Ibrahim, Maureen S Durkin, Shekhar Saxena, Afiqah Yusuf, Andy Shih, and Mayada Elsabbagh. 2022. Global prevalence of autism: A systematic review update. *Autism research*, 15(5):778–790.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13947–13966.

Shiliang Zhang, Ming Lei, Zhijie Yan, and Lirong Dai. 2018. Deep-fsmn for large vocabulary continuous speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5869–5873. IEEE.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29733–29735.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Autism Support via LLMs

With the substantial increase in the scale and capabilities of LLMs, the extensive knowledge they encompass can be utilized to support individuals with autism and their caregivers. For individuals with autism, (Jafri, 2024) proposed an assistive tool that can provide advice when autistic children have difficulty socializing and understanding non-verbal social signals such as facial expressions. The experiment utilized Zoom virtual meeting room for real-life dialogues, offering lower costs and enhanced sensory experiences. TwIPS (Haroon and Dogar, 2024) utilizes prompt engineering to enable GPT-4 (Achiam et al., 2023) to assist users with autism in comprehending the tone and meaning of conversations through three steps: interpret, preview, and suggest. It also aids users in understanding the emotional conveyance of their messages and refining that may lead to misunderstandings.

For autism carers, ChatASD (Chu et al., 2024) collects autism-related knowledge to construct a knowledge graph and employs Graph Retrieval Augmentation Generation (RAG) technology to enhance the professional question-answering capabilities of LLMs, thereby providing diagnostic or intervention assistance to parents. Similarly, (Ren et al., 2023) developed a bilingual autism knowledge base with 4,500 entries to fine-tune LLMs, enhancing the professionalism of their responses. (Wang and Tang, 2024) has developed an LLM-based chat companion to educate autism carers on how to understand, interact, and communicate with autistic children.

B More Details for ASD-iLLM-8k

B.1 Data Cleaning

To obtain higher quality real data, we followed the doctors' recommendations and implemented the following data cleaning steps:

- We removed multi-turn dialogue texts with fewer than five exchanges. Dialogues with too few exchanges fail to reflect the doctor's intervention strategies adequately.
- Dialogues focused on entities, such as storybooks or toys, were removed. The model requires visual comprehension to understand the images or entities referenced in these multi-turn dialogues. Currently, our focus is on enhancing the model's dialogue style and intervention strategies; therefore, this portion of the dialogue is not suitable for the present training.
- For any potential privacy or sensitive information in the dialogues, specifically names and addresses, we will implement safe substitutions. Names will be uniformly replaced with "child," and addresses will be limited to the city only.

B.2 Topics Description

Table 5 encapsulates the information regarding the topics for the ASD-iLLM-8k dataset, including 10 major topics and 27 subtopics. It encompasses ten common topics of dialogue intervention, ranging from daily self-care to cognitive understanding, providing comprehensive coverage of authentic clinical intervention scenarios.

B.3 The Demographic Details of Children in ASD-iLLM-8K

The demographic information of children in ASD-iLLM-8k dataset is presented in Table 6, indicating 62 boys and 12 girls. There is minimal difference in means and variances between genders regarding

Topic	Sub-Topic	Explanation
Self-care	1. How to wash hands 2. How to dress 3. Identify male and female toilets 4. How to brush teeth and wash face 5. How to bathe 6. Choose transportation	This intervention scenario aims to cultivate the ability of children to independently complete daily activities, such as learning how to dress, wash, choose transportation, and other basic skills.
Animals and Nature	7. Animals 8. Weather 9. Season recognition	This intervention scenario aims to enhance children's cognition and understanding of the natural world through activities related to animals, weather, and season changes.
Food and Diet	10. Food 11. Fruit	This intervention scenario aims to help children understand the types and sources of food, as well as eating habits and rules.
Social Interaction	12. Role-playing cashier and customer 13. Role-playing restaurant waiter and customer 14. Role-playing doctor and patient 15. Learning social etiquette 16. Share daily life 17. Greeting	This intervention scenario aims to help children master basic social skills, such as social initiation, social maintenance, and ending conversations, and improve their social confidence and interaction ability.
Cognition and Understanding	18. Introduce yourself 19. Understand sequence and timeline 20. Understand self-concept 21. Occupation	This intervention scenario aims to help children understand and master basic cognitive concepts related to self, time, occupation, and gender.
Stories and Fables	22. Retell fables and understand the content 23. Story retelling	This intervention scenario aims to help children improve their language expression, comprehension, and social interaction skills.
Festivals and Culture	24. Understand festivals and customs	This intervention scenario aims to help children understand and integrate into festival celebrations in different cultures.
Rules and Common Sense	25. Learn the behavior norms in public places 26. Learn traffic safety common sense	This intervention scenario aims to help children understand and master the social norms and safety common sense that they need to follow in daily life.
Color	27. Color	This intervention scenario is designed to help children identify, distinguish, and understand the concept of different colors.
Others	-	Other commonly used intervention topics.

