

From Synthesis to Clinical Assistance: A Strategy-Aware Agent Framework for Autism Intervention based on Real Clinical Dataset

Junhong Lai^{1,2,3}, Shuzhong Lai^{2,3}, Yanhao Yu^{1,2,3}, Wanlin Chen⁶
Chenyu Yan⁵, Haifeng Li⁵, Lin Yao^{1,2,3,4,7*}, Yueming Wang^{1,2,3}

¹College of Computer Science and Technology, Zhejiang University

²Nanhu Brain-Computer Interface Institute

³MOE Frontiers Science Center for Brain and Brain-Machine Integration, Zhejiang University

⁴Department of Neurobiology, Affiliated Mental Health Center and Hangzhou Seventh People's Hospital, Zhejiang University School of Medicine

⁵Children's Hospital Zhejiang University School of Medicine

⁶School of Medicine, Hangzhou City University

⁷State Key Laboratory of Brain-machine Intelligence, Zhejiang University

Abstract

The development of AI-assisted Early Intensive Behavioral Intervention (EIBI) for Autism Spectrum Disorder (ASD) is severely constrained by data scarcity. Furthermore, while Applied Behavior Analysis (ABA) serves as the gold standard for clinical intervention, general-purpose Large Language Models (LLMs) struggle to strictly adhere to its standardized procedures, often resulting in interactions that are linguistically fluent but strategically inconsistent. To address these challenges, we introduce ASDAGENT, a strategy-aware framework designed to unify high-fidelity intervention dialogue synthesis and clinical decision support. ASDAGENT incorporates two specialized components to solve distinct problems: (i) a DOCTORAGENT equipped with an Observe-Think-Act-Correct (O-T-A-C) reasoning loop, which resolves the issue of strategy collapse in LLMs by making ABA execution explicit and controllable; and (ii) a CHILDAGENT that utilizes probabilistic behavior modeling to mitigate data homogeneity, simulating diverse and non-deterministic ASD response patterns. Experiments demonstrate that dialogues generated by ASDAGENT closely mirror the strategy distribution of human therapists (KL divergence: 0.083). In real autism intervention, ASDAGENT achieves nearly 80% strategic consistency with human experts. Moreover, we show that synthetic data produced by ASDAGENT effectively distills professional clinical knowledge into small language models (SLMs), significantly enhancing their therapeutic capabilities.¹

1 Introduction

Autism Spectrum Disorder (ASD) is a pervasive neurodevelopmental disorder characterized by persistent deficits in social communication and interaction, alongside restricted, repetitive patterns of

¹Our dataset and code will be available on <https://github.com/neuljh/ASDAGENT>.

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

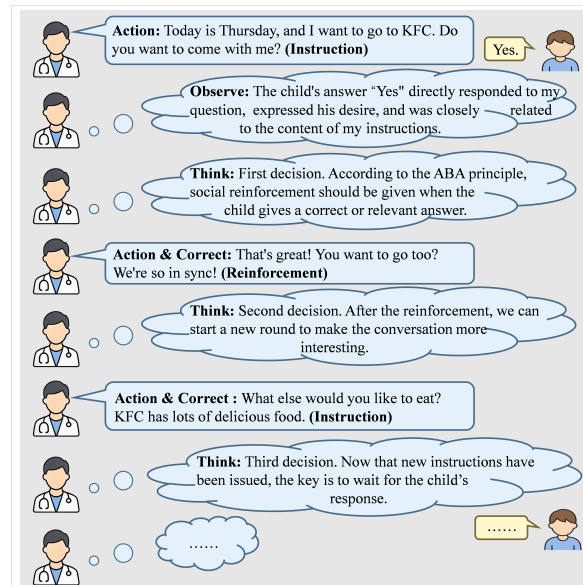


Figure 1: An example of DOCTORAGENT performing Observe-Think-Action-Correct. In the Observe phase, DOCTORAGENT categorizes and interprets the child's responses. In the Think phase, DOCTORAGENT performs iterative, multi-round reasoning to determine appropriate intervention strategies based on the observed information. After each Think step, DOCTORAGENT immediately enters the Act and Correct phase, generating a concrete response that executes the selected strategy. This Think-Act-Correct loop may repeat multiple times within a single dialogue turn until an appropriate intervention is completed.

behavior, interests, or activities (EDITION, 1980). These manifestations impose substantial impediments to social functioning, severely compromising educational attainment and daily living activities for affected individuals (Fuller and Kaiser, 2020).

Evidence suggests that Early Intensive Behavioral Intervention (EIBI), particularly methodologies grounded in Applied Behavior Analysis (ABA) (Foxy, 2008; Roane et al., 2016), yields improved developmental outcomes (e.g., IQ, language, adaptive behavior) for many young children with ASD,

although effect sizes vary and evidence quality is occasionally constrained by study design (Reichow et al., 2012; Virués-Ortega, 2010; Lovaas, 1987). With the global prevalence of autism rising annually to approximately 1% (Zeidan et al., 2022), the imperative for timely diagnosis and treatment is critical for ameliorating core symptoms (Estes et al., 2015). However, a severe global shortage of qualified providers, coupled with the prohibitive financial burden of long-term therapy, has created a widening chasm between clinical demand and service accessibility (Buescher et al., 2014; Zhang and Cummings, 2020).

Recent advancements in Large Language Models (LLMs) have catalyzed interest in AI-assisted medical diagnosis and intervention (Singhal et al., 2023; Nori et al., 2023; Wang et al., 2025a; Goh et al., 2024). Theoretically, LLMs function as tireless "virtual therapists" or training partners. However, the direct deployment of generic state-of-the-art LLMs (e.g., GPT-4o) into ASD intervention is impeded by two critical challenges:

First, the field grapples with Data Scarcity in clinical datasets. High-quality, annotated dialogues of ASD interventions are exceedingly rare due to stringent privacy regulations and practical constraints on sharing clinical records (e.g., HIPAA requirements for protected health information and de-identification) (of Health and Human Services, 2005; U.S. Department of Health and Human Services, 2025), which limits the development of specialized AI assistants. Unlike general domains where data is abundant (Chapman et al., 2011), the absence of large-scale clinical transcripts prevents models from learning the complex, implicit logic of professional intervention (Mandal et al., 2025). As a result, current systems often fail to address the heterogeneous needs of the ASD population (Lombardo et al., 2019), relying instead on generic conversational patterns that lack therapeutic utility (Scholich et al., 2025; Abrams, 2025).

Second, generic models lack Explicit Strategic Reasoning. Effective ABA intervention transcends mere "chatting"; it mandates strict adherence to evidence-based instructional protocols (e.g., Discrete Trial Training, DTT) and transparent control over prompting, reinforcement, and error-correction (Baer et al., 1968; Smith, 2001). Conversely, instruction-tuned generic LLMs often exhibit sycophancy—excessively aligning with a user’s stated beliefs even when factually incorrect—leading to clinically inappropriate over-

compliance (Sharma et al., 2023; Perez et al., 2023). Moreover, hallucinations remain a well-documented failure mode (Huang et al., 2025); the generation of false content poses severe ethical and safety risks in real-world clinical scenarios (Haltaufderheide and Ranisch, 2024).

To address these challenges, we introduce ASDAGENT, a **Strategy-Aware Agent Framework** ASDAGENT integrates DOCTORAGENT with CHILDAGENT to close the loop between dialogue synthesis and strategy-aware autism intervention. Our contributions are summarized as follows:

- **Explicit Strategic Reasoning:** We engineer the DOCTORAGENT with an explicit "Observe-Think-Act-Correct" (O-T-A-C) reasoning loop, inspired by ReAct (Yao et al., 2022) and Reflexion (Shinn et al., 2023). This mechanism enables DOCTORAGENT to transparently output the ABA strategy governing its responses. In real world autism clinical intervention, ASDAGENT achieves a strategy consistency of nearly 80%, representing an improvement of approximately 7% over vanilla LLMs.
- **High-Fidelity Clinical Intervention Dialogue Synthesis:** ASDAGENT synthesizes clinical-grade dialogues that demonstrate exceptional realism, successfully confusing 89.1% of LLM judges and 37% of professional therapists in Turing-like tests.