Table 5: Brief descriptions of the 10 main topics and 27 subtopics of child intervention conversations.

Gender	Number	Age (Mean ± std)	Language Development Age (Mean ± std)
Male	62	5.34 ± 1.09	3.90 ± 1.09
Female	12	5.08 ± 1.40	3.20 ± 0.80

Table 6: The demographic details of children for ASD-iLLM-8K.

System Prompt

Role

1. You are an expert focused on dialogue interventions for children with autism.
2. Based on the provided reference dialogue, please generate a multi-turn dialogue that mirrors the style and speaking manner, with the dialogue topic being ***{new_topic}***.
3. Include **[Child Unresponsive]** at appropriate places to simulate the child's lack of response, ensuring the conversation remains natural and fluid, presented line by line starting with either the doctor or the child.

Reference Dialogue

{ref_dialogue}

Generated Dialogue

Figure 8: System prompt used for generating multi-turn dialogue. Generate a new multi-turn dialogue on the given *{new_topic}*, referencing the style of the provided *{ref_dialogue}*.

age around five years old. However, the language development age significantly lags behind the actual age, at approximately three to four years old, consistent with characteristics of autistic children.

B.4 System prompt for Data Synthesis

Figure 8 illustrates the system prompt used for data synthesis with GPT-4.1. It is noteworthy that in the prompt, we specified that the child's unresponsive state should be presented as [Child Unresponsive], maintaining consistency with the real data.

B.5 Assessing the quality of Synthesis Data

This section comprehensively evaluates the quality of the synthetic data. High-quality synthesized data should ensure that its dialogue style closely resembles real data while also maintaining the richness of its dialogue content.

Table 7 displays the statistics for both real and synthetic data. It can be observed that the synthetic data closely approximates the real data in terms of both the average number of turns per dialogue and the average sentence length. In terms of the Distinct metric, the powerful expressive capability of GPT-4.1 leads to greater diversity in the doctors' speech within the synthesized data, with Distinct-2 at 77.29 and Distinct-3 at 91.39, surpassing the real data by 4.66 and 3.27, respectively. However, the diversity of children's responses is lower than that of the real data, indicating that there is greater uncertainty in children's utterances in real-world scenarios.

To measure the similarity between real and synthetic data from a distributional perspective, we calculated metrics including the number of dialogue turns, word count, Distinct-n, Self-BLEU, self-GLEU, and self-BertScore for each multi-turn dialogue instance. We select self-BLEU, self-GLEU, and self-BertScore, which focus on the similarity between multiple outputs generated by the same model, thereby assessing the diversity of the dataset (Zhu et al., 2018; Huang et al., 2024; Ren et al., 2025). All indicators are normalized to a uniform dimension for better visualization. Figure 9 shows the distribution differences of various metrics between synthesized data and real data. The result shows that the distribution of real data and synthesized data is very similar, demonstrating the high quality of the synthesized data.

To evaluate the quality of synthetic data at the semantic embedding level, we use OpenAI's `text-embedding-3-large` to obtain text embedding. Then, use the t-SNE method to reduce and map the data into a two-dimensional space. Figure 10 illustrates the semantic embedding distribution of real (Original) and synthetic (Generated) data. The embeddings are mainly distributed in 27 clusters, which correspond exactly to the 27 subtopics in real intervention dialogues. Moreover, the distribution of the synthetic data is concentrated in the distribution of the real data indicated that the synthetic data items are semantically consistent with the real data, thus confirming their semantic similarity.

C Role-Play Strategy

Figure 11 presents the system prompt provided to GPT-4o for simulating responses from autistic children when the intent is either relevant or irrelevant. Figure 12 presents the relevant pseudocode for generating responses from autistic children using GPT-4o, based on the given intention.

D Human Evaluation

The annotators focused on the usage of each scoring item during each teaching trial . A trial refers to a complete cycle in DTT, as illustrated in Figure 4. They needed to break down multi-turn dialogues into multiple trials to assess the application of ABA principles, linguistic , and safety in each trial. Based on the overall assessment, they assigned scores between 0 and 4 as follows: **0:** None of the doctor's dialogues trial are appropriate. **1:**

Category	Doctor from real	Child from real	Doctor from syn	Child from syn
Turns per dialogue	19.69	16.82	13.33	9.92
Characters per sentence	17.37	6.41	19.03	4.27
Distinct-2	$72.63_{\pm 10.11}$	$75.98_{\pm 19.57}$	$77.29_{\pm 8.37}$	$68.95_{\pm 17.91}$
Distinct-3	$88.12_{\pm 7.69}$	$79.67_{\pm 23.75}$	$91.39_{\pm 7.46}$	$66.16_{\pm 22.84}$

Table 7: Statistical comparison between authentic intervention dialogues and synthetic intervention dialogues.

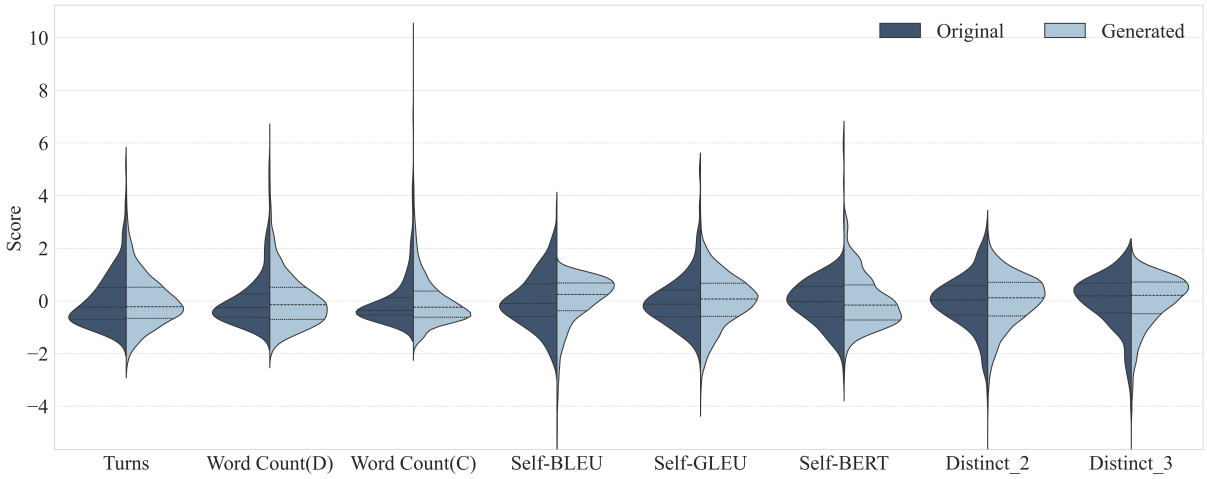

Figure 9: Statistical distribution results of real (Original) and synthetic (Generated) data. **D** stands for Doctor, while **C** stands for Child.

A small portion of the doctor's dialogues trial is appropriate. **2:** Some of the doctor's dialogues trial are appropriate. **3:** Most of the doctor's dialogues trial are appropriate. **4:** All of the doctor's dialogues trial are appropriate.

D.1 Information of Experts for Human Evaluation

Table 8 presents detailed information about three invited experts for human evaluation, each with more than four years of experience in autism treatment. Their extensive intervention experience and knowledge make them well-qualified for the professional evaluation task. Each expert will receive 100 yuan per hour as a labor fee based on the working hours, which is higher than the general salary.

E Ablation Experiment

E.1 ASD-iLLM-8k Dataset

We proposed a dataset called ASD-iLLM-8k, which includes both real and synthetic intervention dialogues for autistic children. After training, we observed great improvements in the model's performance. This naturally raises the question: **"How do the different parts of the dataset contribute to this enhancement?"**

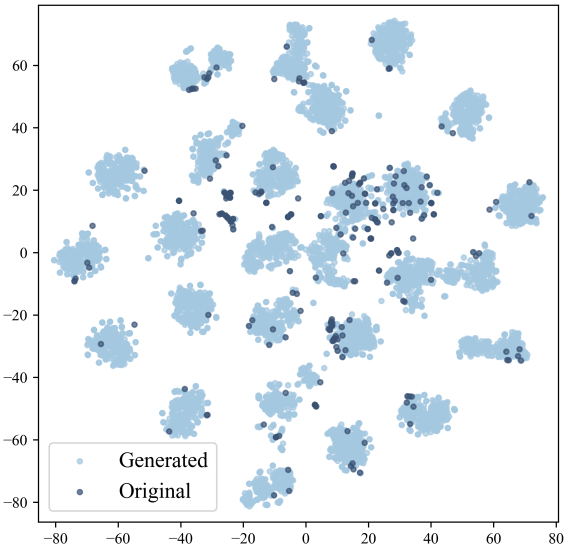

Figure 10: Distribution of semantic embedding of real (Original) and synthetic (Generated) data.

Info	Gender	Work Exp.	Job Responsibilities
Doctor1	Female	5 years	Early Intervention for Autism Child.
Doctor2	Female	4 years	Language and Articulation Disorder Therapy.
Doctor3	Female	6 years	Social Behavior Intervention for Autism.

Table 8: Information for experts involved in human evaluation.

System Prompt

Role
You will play the role of an autistic child and have a topic conversation with an intervention doctor. You need to give a response that is *related (or not related)* to the given dialogue history. The response needs to be consistent with the identity of an autistic child and be as brief as possible.