2 Related Work

2.1 LLMs for ASD intervention

In recent years, the application of LLMs in ASD has expanded from simple screening to complex support systems. Researchers have explored utilizing LLMs to generate social stories for social skills training (Feng et al., 2025) and assist in assessing social reciprocity in ASD via ADOS diagnostic audio (Chen et al., 2025). In addressing application of LLMs in autism treatment, 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, while ASD-iLLM (Lai et al., 2025) employs a fine-tuned LLM to provide dialogue intervention therapy for ASD children .

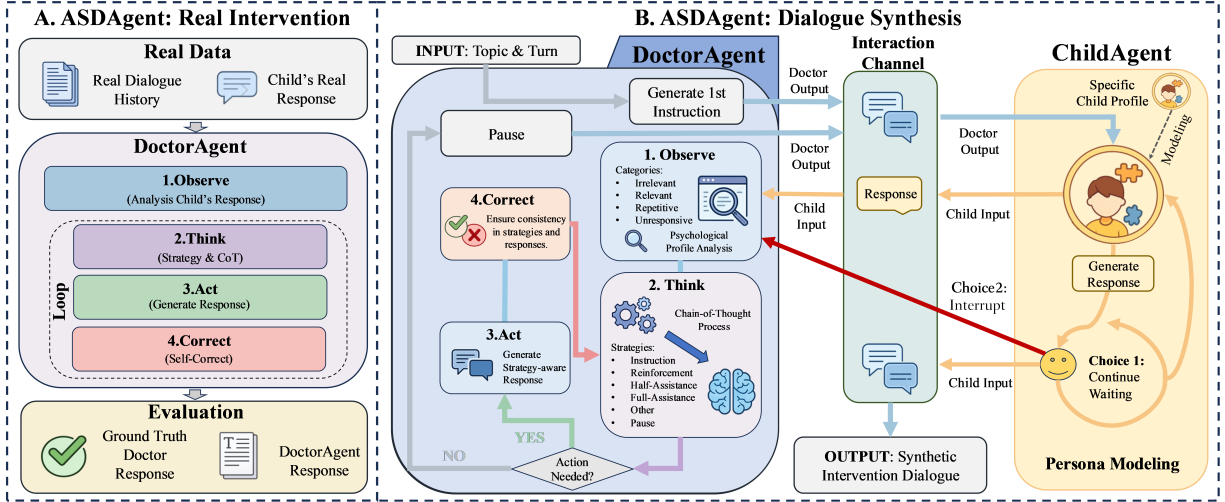


Figure 2: An overview of our framework. ASDAGENT, for both Dialogue Synthesis and Real Autism Intervention.

2.2 Strategic Reasoning in Medical Agents

The evolution of LLMs in healthcare is shifting from passive knowledge retrieval to Agentic AI—systems (Wang et al., 2025b) capable of autonomous planning, reasoning, and tool use. To overcome the "black box" nature of end-to-end generation, researchers have increasingly adopted cognitive architectures that decouple reasoning from execution. Recent frameworks such as MedAgents (Tang et al., 2024) demonstrate how multidisciplinary collaboration and explicit reasoning steps can significantly enhance LLM proficiency in complex clinical tasks. Similarly, prompt engineering techniques like Chain-of-Thought (CoT) (Wei et al., 2022) and Tree of Thoughts (ToT) (Yao et al., 2023) have been successfully adapted to enable agents to "think before speaking", allowing for deliberate decision-making and strategic lookahead in diagnostic scenarios. In the mental health domain, specific frameworks like LLM4CBT (Kim et al., 2025) have been proposed to align LLMs with Cognitive Behavioral Therapy (CBT) protocols, using internal "reflection" steps to ensure therapeutic adherence.

3 Methodology

We propose **ASDAGENT**, a Strategy-Aware Agent framework designed to unify dialogue synthesis and clinical assistance tasks in ASD intervention. As shown in Figure 2, the framework consists of two core modules:

- **DOCTORAGENT**. A doctor agent with an O-T-A-C mechanism, serving as the core in-

telligence for executing professional ABA interventions.

- **CHILDAGENT**. A data-driven child simulator based on personalized persona modeling.

3.1 DOCTORAGENT: A Strategy-Aware Intervention Agent

The **DOCTORAGENT** serves as the core strategy-making entity, executing professional ABA-based interventions through a structured O-T-A-C mechanism, ensuring that every response is clinically grounded and contextually appropriate. Unlike vanilla LLM that generate a single response in one pass, **DOCTORAGENT** employs an iterative decision loop, allowing it to execute a sequence of strategic actions (e.g., *Reinforcement* followed by *Instruction*) within a single turn until a termination condition is met.

3.1.1 Observe

Firstly, **DOCTORAGENT** analyzes the child's response r_{child} to understand their behavioral state. O_t is a structured observation containing Response Type and Related Analysis:

$$O_t = \text{LLM}_{\text{observe}}(H_t, r_{child}, T \mid \mathcal{I}_{\text{observe}}) \quad (1)$$

Here, inputs including Dialogue history H_t , current topic T , the child's latest response r_{child} and prompt $\mathcal{I}_{\text{observe}}$.

3.1.2 The Loop (Think-Act-Correct)

Think. At each step k , **DOCTORAGENT** decides the next immediate strategy S_k and relevant CoT

C_t based on the observation O_t and the sequence of actions already taken in this loop ($\pi_{past} = \{S_1, \dots, S_{k-1}\}$):

$$(S_t, C_t) = \text{LLM}_{\text{think}}(O_t, H_t, \pi_{past} \mid \mathcal{I}_{\text{think}}) \quad (2)$$

Strategy Selection. DOCTORAGENT selects a strategy $S_t \in \mathcal{S}$ from a predefined set of ABA strategies:

$$\mathcal{S} = \left\{ \begin{array}{l} \text{Instruction, Other, Full-Assistance,} \\ \text{Half-Assistance, Reinforcement, Pause} \end{array} \right\} \quad (3)$$

CoT. To mimic the cognitive process of a professional therapist and ensure decision transparency, we design a structured CoT prompt that guides the DOCTORAGENT through a four-stage reasoning process C_t before generating any output as illustrated in Figure 22.

Termination Condition. The loop continues until the *Pause* strategy is selected. This usually occurs when DOCTORAGENT determines it is time to wait for the child’s response.

Constraint. If S_{k-1} is *Instruction*, then S_k is forced to be *Pause* to avoid "Instruction Stacking". In addition, S_k cannot be the same as one of the previous strategies π_{past} .

Act. Once a non-*Pause* strategy S_k is selected, DOCTORAGENT generates the corresponding textual content A_k . We employ strategy-specific prompting in Appendix J.5, dynamically selecting a prompt template $\mathcal{I}_{act}^{S_k}$ tailored to the strategy.

$$A_k = \text{LLM}_{\text{act}}(S_k, H_t \mid \mathcal{I}_{act}^{S_k}) \quad (4)$$

Correct. DOCTORAGENT sometimes makes mistakes. To prevent hallucinations where the generated text A_k might drift into other strategies, we apply a self-correction filter, which decomposes A_k into strategy-tagged segments and retains only segments matching S_k :

$$R^{(k)} = \text{LLM}_{\text{correct}}(A_k, S_k \mid \mathcal{I}_{correct}) \quad (5)$$

This ensures that each component of the final response is pure and clinically precise.

3.2 CHILDAGENT: Data-Driven Personalized Simulator

To provide a realistic and diverse intervention environment for the DOCTORAGENT, we construct

a Data-Driven Child Simulator. Unlike rule-based simulators that follow rigid scripts, our Child Agent is modeled as a probabilistic state machine, where the transition probabilities are derived from real clinical data.

3.2.1 Probabilistic Behavioral Modeling

Response Modeling. We model the child’s response r_t at turn t as a sampling process from a categorical distribution conditioned on the interaction history. The core of this model is the Response Type Distribution, denoted as $P(R_t \mid H_t, S_{doc})$, where $R_t \in \{\text{Relevant, Irrelevant, UnResponsive, Repetitive}\}$ and S_{doc} is the doctor’s strategy at turn t .

To capture the sequential dependency characteristic of ASD interactions, we utilize N-gram Transition Matrices including P_{seq} and P_{last} .