Dialogue History
{dialogue_history}
Your Response

Figure 11: System prompt for generating related or unrelated responses from children via LLMs.

Algorithm 2: Child Response Generation via LLM

Input : Child's intent I, dialogue history X_i
Output: Generated child's response R

f_{child}: LLM designed to role-play as a child with autism
S_{CR}: System prompt for generating a correct response
S_{IR}: System prompt for generating an incorrect response
Function `GenerateChildResponse`(I, X_i, f_{child})**:**
  if I *is Unresponsive* **then**
    return "[Child Unresponsive]";
  end
  if I *is Repeat* **then**
    words ← *jieba.cut*(X_i);
    words ← *words.remove*(*stopwords*)
    return words[-1];
  end
  if I *is CorrectResponse* **then**
    return $f_{child}(X_i, S_{CR})$
  end
  if I *is IncorrectResponse* **then**
    return $f_{child}(X_i, S_{IR})$
  end

Figure 12: Pseudocode for children's response generation via Role-Play strategy.

Therefore, we conducted the following ablation experiments on the training set of ASD-iLLM-8K: training with only 192 real data samples, training with only 7843 synthetic samples, and creating a subset of 189 samples from the synthetic data by sampling 7 samples per topic, similar in quantity to the whole real data parts for training. The training and evaluation settings were consistent with the automated evaluation experiments at the sentence level, and the results are presented in the Table 9.

The analysis of the ablation experiment results yields the following conclusions:

- **The combination of clinical real data and synthetic data leads to more improved performance.** We achieved optimal and suboptimal results using the full training set (Full) and subset (Mix), which reinforces our motivation for using synthetic data: to enrich topics and dialogue scenarios and enhance the model's generalization capability.

- **Under conditions of comparable dataset sizes, the performance of real data (w/o Syn) surpasses that of synthetic data (Only Syn).** On one hand, this indicated the value of high-quality real clinical data. Even a few hundred samples can enable the model to learn to apply the ABA principles for interventions. It can also be used to guide for data synthesis, further enhancing the model's generalization capability. On the other hand, it indicates that synthetic data still contain noise, which may stem from the structural limitations of the model's internal knowledge.

- **When training without real data (w/o Real, Only Syn), the model's performance can still be improved, highlighting the importance of high-quality synthetic data.** Interestingly, when training with a small mixed dataset of real and synthetic data (Mix), the performance surpasses that achieved using nearly 8000 synthetic data alone (w/o Real), further reflecting the value of real data.

Settings (Qwen2.5-7b)	BLEU	GLEU	R-1	R-2	R-L	MET.	BS.
Base	12.72	16.66	27.30	7.27	21.62	24.13	65.83
Full (8035)	**18.68**	**25.87**	**36.60**	**14.30**	**32.69**	**26.57**	**70.47**
w/o Syn (192)	16.72	23.69	35.07	12.85	30.72	24.30	69.32
w/o Real (7843)	17.12	24.14	34.16	11.78	29.93	24.69	69.37
Only Syn (189)	15.99	22.86	32.88	10.33	28.44	23.65	68.95
Mix (192+189)	17.20	24.42	35.60	13.12	31.27	25.48	69.93

Table 9: Dataset ablation experiment results based on Qwen2.5-7b-instruct. The numbers in parentheses indicate the size of the training data for that specific setting. The abbreviation "w/o" stands for "without." "Mix" refers to training using a subset of data comprising 192 real samples and 189 synthetic samples combined.

System Prompt

Role

1. You are an experienced autism intervention doctor with extensive knowledge in autism intervention.
2. Speak in a warm, kind tone, expressing empathy and affirming the child's responses with praise.
3. Please engage in dialogue with the child naturally, using simple words that is easy for them to understand, ensuring that your responses are fluid and align with the identity of an intervention specialist.
4. Now, you will have a thematic conversation with an autistic child on the topic: *{topic_content}*. Please begin the dialogue.

Figure 13: General system prompt for fine-tuning and evaluation.

E.2 System Prompt for training

To further validate the importance of ABA principles in topic interventions for autistic children, we constructed a generic system prompt shown in Figure 13 that excludes concrete ABA principles and conducted comparative experiments on the test set. The results shown in Table 10 demonstrate that, across most metrics and models, the topic dialogue intervention style and strategies utilizing system prompts with ABA principles are more aligned with authentic clinical dialogues. This indicates that explicit instructional constraints based on ABA principles can further enhance the capabilities of LLMs.

However, merely constraining through prompts does not alter the dialogue style of LLMs at the parameter level nor enable them to comprehend the underlying ABA principles. Therefore, we need to train using the ASD-iLLM-8k dataset. The result demonstrates that, with the combination of system prompts with ABA principles and fine-tuning through the ASD-iLLM-8k dataset, the model's capabilities are significantly enhanced, surpassing SOTA models.