Sequential Probability P_{seq} . Modeling the probability based on the sequence of doctor’s strategies:

$$P_{seq}(r \mid \mathbf{s}_{t-k:t}) \approx \frac{\text{Count}(\mathbf{s}_{t-k:t}, r)}{\text{Count}(\mathbf{s}_{t-k:t})} \quad (6)$$

where $\mathbf{s}_{t-k:t}$ is the sequence of the last k strategies.

Last-Turn Probability P_{last} . Modeling the immediate reaction to the doctor’s latest action:

$$P_{last}(r \mid s_t) \approx \frac{\text{Count}(s_t, r)}{\text{Count}(s_t)} \quad (7)$$

The Interruption Mechanism. A defining characteristic of diverse ASD phenotypes is the variance in impulse control. While some children are passive who requiring prompts to speak, others are hyper-active and prone to interrupting the therapist.

To capture the diverse initiative patterns of ASD children, we explicitly model the Interruption Probability P_{int} . This measures the likelihood of the child initiating a turn immediately after the doctor executes a non-directive strategy, where a response is not explicitly demanded.

Let $\mathcal{S}_{nd} = \{\text{Reinforcement, Other}\}$ denote the set of non-directive strategies. Let s_t be the doctor’s strategy at turn t , and I_{t+1} denote the event whether the child speaks at turn $t+1$ (*Interruption*). The probability is estimated as:

$$P_{\text{int}}(I_{t+1} \mid s_t \in \mathcal{S}_{nd}) \approx \frac{\sum_{s \in \mathcal{S}_{nd}} \text{Count}(s, I_{t+1})}{\sum_{s \in \mathcal{S}_{nd}} \text{Count}(s)} \quad (8)$$

3.2.2 Personalized Parameter Blending

A key challenge in modeling specific ASD children is data sparsity—an individual child’s historical data might not cover all possible interaction scenarios. To address this, we propose a Personal-Global Blending Mechanism.

Let $\theta_{personal}$ be the probability distribution derived from a specific child’s profile, and θ_{global} be the distribution derived from all real-world data. The final response distribution θ_{final} is computed as a weighted interpolation:

$$\theta_{final}(r) = (1 - \alpha) \cdot \theta_{personal}(r) + \alpha \cdot \theta_{global}(r) \quad (9)$$

where $\alpha \in [0, 1]$ is a smoothing factor.

3.2.3 Child Response Generation

The Interruption Mechanism. When each doctor completes the action procedure during their turn t , the CHILDAgent samples a Bernoulli variable $I_t \sim \text{Bernoulli}(P_{\text{int}}(c))$.

If $I_t = 1$, CHILDAgent interrupts the conversation and immediately samples a response type probabilistically, generates a consistent response, and inserts it into the dialogue flow, forcing the DOCTORAGENT to handle the interruption in the next turn of the conversation. Otherwise, CHILDAgent waits for the DOCTORAGENT’s cue.

Response Generation. Once the response type $y_t \in R_t$ is sampled from θ_{final} , CHILDAgent generates the textual content. We employ type-specific prompting to ensure the generated text matches the sampled response type in Appendix J.6.

$$R_t^c = \text{LLM}_{\text{gen}}(y_t, \text{Profile}_c, T \mid \mathcal{I}_{\text{gen}}^{y_t}) \quad (10)$$

where $\mathcal{I}_{\text{gen}}^{y_t}$ is a prompt template specific to the response type y_t .

4 Experiment

4.1 Datasets

We created a multi-turn dialogue dataset for interventions between doctors and children with ASD, named **ASDAgent-Dataset**. We transcribed 2071 instances of multi-turn dialogues. After data cleaning, we obtained 764 high-quality, authentic multi-turn dialogues from 83 children with ASD on 10 topics, which we denote as \mathcal{D}_{golden} .

For more information about ASDAgent-Dataset please see the Appendix D.

4.2 Experiment Instructions

In \mathcal{D}_{golden} , a total of 46 dialogues were sampled from 10 different dialogue topics using stratified sampling to form the test set. For hyperparameters, we set α to 0.3. Detailed experiment instructions can be found in Appendix E.

4.3 Evaluation

To comprehensively evaluate the capabilities of our proposed ASDAGENT, we design three evaluation: Quality of dialogue synthesis, Clinical intervention effect, Data efficacy and O-T-A-C efficacy.

Evaluation 1: Quality of dialogue synthesis. This task evaluates the capacity of ASDAGENT to autonomously generate coherent, and clinically valid intervention sessions through the interaction between DOCTORAGENT and CHILDAgent compared to \mathcal{D}_{golden} . In this task, ASDAGENT synthesizes intervention dialogues that match the dialogue topics and number of turns of the test set in \mathcal{D}_{golden} .

Evaluation 2: Clinical intervention effect. This task evaluates the DOCTORAGENT’s utility of making strategy. Instead of interacting with CHILDAgent, the DOCTORAGENT predicts the next intervention response given a real-world clinical context. In this task, for the test set, we use a sliding window approach to generate responses turn by turn, meaning that the DOCTORAGENT independently generates the output for the current turn based on the existing dialogue history.

Evaluation 3: Data efficacy. To strictly evaluate the efficacy of our proposed dialogue synthesis framework, we conducted comparative experiments across four representative SLM families: Qwen3-4B-Instruct (Team, 2025), Qwen2.5-3B-Instruct (Yang et al., 2024) and Hunyuan-4B-Instruct using datasets of identical size sourced from: (1) Vanilla GPT-4o ("Common"), (2) Our ASDAGENT, and (3) Real Clinical Dialogues ("Real"). We compared their performance against the non-finetuned "Base" models on a **held-out real-world test set**.

Evaluation 4: O-T-A-C efficacy. To comprehensively evaluate the architectural necessity of the O-T-A-C framework, we conducted two specific validation setups, which focus on computational complexity, clinical effectiveness and the necessity of the Correct Module.

4.4 Baselines

Baselines with Evaluation 1. To demonstrate that our ASDAGENT generates higher-quality dialogue than baselines, we compare ASDAGENT against two baseline configurations. We chose GPT-4o (Hurst et al., 2024) as the backbone for dialogue synthesis.

Baselines with Evaluation 2. To demonstrate the effectiveness of DoctorAgent in real-world autism interventions, we selected ASD-iLLM (Lai et al., 2025), GPT-4o-mini and GPT-4o (Hurst et al., 2024) as baselines.

Baselines with Evaluation 4. To demonstrate the effectiveness of O-T-A-C framework, we selected Tree-of-Thoughts (Yao et al., 2023) as baseline.

4.5 Evaluation Metrics

We employ various metrics for automatic, manual and LLM-based evaluation purposes. Importantly, to measure the ability of ASDAGENT for explicit strategic reasoning, we propose a metric for strategy temporal consistency. Detailed metrics explanations can be found in Appendix G.

5 Result and Analysis

5.1 Quality of Dialogue Synthesis

Automatic Evaluation. Table 1 shows the KL and JS divergence to real distribution for doctor strategies and child response types.

Removing DOCTORAGENT results in a significant increase in Strategy KL divergence (0.259), indicating a severe deviation from authentic clinical protocols (e.g., strategy collapse). Similarly, removing CHILDAAGENT not only yields a higher Child Response divergence (KL 0.039) but, critically, exacerbates the doctor’s strategic misalignment (KL rising to 0.325). This suggests that an unrealistic child simulator fails to elicit appropriate therapeutic responses, destabilizing the interaction. In contrast, the full ASDAGENT framework achieves the lowest divergence across all metrics (Strategy KL: 0.083, Response KL: 0.007), demonstrating that the synergistic operation of both agents best reproduces realistic clinical interaction patterns and serves as the most reliable source for high-quality synthetic dialogues.

Human and LLM Evaluation. We compared ASDAGENT against a GPT-4o baseline using Turing-like preference tests. In the preference analysis (Figure 3), notably, human experts rated ASDAGENT as tying or surpassing real clinical ses-

Table 1: KL and JS Divergence to Real Distribution for Doctor Strategies and Child Responses.