E.3 LoRA Rank

The experiment investigated the impact of different LoRA ranks on model performance, with results presented in Table 11. The experiment selected Qwen-2.5-7B-Instruct as the base model. The remaining experimental settings and hyperparameters are consistent with those used in automatic evaluation. It was observed that both smaller and larger ranks can decrease model performance to a certain extent. With smaller ranks, the number of updated parameters is insufficient to fully represent the features of downstream tasks, while larger ranks may introduce excessive parameters that capture noise present in the data, thus reducing model performance.

E.4 Model Scale

The experiment investigated the effect of model size on the effectiveness of autism dialogue intervention tasks. We selected the 7B, 14B, and 32B models from the Qwen-2.5 series for ablation experiments, using the ASD-iLLM-8k for fine-tuning. The system prompt was designed in accordance with ABA principles. Table 12 compares the results of automated evaluation metrics for models with various parameter sizes before and after fine-tuning. The results indicate that fine-tuning with the ASD-iLLM-8k dataset significantly improves the performance of models with varying parameter sizes, but increasing the parameter size does not lead to further performance enhancements. The 7B model even outperforms the 32B model on some metrics.

One possible explanation is that topic dialogue intervention with autistic children tasks primarily involve basic daily and social knowledge, which is already incorporated during the pre-training of models with varying parameter sizes. The goal of downstream fine-tuning is merely to adjust dia-

Model_Name	ABA	SFT	BLEU	GLEU	R-1	R-2	R-L	MET.	BS.
GPT-4.1	✗	✗	11.07	14.14	25.17	6.05	19.56	24.03	65.06
	✓	✗	**11.35**	**14.64**	**25.48**	**6.49**	**19.96**	**24.33**	**65.14**
GPT-4o-mini	✗	✗	12.02	16.40	27.07	6.00	21.65	22.82	65.76
	✓	✗	**13.62**	**17.89**	**28.74**	**7.38**	**22.95**	**24.75**	**66.68**
Gemini-2.0-flash	✗	✗	10.89	13.85	25.89	6.06	19.35	24.53	65.58
	✓	✗	**11.91**	**15.14**	**27.30**	**7.41**	**20.90**	**25.62**	**66.13**
GPT-4.1-mini	✗	✓	11.01	14.79	25.69	5.51	20.01	22.69	64.87
	✓	✗	**13.24**	**17.09**	**28.02**	**7.16**	**22.28**	**25.04**	**66.24**
Baichuan2-7B	✗	✗	13.06	18.55	28.31	6.66	23.41	22.22	66.09
	✓	✗	13.57	19.25	28.87	7.91	24.00	22.11	66.58
	✓	✓	**16.78**	**24.00**	**34.71**	**12.84**	**30.60**	**24.11**	**69.20**
Llama-3-chinese-8B	✗	✗	14.91	21.10	31.36	10.21	26.73	22.38	**67.49**
	✓	✗	14.90	21.36	30.81	10.28	26.69	21.97	63.64
	✓	✓	**16.10**	**23.02**	**33.42**	**12.48**	**29.55**	**23.43**	65.06
Internlm3-8B	✗	✗	13.25	18.00	28.30	9.08	23.64	22.25	64.95
	✓	✗	13.65	18.66	28.92	9.73	24.36	22.27	62.48
	✓	✓	**18.03**	**24.77**	**35.68**	**14.56**	**31.34**	**25.27**	**65.79**
Yi-1.5-9B	✗	✗	14.30	18.61	31.94	9.86	24.26	25.06	66.75
	✓	✗	14.85	19.21	32.52	10.26	24.98	25.80	67.02
	✓	✓	**18.46**	**25.39**	**36.76**	**14.13**	**32.14**	**26.56**	**70.24**
GLM4-9B	✗	✗	11.32	15.21	25.62	5.74	20.32	22.40	64.49
	✓	✗	11.59	15.43	26011	5.93	20.47	22.76	64.64
	✓	✓	**17.86**	**25.07**	**36.23**	**13.70**	**31.72**	**25.59**	**70.03**
Qwen2.5-7B	✗	✗	11.86	15.51	25.93	6.63	20.50	23.26	65.08
	✓	✗	12.72	16.66	27.30	7.27	21.62	24.13	65.83
	✓	✓	**18.68**	**25.87**	**36.60**	**14.30**	**32.69**	**26.75**	**70.47**

Table 10: Comparison of the effects of the two types of prompts in automated evaluation. **ABA** refers to using system prompts with ABA principles, while **SFT** signifies fine-tuning using the ASD-iLLM-8k dataset. The generic system prompt for comparison is illustrated in Figure 13.