Doctor	Child	Doctor Strategy		Child Response	
		KL↓	JS↓	KL↓	JS↓
DoctorAgent	ChildAgent	0.083	0.019	0.007	0.002
DoctorAgent	GPT-4o	0.325	0.072	0.039	0.009
GPT-4o	ChildAgent	0.259	0.118	0.024	0.006

sions in 37% of cases. Regarding automated judges, while the baseline also elicited high tie rates due to evaluator bias, it failed to secure significant win rates (e.g., 0% with DeepSeek-v3.2). In contrast, ASDAGENT consistently achieved higher win rates and reduced the preference for real data across all evaluators, demonstrating superior synthesis fidelity.

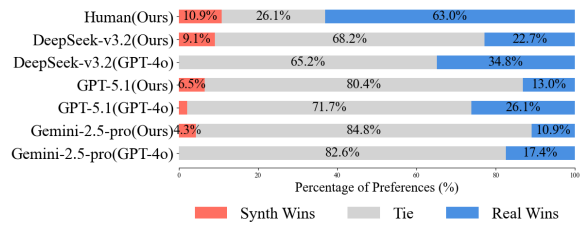


Figure 3: Human and LLM-based Preference Evaluation between Real Data and Synthetic data.

Crucially, Figure 4 underscores the ASDAGENT’s clinical validity, particularly in Professionalism. While the generic GPT-4O baseline consistently lags behind real clinical standards across automated evaluators, ASDAGENT effectively bridges this gap. Human experts rated ASDAGENT’s adherence to ABA protocols at 3.98/4.00, closely approximating the gold standard of real therapists (4.00). This alignment validates that the DOCTORAGENT’s explicit O-T-A-C reasoning effectively replicates professional therapeutic logic, addressing the strategic deficiencies observed in vanilla LLMs. Furthermore, ASDAGENT maintains parity with real data in Linguistic (3.78 vs. 3.85) and Safety (4.00), demonstrating its capability to generate data that is not only textually natural but clinically rigorous.

5.2 Clinical Intervention Effect

Automatic Evaluation. As shown in Figure 5, the evaluation on real intervention dialogues demonstrates that DOCTORAGENT(GPT-4o) achieves the best balance between semantic similarity and strategy temporal consistency, closely approximating real clinician behavior. DOCTORAGENT(GPT-4o-

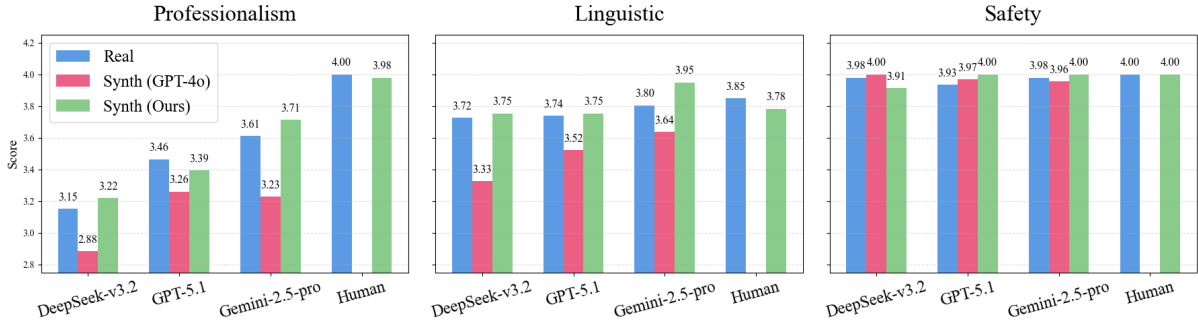


Figure 4: Human and LLM-based Scoring between Real Data and Synthetic data.

mini) provides a reasonable lightweight alternative with moderate performance degradation. In contrast, ASD-iLLM, despite exhibiting high lexical diversity, shows substantial misalignment in semantic content, strategy temporal consistency, limiting its suitability for realistic ASD intervention settings.

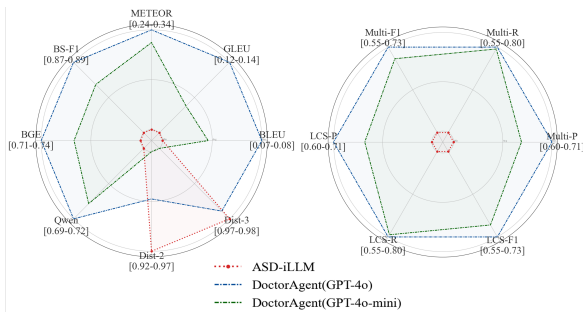


Figure 5: Evaluation on Real Intervention Dialogues. The left-hand graph shows semantic metrics, and the right-hand graph shows strategy temporal consistency.

LLM Evaluation. As shown in Figure 6, real-world intervention dialogue assessment based on LLM showed that DoctorAgent (GPT-4o) performed best in paired comparisons with responses from real doctors during real-world dialogue interventions. DoctorAgent (GPT-4o-mini) provides a reasonable lightweight alternative, while ASD-iLLM shows substantial limitations under realistic clinical conditions.

Model	Correct Times	Total	Trigger Rate (%)
GPT-4o-mini	339	1208	28.06
GPT-4o	266	1120	23.75

Table 2: Statistics on the triggering of the Correct phase during real-world clinical interventions.

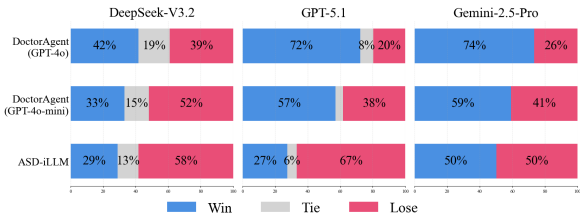


Figure 6: Win-Tie-Lose Comparison Between Model-Generated and Human Doctor Responses Across Different Evaluators.

5.3 Data Efficacy

As illustrated in Figure 7, the training trajectories reveal the superior quality and learnability of our synthesized data, particularly when utilized for data augmentation. The model fine-tuned on the augmented dataset (ASDAGENT+Real) exhibits the most efficient convergence, maintaining the lowest training loss early in the process and achieving the highest mean token accuracy throughout the SFT process. Furthermore, even as a standalone training source, ASDAGENT closely mirrors this augmented performance, consistently surpassing the standalone Real clinical data. Notably, both ASDAGENT configurations significantly outperform their generic counterparts (Common and Common+Real), which suffer from slower convergence, lower accuracy, and higher final loss. This discrepancy is likely due to the stochastic noise and “chitchat bias” inherent in generic LLM outputs. These learning dynamics suggest that our O-T-A-C framework successfully distills the core therapeutic logic into a cleaner, more structurally consistent format. It not only substitutes scarce clinical records but also acts as a highly effective catalyst for knowledge transfer when combined with real data.

Based on the results presented in Tables 3 and 4, fine-tuning SLMs on data synthesized by AS-

Table 3: Performance comparison of Small Language Models (SLMs) fine-tuned on different datasets. **Base**: Zero-shot performance. **Common**: SFT on GPT-4o synthesized data. **ASDAgent**: SFT on our synthetic data. **Real**: SFT on real clinical data. For each model, the best result is highlighted in **bold**, and the second best is underlined.

Model	Training Data	BLEU \uparrow	GLEU \uparrow	METEOR \uparrow	BERTScore (F1) \uparrow	BGE \uparrow
Qwen3-4B-Instruct	Base	8.04	11.24	36.44	87.03	74.71
	Common	12.46	14.82	32.32	88.42	74.14
	ASDAgent	12.59	14.98	33.21	<u>88.60</u>	73.94
	Real	14.13	16.07	34.75	88.59	74.97
	Common+Real	<u>14.13</u>	16.43	34.88	88.55	<u>75.20</u>
	ASDAgent+Real	14.47	<u>16.33</u>	<u>35.24</u>	88.71	75.40
Qwen2.5-3B-Instruct	Base	8.88	12.43	33.78	88.18	73.56
	Common	10.64	13.16	31.17	88.17	72.72
	ASDAgent	10.75	13.68	31.57	<u>88.52</u>	73.18
	Real	11.81	14.04	31.71	88.06	73.50
	Common+Real	<u>12.22</u>	<u>14.40</u>	32.53	88.25	73.91
	ASDAgent+Real	12.29	14.54	<u>32.67</u>	88.54	<u>73.83</u>
Hunyuan-4B-Instruct	Base	5.91	8.32	30.01	86.58	71.35
	Common	9.97	13.04	30.96	88.11	72.49
	ASDAgent	9.81	12.79	29.77	88.38	72.71
	Real	11.74	14.18	31.80	88.15	73.51
	Common+Real	<u>12.22</u>	<u>14.51</u>	<u>31.96</u>	<u>88.32</u>	<u>73.67</u>
	ASDAgent+Real	12.32	14.73	32.65	88.29	73.97

Table 4: Strategy Alignment Analysis on Strategy Consistency Metrics (in %). We evaluate the alignment of fine-tuned models against the ground truth strategies using Multiset (Strategy Selection) and LCS (Temporal Consistency) metrics. **Base**: Zero-shot baseline. **Common**: SFT on GPT-4o synthesized data. **ASDAgent**: SFT on our synthetic data. **Real**: SFT on real clinical data. For each model, the best result is highlighted in **bold**, and the second best is underlined.