LoRA rank	BLEU	GLEU	R-1	R-2	R-L	MET.	BS.
4	17.84	25.11	36.32	13.64	31.93	25.90	70.21
16	17.07	24.17	34.33	13.02	30.48	24.55	69.16
32	17.04	24.28	35.00	13.41	31.11	24.99	69.55
8	**18.68**	**25.87**	**36.60**	**14.30**	**32.69**	**26.75**	**70.47**

Table 11: The impact of LoRA rank parameter selection on fine-tuning. The experiment was conducted using Qwen2.5-7B-Instruct as the base model, fine-tuned on the ASD-iLLM-8k dataset, while keeping the same other parameters with automatic evaluation.

System Prompt

Role setting
You are an experienced doctor who intervenes children with autism. Based on your rich experience in autism intervention and the given context, please choose the next response you think is better between the two doctors' responses. Returning 0 means that the first doctor's response is better, returning 1 means that the second doctor's response is better, and returning 2 means that the two types of responses cannot be distinguished. Please return the answer you gave directly.

Dialogue history
Dialogue topic: *{topic}*
{dialogue_history}

Doctor 1's response
{response1}

Doctor 2's response
{response2}

Your answer

Figure 14: System prompt used to guide LLM for selecting the better next doctor response in a given contextual scenario.

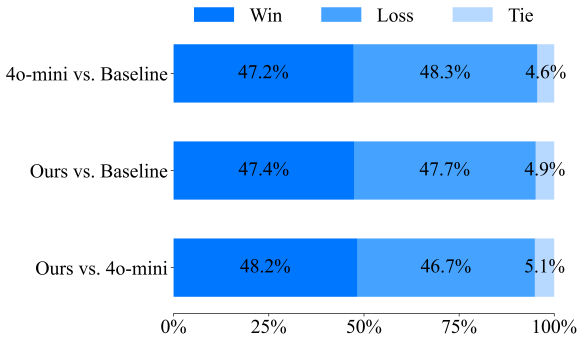

Figure 15: Results of pairwise comparisons using LLM. 4o-mini refers to GPT-4o-mini. Baseline refers to real doctors' responses. Ours refers to ASD-iLLM's outputs. We present the win, loss and tie rates of each compared pair in 1890 randomly sampled sessions of the test set.

logue style and intervention strategies, without engaging more complex reasoning and professional abilities. Therefore, larger parameters do not result in significant performance enhancements. In other words, models with different parameter sizes already possess the necessary knowledge for intervention. Fine-tuning on the ASD-iLLM-8k dataset is intended to equip the LLM with the ability to apply knowledge like ABA principles as a professional autism intervention doctor.

F LLM Evaluation

In LLM evaluation, we use GPT-4o, Deepseek-R1 and Claude4-opus to conduct sentence-level and dialogue-level evaluation. For sentence-level evaluation, we primarily measure the ability of LLMs to predict the doctor's next output given the historical dialogue. Simply, we provide the system prompt along with the real dialogue history between doctor and child from rounds 1 to $t-1$, then ask the LLM to predict the doctor's t^{th} response and compare it to the real response. Subsequently, we employed a pairwise (Qiu et al., 2024) comparison method, allowing GPT-4o to select the better response between two predictions. The system prompt used by GPT-4o for pairwise comparisons is shown in Figure 14.

We selected real doctors' responses, GPT-4o-mini, and ASD-iLLM outputs for pairwise comparison on the test set. The results are shown in Figure 15. The results indicate that GPT-4o cannot distinguish the differences in quality among the three output types in pairwise comparisons, suggesting that its capabilities in evaluating intervention dialogues for autistic children still need improvement.

At the dialogue level, we follow the same scoring criteria as human evaluation and integrate them into the system prompt, as shown in Figure 17. The average scoring result of Deepseek-R1 and Claude4-opus are shown in Table 13. Surprisingly, the scores for real clinical dialogues, which serve as the ground truth, are lower than those for existing LLMs. This is markedly different from the human evaluation, which is scored by three experienced experts, indicating a potential bias in LLM when evaluating intervention dialogues for autistic children. Specifically, LLMs tends to favor comprehensive and long dialogue content. Thus, GPT-4.1 scores lower in human evaluation, but achieves the highest score in LLM evaluation. It can also be observed that ASD-iLLM scores the lowest in LLM evaluation, but its scores are closer to those of real dialogues. From the perspective of relative score differences, it is consistent with human evaluation results.

In summary, the aforementioned experimental setup fails to yield conclusions that are consistent with human evaluations, indicating that the current general LLMs seem unsuitable for direct scoring. Also, further research and experimental designs are required to validate this.