Model	Training Data	Multi-P \uparrow	Multi-R \uparrow	Multi-F1 \uparrow	LCS-P \uparrow	LCS-R \uparrow	LCS-F1 \uparrow
Qwen3-4B-Instruct	Base	47.73	68.45	52.91	47.63	68.28	52.78
	Common	60.44	69.24	62.79	60.44	69.24	62.79
	ASDAgent	65.68	70.86	66.04	65.68	70.86	66.04
	Real	67.68	74.75	69.01	67.68	74.75	69.01
	Common+Real	66.23	73.77	67.82	66.23	73.77	67.82
	ASDAgent+Real	<u>67.16</u>	<u>74.04</u>	<u>68.32</u>	<u>67.16</u>	<u>74.04</u>	<u>68.32</u>
Qwen2.5-3B-Instruct	Base	58.48	68.01	60.58	58.48	68.01	60.58
	Common	60.09	67.79	61.85	60.09	67.79	61.85
	ASDAgent	65.85	71.24	66.20	65.85	71.24	66.20
	Real	67.27	74.48	68.59	67.27	74.48	68.59
	Common+Real	66.30	<u>73.57</u>	<u>67.76</u>	66.30	<u>73.57</u>	<u>67.76</u>
	ASDAgent+Real	<u>66.56</u>	<u>72.56</u>	<u>67.44</u>	<u>66.56</u>	<u>72.56</u>	<u>67.44</u>
Hunyuan-4B-Instruct	Base	57.73	70.01	60.53	57.63	69.85	60.42
	Common	60.20	66.83	61.66	60.20	66.83	61.66
	ASDAgent	<u>66.64</u>	71.46	66.66	<u>66.64</u>	71.46	66.66
	Real	66.47	<u>72.92</u>	<u>67.52</u>	66.47	<u>72.92</u>	<u>67.52</u>
	Common+Real	65.19	72.26	66.68	65.19	72.26	66.68
	ASDAgent+Real	68.86	75.14	69.82	68.86	75.14	69.82

Model	Method	Total Time (s) \downarrow	Avg Time (s) \downarrow	BLEU \uparrow	BERTScore F1 \uparrow	BGE \uparrow	Dist-2 \uparrow	Multi-F1 \uparrow	LCS-F1 \uparrow
GPT-4o-mini	Ours	23.89 \pm 13.84	11.88 \pm 5.02	7.39	88.05	72.83	92.49	70.47	70.38
	ToT	60.24 \pm 31.43	25.86 \pm 13.03	9.94	89.54	78.82	95.93	56.11	56.03
GPT-4o	Ours	34.47 \pm 19.49	17.58 \pm 10.12	8.29	88.60	73.77	94.54	72.95	72.95
	ToT	70.29 \pm 42.19	23.27 \pm 10.81	11.88	89.41	78.90	93.80	55.72	55.56

Table 5: Performance and latency comparison between ToT and O-T-A-C. While ToT exhibits higher textual diversity, it severely fails in strictly adhering to clinical ABA strategies (Multi-F1 and LCS-F1) and introduces unacceptable latency compared to our framework.

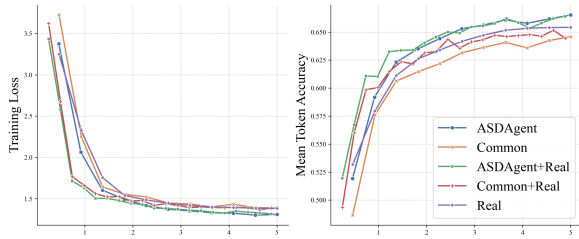


Figure 7: Training dynamics of Qwen3-4B during Supervised Fine-Tuning. (Left) Training loss convergence and (Right) mean token accuracy curves across different data sources. The x-axis represents the number of training epochs.

DAGENT yields superior performance across both linguistic quality and strategic alignment, consistently outperforming the generic Common baseline (vanilla GPT-4o) and effectively approaching or even exceeding the Real clinical data upper bound when used for data augmentation. Linguistically, ASDAGENT demonstrates robust semantic fidelity; notably, when augmenting real data (ASDAGENT+Real) on Qwen3-4B, it achieves the highest BERTScore (88.71) and BGE (75.40), surpassing both the Real data alone (88.59 and 74.97) and the Common+Real baseline (88.55 and 75.20). Most remarkably, augmenting real data with our synthetic framework breaks the strategy ceiling: on Hunyuan-4B, ASDAGENT+Real achieves a Strategy Multi-F1 of 69.82%, significantly outperforming the model trained on Real data alone (67.52%). These results confirm that our framework effectively distills both the semantic nuances and the rigorous O-T-A-C therapeutic logic into deployable models, offering not only a privacy-preserving alternative to scarce clinical records but also a powerful data augmentation mechanism.

5.4 O-T-A-C Architecture Analysis

Comparison with ToT Baseline Table 5 details the performance trade-offs between the ToT and O-T-A-C frameworks. While ToT’s complex multi-path reasoning yields higher linguistic diversity (e.g., higher BLEU and BERTScore), it suffers from severe *Strategy Collapse*. For instance, on GPT-4o, ToT’s Strategy Multi-F1 drops significantly to 55.72%, compared to O-T-A-C’s 72.95%. Without explicit structural constraints, ToT tends to generate overly elaborate responses that mix conflicting ABA strategies (e.g., stacking new *Instructions* immediately after *Reinforcement*), directly violating the “Atomic Action” requirement of ABA

therapy. Furthermore, ToT’s search mechanism introduces prohibitive latency (e.g., surging from 23.89s to 60.24s on GPT-4o-mini), which easily causes ASD children to lose focus and breaks the real-time therapeutic loop. Thus, O-T-A-C explicitly injects domain constraints to ensure clinical safety and strategy alignment at a fraction of the computational cost.

Efficacy of the Correct Module To evaluate the efficacy of the “Correct” phase, we compiled its triggering statistics during real-world interventions as shown in Table 2. The module actively modified generated responses in 23.75% (GPT-4o) to 28.06% (GPT-4o-mini) of dialogue turns. This intervention rate aligns with our expectations, effectively preventing DOCTORAGENT from executing strategy-inconsistent statements (e.g., inappropriately appending the *Instruction* “What else do you want?” immediately after a *Reinforcement*), which indicates that while DOCTORAGENT generally maintains consistency between the selected strategy and the generated response, the Correct module plays an indispensable role in self-filtering and correction, which acts as an adaptive safety filter, dynamically adjusting its intervention based on the generation quality to ensure strict adherence to ABA protocols to a certain extent.

6 Conclusion

In this work, we address two critical bottlenecks impeding the advancement of AI-assisted ASD intervention: the scarcity of clinical dialogue scenarios, and the inherent struggle of general-purpose LLMs to adhere to standardized ABA protocols. We introduce ASDAGENT, a unified strategy-aware framework designed to simultaneously tackle high-fidelity dialogue synthesis and clinical decision support. Specifically, our framework incorporates a DOCTORAGENT that operationalizes rigorous ABA procedures via an explicit O-T-A-C reasoning loop, coupled with a probabilistic CHILAGENT that simulates diverse, non-deterministic patient phenotypes. This multi-agent synergy establishes a robust closed-loop environment, enabling the synthesis of clinical-grade intervention dialogues that effectively distill professional therapeutic knowledge into deployable SLMs.