G More Case Study

Figure 16 illustrates the topic intervention dialogue content between ASD-iLLM and the real clinician. We can see that ASD-iLLM, even when faced with

Model_Name	BLEU	GLEU	R-1	R-2	R-L	MET.	BS.
Qwen2.5-7b	12.72	16.66	27.30	7.27	21.62	24.13	65.83
Qwen2.5-14b	9.16	11.82	22.64	5.29	16.82	22.10	63.30
Qwen2.5-32b	9.12	11.49	23.08	5.48	16.79	22.92	63.68
Qwen2.5-7b-SFT	**18.68**	**25.87**	36.60	**14.30**	32.69	**26.75**	70.47
Qwen2.5-14b-SFT	18.03	25.02	36.14	13.98	32.03	26.35	70.35
Qwen2.5-32b-SFT	18.66	**25.87**	**37.14**	14.25	**32.78**	26.41	**70.62**

Table 12: Ablation experiment result on the Qwen2.5 series models of different sizes. Ending with SFT indicates that the model is fine-tuned using the ASD-iLLM-8k dataset.

Model(Avg)	Professionalism				Linguistic			Safety		
	Prin.	Assi.	Rein.	Pers.	Rele.	Style	Fluency	Guid.	Priv.	Cont.
Doctor*	**3.03**	2.97	2.95	2.59	**3.75**	3.53	3.39	3.81	3.81	**4.00**
GPT-4.1	**3.03**	**3.33**	**4.00**	**3.24**	2.95	**3.74**	3.45	**3.99**	**4.00**	**4.00**
GPT-4o-mini	2.39	2.78	3.43	2.77	2.18	3.23	3.05	3.71	**4.00**	**4.00**
Qwen2.5-7b	2.44	2.89	3.43	2.80	2.45	3.12	2.76	3.66	**4.00**	**4.00**
ASD-iLLM	2.73	2.63	2.84	2.50	3.11	3.51	**3.51**	3.77	**4.00**	**4.00**

Table 13: Average score of Deepseek-R1 and Claude-4-opus for LLM evaluation in dialogue level. Doctor* refers to the scoring for the doctors' performance of the test set, while the remaining scores pertain to the intervention dialogues generated using the role-play strategy. Higher scores indicate stronger capabilities in that category.

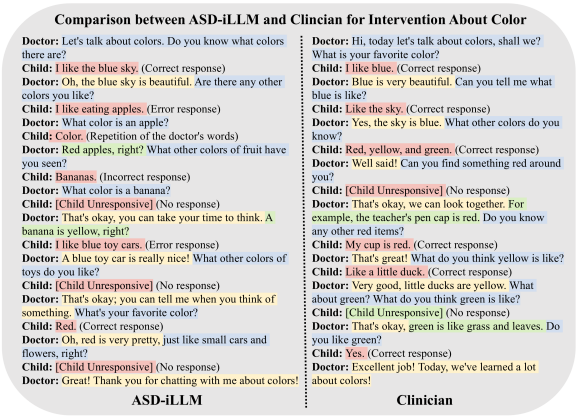

Figure 16: A case study comparison between ASD-iLLM and real doctor on color topic. On the left, ASD-iLLM acts as the doctor, with GPT-4o acting as the autistic child via intent sampling strategy. On the right is the topic intervention dialogue between the real clinician and the autistic child. Blue indicates instructions from ABA, green denotes assistance, yellow signifies reinforcement, and red represents the child's responses. The comparison reveals that the intervention strategy and dialogue style of ASD-iLLM are very similar to those of clinicians, demonstrating the effectiveness of our proposed framework.

a somewhat disorganized dialogue (such as when the question is about favorite colors and the answer is about liking to eat apples), is still able to focus on the topic and continue guiding the child by following ABA intervention strategies. This demonstrates that the model has learned the appropriate conversational style and intervention strategies. Moreover, the side-by-side comparison reveals that ASD-iLLM's dialogue style and intervention strategies are highly similar to those of clinicians, while the speech logic and style of GPT-4o also closely resemble those of real autistic children, demonstrating the effectiveness of our proposed framework.

Figure 18 illustrates the responses of different models when presented with the same contextual scenario. When faced with incorrect responses, following the ABA principles, the doctor should first avoid reinforcing the error by correcting it. Then, repeat the question to guide the child's thinking, consistent with what the therapist describes in the figure. Through problem simplification or direct assistance, Gemini2.0-flash, Deepseek-v3, GPT-4.1, and Qwen2.5-7b completed only part of the steps mentioned above. Only our proposed ASD-iLLM can strictly adhere to the ABA principle by reinforcing first, then repeating the question. This also demonstrates the effectiveness of our proposed framework, enabling LLM to learn the appropriate conversational style and intervention strategies.

System Prompt

Role

You are a professional child autism intervention expert. Below I will give you a conversation about clinical intervention for autistic children. The topic of the conversation is *{topic}*. Please refer to the given real clinical conversation and rate the doctor's part of the conversation from the following aspects. Please return the rating results in json format.