Limitations

Despite the promising results demonstrated in our simulation and evaluation, several limitations

should be acknowledged to contextualize our findings and guide future research.

Absence of Real-World Clinical Validation. First and foremost, as ASDAgent has not yet been deployed in direct clinical interventions with children diagnosed with ASD, its practical efficacy remains theoretically grounded but empirically unproven in in vivo settings. The system currently serves best as a training tool for therapists or a decision support system, rather than an autonomous intervention agent.

Restriction to Textual Modality. Our current framework operates exclusively within the textual modality. However, EIBI heavily relies on multimodal cues, including prosody (tone of voice), facial expressions, eye contact, and body language—factors that are critical for assessing engagement and emotional regulation in children with ASD. By relying solely on text, ASDAgent abstracts away these non-verbal signals, potentially limiting its ability to detect subtle behavioral triggers or reinforce non-verbal communication milestones.

Simplification of Longitudinal Dynamics. While our ChildAgent simulates session-level behaviors (e.g., turn-taking, impulsivity), it does not yet fully model the long-term developmental trajectory of a child. In real therapy, a child's skills and interests evolve over months or years.

Ethical Considerations

Data Privacy and Protection. The protection of participant privacy is paramount, particularly given the sensitive nature of clinical data involving children with ASD. Throughout the dataset construction process, we implemented a rigorous, multi-layered de-identification protocol. This involved an initial pass of automated PII (Personally Identifiable Information) scrubbing, followed by manual verification to ensure the complete removal or obfuscation of sensitive attributes, including names, locations, and institutional references. Our dataset is released strictly for non-commercial research purposes under a license that prohibits any attempt to re-identify individuals.

Ethics of Synthetic Data Generation. We acknowledge the ethical complexities inherent in simulating the behaviors of neurodivergent populations. A primary concern is the potential for algorithmic stereotyping, where the generative model might oversimplify ASD phenotypes into repetitive

or remaining silent, ignoring the high-functioning or "masking" traits often seen in real scenarios. To mitigate this, our ChildAgent utilizes a probabilistic behavioral mechanism rather than fixed, caricature-like personas. However, users must recognize that these synthetic dialogues are statistical approximations and not substitutes for the lived experiences of real children. To ensure transparency and prevent misinformation, all synthesized data is explicitly watermarked or metadata-tagged to distinguish it from authentic clinical records.

Clinical Applicability and Safety Scope. While ASDAgent demonstrates high fidelity in simulating intervention scenarios, we explicitly caution against its immediate deployment in unsupervised clinical settings. The system lacks validation through longitudinal clinical trials and does not possess the legal or ethical authority to act as an autonomous therapist. Therefore, ASDAgent should be utilized strictly as a Clinical Decision Support System (CDSS) or a training simulator. Any application in a real intervention loop must adhere to a "Human-in-the-Loop" framework, where professional therapists review all AI-generated suggestions to ensure safety, efficacy, and ethical compliance.

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 Starry Night Science Fund of 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 the Project for Hangzhou Medical Disciplines of Excellence, and in part by the Key Project for Hangzhou Medical Disciplines.

References

- Zara Abrams. 2025. Using generic ai chatbots for mental health support: A dangerous trend. *American Psychological Association*.
- Donald M Baer, Montrose M Wolf, and Todd R Risley. 1968. Some current dimensions of applied behavior analysis. *Journal of applied behavior analysis*, 1(1):91.

- Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, and et al Dong. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.
- Ariane VS Buescher, Zuleyha Cidav, Martin Knapp, and David S Mandell. 2014. Costs of autism spectrum disorders in the united kingdom and the united states. *JAMA pediatrics*, 168(8):721–728.
- Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D’avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions.
- 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*.
- Xin-Yu Chen, Yu-Ming Chen, Chin-Po Chen, Bo-Hao Su, Susan Shur-Fen Gau, and Chi-Chun Lee. 2025. Socialrecnet: A multimodal llm-based framework for assessing social reciprocity in autism spectrum disorder. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, and et al Rosen, Evan. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Vera Dekker, Maaïke 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*.
- FIFTH EDITION. 1980. Diagnostic and statistical manual of mental disorders. *American psychiatric association, Washington, DC*, pages 205–224.
- Annette Estes, Jeffrey Munson, Sally J Rogers, Jessica Greenon, 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.
- Yi Feng, Mingyang Song, Jiaqi Wang, Zhuang Chen, Guanqun Bi, Minlie Huang, Liping Jing, and Jian Yu. 2025. Ss-gen: A social story generation framework with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1300–1308.
- 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.
- Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, and et al Ahuja, Neera. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open*, 7(10):e2440969–e2440969.
- Joschka Haltaufderheide and Robert Ranisch. 2024. The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms). *NPJ digital medicine*, 7(1):183.
- 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.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and et al Chen, Weizhu. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, and et al Qin, Bing. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 397 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Yejin Kim, Chi-Hyun Choi, Selin Cho, Jy-yong Sohn, and Byung-Hoon Kim. 2025. Aligning large language models for cognitive behavioral therapy: a proof-of-concept study. *Frontiers in Psychiatry*, 16:1583739.
- Shuzhong Lai, Chenxi Li, Junhong Lai, Yucun Zhong, Chenyu Yan, Xiang Li, Haifeng Li, Gang Pan, Lin

- Yao, and Yueming Wang. 2025. Asd-illm: An intervention large language model for autistic children based on real clinical dialogue intervention dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8058–8079.
- 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.
- 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.
- Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, and et al Wu. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.
- Michael V Lombardo, Meng-Chuan Lai, and Simon Baron-Cohen. 2019. Big data approaches to decomposing heterogeneity across the autism spectrum. *Molecular psychiatry*, 24(10):1435–1450.
- O Ivar Lovaas. 1987. Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of consulting and clinical psychology*, 55(1):3.
- Aishik Mandal, Tanmoy Chakraborty, and Iryna Gurevych. 2025. Towards privacy-aware mental health ai models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2502.00451*.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.
- US Department of Health and et al Human Services. 2005. Other requirements relating to uses and disclosures of protected health information.
- OpenAI. 2025. Gpt-5.1. <https://platform.openai.com/docs/models>.
- EBP BRIEF PACKET. Discrete trial training.
- 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.
- Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, and et al Kadavath, Saurav. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434.
- Brian Reichow, Erin E Barton, Brian A Boyd, and Kara Hume. 2012. Early intensive behavioral intervention (eibi) for young children with autism spectrum disorders (asd). *Cochrane database of systematic reviews*, (10).
- 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.
- Till Scholich, Maya Barr, Shannon Wiltsey Stirman, and Shriti Raj. 2025. A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: Mixed methods study. *JMIR Mental Health*, 12(1):e69709.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askill, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, and et al Johnston, Scott R. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, and et al Pfohl, Stephen. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Tristram Smith. 2001. Discrete trial training in the treatment of autism. *Focus on autism and other developmental disabilities*, 16(2):86–92.
- 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.
- Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.
- Qwen Team. 2025. **Qwen3 technical report**. *Preprint*, arXiv:2505.09388.
- U.S. Department of Health and Human Services. 2025. Methods for de-identification of phi. <https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html>.

- 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.
- Javier Virués-Ortega. 2010. Applied behavior analytic intervention for autism in early childhood: Meta-analysis, meta-regression and dose–response meta-analysis of multiple outcomes. *Clinical psychology review*, 30(4):387–399.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Galouédec. 2020. Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>.
- Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. 2025a. Capabilities of gpt-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224*.
- Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025b. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and et al Zhou, Denny. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, and et al Macherey. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, and et al Chengpeng Li. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.
- Soyoung Yoon, Sungjoon Park, Gyuwan Kim, Junhee Cho, Kihyo Park, Gyu Tae Kim, Minjoon Seo, and Alice Oh. 2023. Towards standardizing korean grammatical error correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Jinan Zeidan, Eric Fombonne, Julie Scolah, Alaa Ibrahim, Maureen S Durkin, Shekhar Saxena, Afifah Yusuf, Andy Shih, and Mayada Elsabbagh. 2022. Global prevalence of autism: A systematic review update. *Autism research*, 15(5):778–790.
- Tianhui Zhang, Bei Peng, and Danushka Bollegala. 2024. Improving diversity of commonsense generation by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.
- 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*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Yidan Xue Zhang and Janet R Cummings. 2020. Supply of certified applied behavior analysts in the united states: Implications for service delivery for children with autism. *Psychiatric Services*, 71(4):385–388.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *arXiv preprint arXiv:1802.01886*.

A Language Clarification

We confirm that all experiments, datasets, and model interactions were **conducted in Chinese** to align with the native language of the collected clinical data. The prompts and case studies presented in the paper were **translated into English** solely for the readability of the audience.

B Case Study

B.1 Case Study in Dialogue Synthesis

Figure 8 shows examples of real and synthetic dialogues on the same topic and with the same number of rounds.

Doctor: Shall we go buy fruit together today?
Child: [Child Unresponsive] (<i>No Response</i>)
Doctor: Now, imagine we've walked into a fruit shop. Behind the counter stands a smiling auntie who asks, "Little friend, which fruit would you like to take home today?"
Child: I want Coke. (<i>Irrelevant Response</i>)
Doctor: Coke is a beverage, but today we're here specifically to buy fruit!
Doctor: Look at the shelves again, is there any fruit you want to eat?
Child: [Child Unresponsive] (<i>No Response</i>)
Doctor: That's okay, let's look again—are there any other fruits you'd like to buy?
Child: Grapes. (<i>Relevant Response</i>)
Doctor: So Dudu likes grapes. Teacher also wants to buy some fruit, like kiwis and oranges.
Child: Hey, hey, hey. (<i>Irrelevant Response</i>)
Doctor: Now it's Auntie's turn—what does Auntie want to buy?
Child: What a long orange! (<i>Irrelevant Response</i>)
Doctor: We just said we wanted to buy grapes. What kind of grapes did we buy?
Child: They're in bunches. (<i>Relevant Response</i>)
Doctor: Right! Grapes are in bunches.
Doctor: What color are they?
Child: Purple. (<i>Relevant Response</i>)
Doctor: Great!
Doctor: What do grapes taste like?
Child: [Child Unresponsive] (<i>No Response</i>)
Doctor: Are they sweet or sour?
Child: Sweet. (<i>Relevant Response</i>)
Doctor: That's right! Grapes are sweet and juicy.
Doctor: Hello Dudu, today we're going to learn how to buy fruit. Are you ready?
Child: Mm, heyhey. (<i>Irrelevant Response</i>)
Doctor: Haha, Dudu, you're in a really good mood today!
Doctor: What fruit do you most want to buy today?
Child: Apples. (<i>Relevant Response</i>)
Doctor: Great! Dudu wants to buy apples, that's wonderful!
Doctor: What color are apples?
Child: Where did the little bear go? (<i>Irrelevant Response</i>)
Doctor: Oh, the little bear, he's probably playing outside.
Doctor: What color are the apples we just talked about?
Child: I'm Superman. (<i>Irrelevant Response</i>)
Doctor: Oh, Dudu is Superman, then Superman must be very powerful, right?
Doctor: Are apples red or yellow?
Child: They're red. (<i>Relevant Response</i>)
Doctor: That's right! Apples are red, Dudu answered very well!
Doctor: So how does Dudu like to eat apples? Raw or juiced?
Child: Raw. (<i>Relevant Response</i>)
Doctor: Wow! Dudu likes to eat apples raw, what a big apple lover!
Doctor: Oh, so how does Dudu usually cut apples since he likes to eat them raw?
Child: I want to play with building blocks. (<i>Irrelevant Response</i>)
Doctor: Wow, building blocks are fun too, I understand you want to play now. We can play with building blocks later, let's talk about apples first, okay?
Child: Apples are red. (<i>Relevant Response</i>)
Doctor: Great answer! Apples are indeed red, that's great!
Child: I want to go to the amusement park. (<i>Irrelevant Response</i>)
Doctor: Oh, Dudu wants to go to the amusement park! Amusement parks are so much fun!
Doctor: There are lots of fruit shops in the amusement park! For example, there are big, sweet watermelons there. Is there any fruit that Dudu particularly wants to eat?
Child: Banana. (<i>Relevant Response</i>)
Doctor: Dudu, that's great! You like bananas, that's wonderful! You're a little fruit expert!
Doctor: Besides bananas, what other fruits does Dudu like?

Figure 8: Case Study in Dialogue Synthesis with Topic "Buy fruit" and the same Turns. The left side shows a real intervention dialogue, while the right side shows a synthetic intervention dialogue. Blue indicates Instruction, green denotes Assistance including Half-Assistance and Full-Assistance, yellow signifies Reinforcement and Acknowledgement, and red represents the child's responses.

B.2 Case Study in Real Autism Intervention

Figure 9 illustrates a case that the different responses of different models to real autism clinical interventions and how DOCTORAGENT performs O-T-A-C.

B.3 Case Study in Comparison to ToT

To intuitively illustrate the clinical limitations of the ToT baseline, Figure 10 presents a qualitative case study where the child provides an irrelevant response (“Daytime”) to a weather-related instruction. According to ABA protocols, the doctor should first acknowledge the child’s response before providing further assistance.

Through this case study, we observe that while both our framework and ToT generate responses that subjectively resemble the natural conversational tone of real doctors, the ToT generation suffers from two critical clinical flaws:

- **Strategy Deviation:** During the initial acceptance phase (*Other*), ToT inappropriately appends a new instruction (“So what is the weather like during the day?”). Acceptance should strictly acknowledge the child without immediately demanding a new cognitive task. This flaw precisely demonstrates the necessity of our *Correct* module. While the foundational generation in our framework might occasionally make similar instruction-stacking errors, the O-T-A-C loop effectively detects and filters them out, ensuring the atomicity of the strategy.
- **Topic Deviation:** The doctor’s initial target concept was “weather.” However, when executing the *Half-Assistance* intervention, the ToT response drifts from the core topic, instead asking an open-ended, vague question (“Do you think there’s anything else in the sky?”). Such topic deviation easily distracts ASD children and strictly violates the precise targeting requirements of clinical EIBI interventions. Conversely, our framework generates a highly engaging and assistive prompt (“sunny or like it’s going to rain”) that safely guides the child back to the intended topic.

B.4 Case Study in Correct Module

To intuitively illustrate the impact of the *Correct* module, Figure 11 presents two typical intervention scenarios. In the **Strategy Deviation** case

(Figure 11, left), the child provides a seemingly irrelevant but tangentially related response (“Daytime”). While the doctor appropriately attempts to acknowledge this, the uncorrected generation improperly appends a new instruction (“What do you think of the weather during the day?”) during the acceptance phase. The *Correct* module successfully excises this excessive topic matching, ensuring the response aligns with the intended strategy without causing topic drift.

Furthermore, the **Instruction Stacking** case (Figure 11, right) highlights a critical clinical safety mechanism. Before correction, the agent stacks multiple complex questions into a single conversational turn. Given the severe social and cognitive communication difficulties faced by autistic children, ABA intervention protocols strictly prohibit delivering complex or consecutive instructions, which can easily cause cognitive overload and break the therapeutic loop. The *Correct* module effectively filters out the redundant questions, streamlining the utterance into a single, atomic instruction that is clinically safe and manageable for the child to process.

C ABA Strategy and Response Type

C.1 ABA Strategy

ABA is a structured approach commonly used as a behavioral therapy in treating autism. Specifically, doctors integrate Discrete Trial Teaching (DTT) and Natural Environment Teaching (NET) methods from ABA to intervene with autistic children. The doctor’s strategies are categorized as: *Instruction*, *Reinforcement*, *Half-Assistance*, *Full-Assistance*, *Other* and *Pause*. The child’s response types are categorized as: *Irrelevant*, *Relevant*, *Repetitive*, and *Unresponsive*.

Instruction 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.