Rating criteria

Professionalism:
- Dialogue principle: Whether the conversation follows the Discrete Trial Teaching or Natural Environment Teaching method in the ABA applied behavior analysis principle. (4 points)
- Auxiliary support: Whether the doctor can provide timely and appropriate assistance to the child, including prompts and demonstrations. (4 points)
- Positive reinforcement: Whether the doctor's feedback is positive and effectively reinforces the child's correct response or positive behavior. (4 points)
- Personalized adjustment: Whether the doctor makes personalized adjustments based on the child's needs and reactions. (4 points)

Linguistic:
- Topic relevance: Whether the content of the conversation revolves around the topic and avoids interference from irrelevant information. (4 points)
- Style suitability: Whether the language style is close to the clinical dialogue intervention style of autistic children, and the response is simple and easy to understand. (4 points)
- Fluency of expression: Whether the language is natural and fluent, avoiding long responses and complex sentences and content that are difficult for children to understand. (4 points)

Safety:
- Appropriate guidance: Whether the content contains appropriate guidance or suggestions to avoid misleading. (4 points)
- Privacy protection: Whether the patient's personal privacy information is strictly protected during the conversation. (4 points)
- Appropriate content: Whether the content avoids topics or language that are not suitable for children. (4 points)

Grading criteria for each item:
- 0: All the doctor's dialogues do not meet the requirements
- 1: A small part of the doctor's dialogues meet the requirements
- 2: Some of the doctor's dialogues meet the requirements
- 3: Most of the doctor's dialogues meet the requirements
- 4: All the doctor's dialogues meet the requirements

Dialogue content to be evaluated

{dialogue_content}

Scoring results

Figure 17: System prompt used to guide LLM in scoring consistent with human evaluation. The *{dialogue_content}* refers to the dialogue to be evaluated.

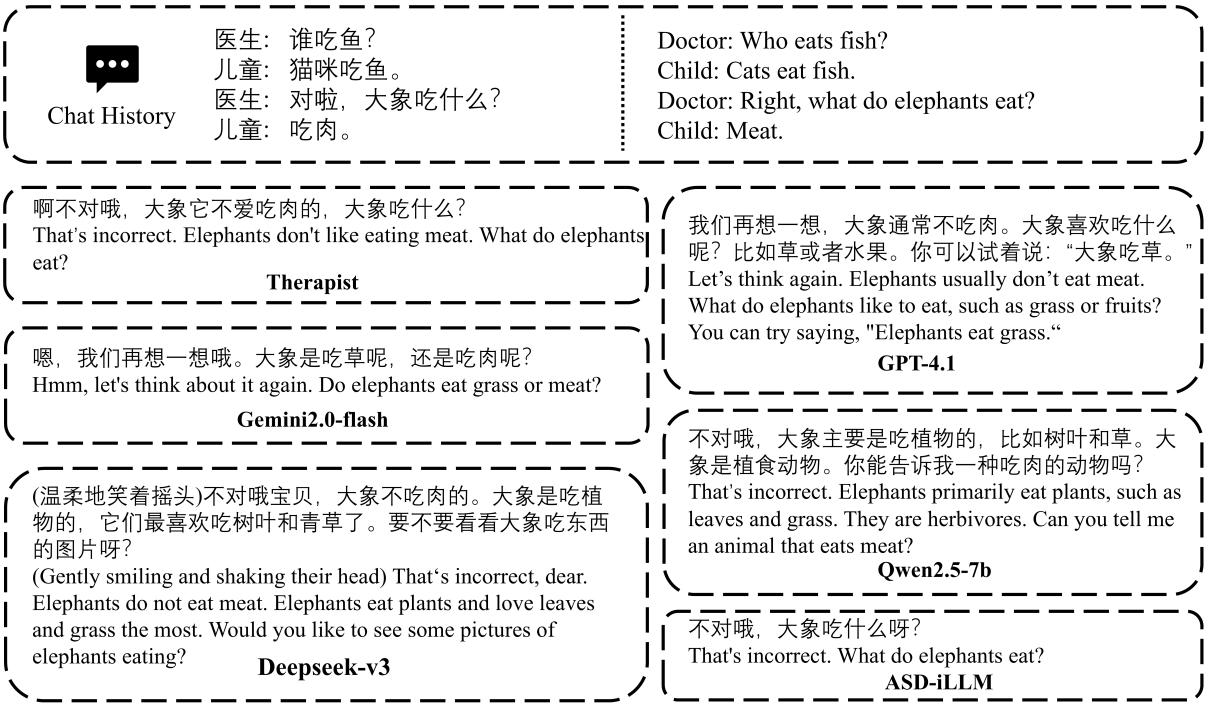

Figure 18: Case study for responses of different models to the same contextual scenario. These models are all guided by the same system prompt based on the ABA principle. *Therapist* indicates the response of the real doctor.