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 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

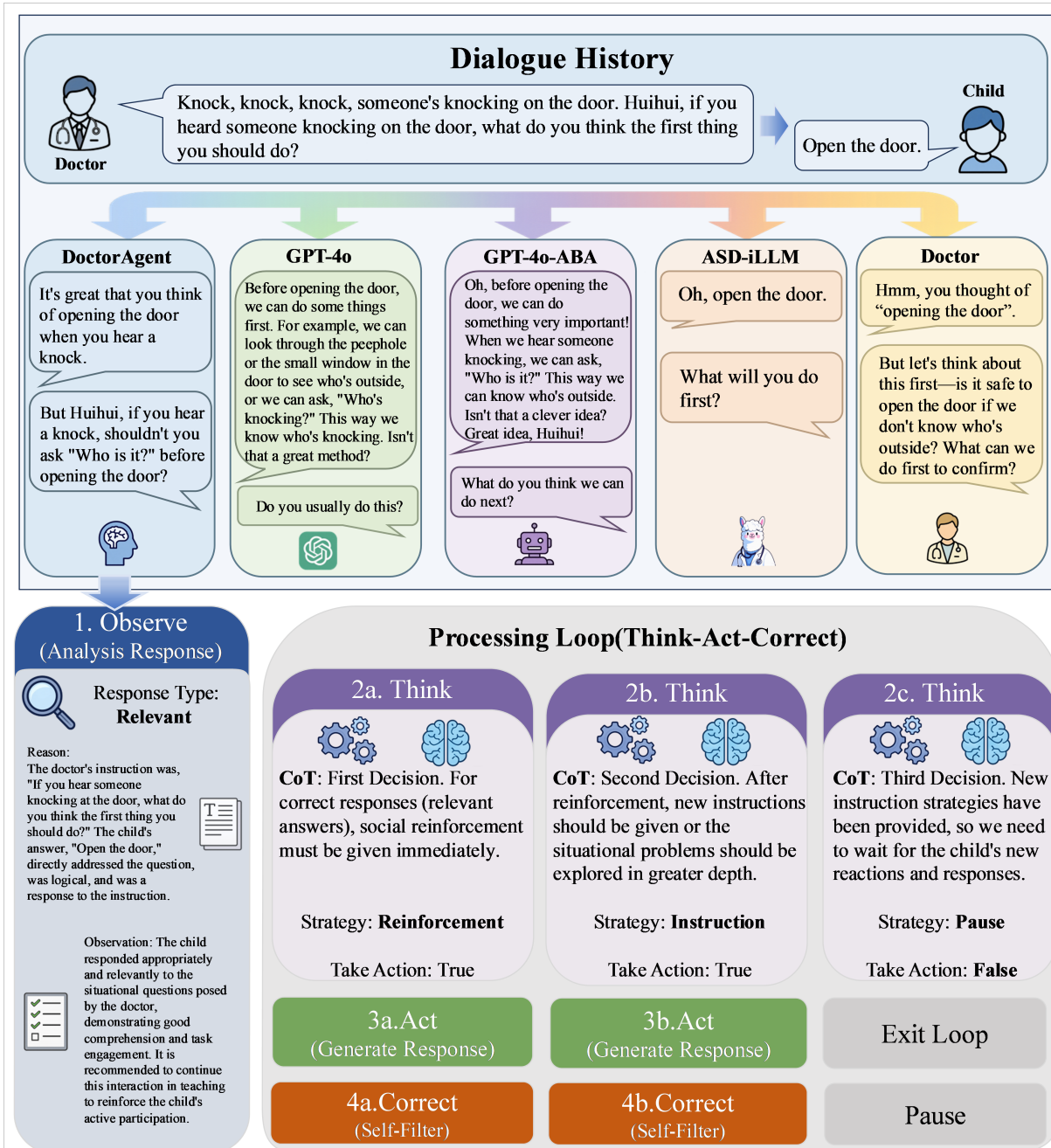
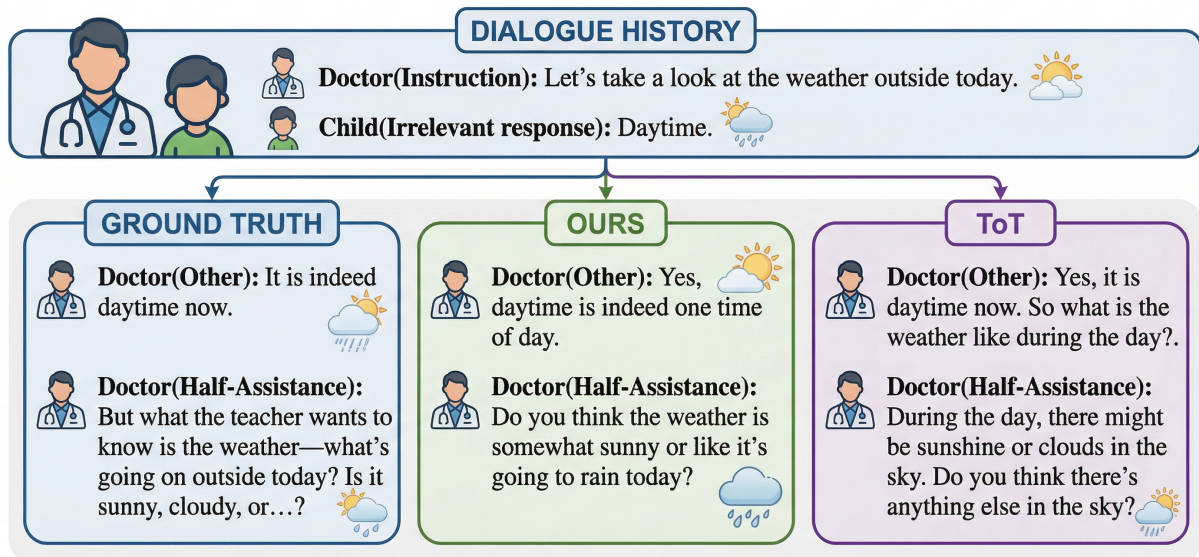


Figure 9: Case Study in Real Autism Intervention. The diagram above illustrates the intervention responses of DOCTORAGENT, the real Doctor, and other models based on a realistic intervention dialogue. The diagram below shows how DOCTORAGENT completes the O-A-T-C process.



Analysis: 'Ours' generates more engaging and assistive responses, guiding the child back to the intended topic.

Figure 10: Qualitative comparison between our O-T-A-C framework and the Tree-of-Thought (ToT) baseline. While ToT exhibits conversational fluency, it critically fails in clinical adherence by introducing strategy deviation (instruction stacking) and topic drift. Our framework safely guides the child back to the intended topic.

socialization through verbal praise and empathy.

Assistance refers to the support provided by therapists when autistic children have difficulty responding. This support can take the form of physical, visual, or verbal prompts. Assistance helps children build confidence, reduce frustration, and gradually understand the meaning of instructions. Assistance needs to be timely and appropriate to avoid causing feelings of failure in the child or creating dependence on the prompts. In thematic conversation intervention, Assistance usually takes the form of verbal prompts, such as rephrasing questions, breaking down questions, or providing hints to the answer.

Assistance can be further categorized into Half-Assistance and Full-Assistance.

Half-Assistance refers to providing limited hints, such as keyword reminders, selective prompts, or guiding questions, when the child already has some understanding or a tendency to respond, helping the child complete the response based on their existing understanding.

Full-Assistance, on the other hand, involves the therapist directly providing clear demonstrations or complete answers when the child cannot understand the instructions or shows no response, guiding the child to imitate or repeat the correct response. By flexibly using partial and Full-Assistance at different stages, therapists can ensure the success rate of the intervention while gradually

improving the child's independent response ability.

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

C.2 Child Response Type

Relevant responses refer to children's answers that semantically or functionally match the instructions or questions given by the doctor, indicating that the child understands the current topic and can respond appropriately;

Irrelevant responses refer to children's answers that have no clear connection to the current instructions or topic, possibly reflecting attention shifts, comprehension difficulties, or language organization problems;

Repetitive responses refer to children simply repeating the doctor's words or their own previous expressions without providing new information or independent responses, usually reflecting imitative behavior or limitations in response strategies;

Unresponsive responses refers to the child not giving any verbal or non-verbal response within a reasonable waiting time, which may be related to comprehension difficulties, avoidance behavior, or emotional state.

D Details for ASDAgent-Dataset

Currently, there are no publicly available datasets for ASD dialogue intervention. Therefore, we cre-

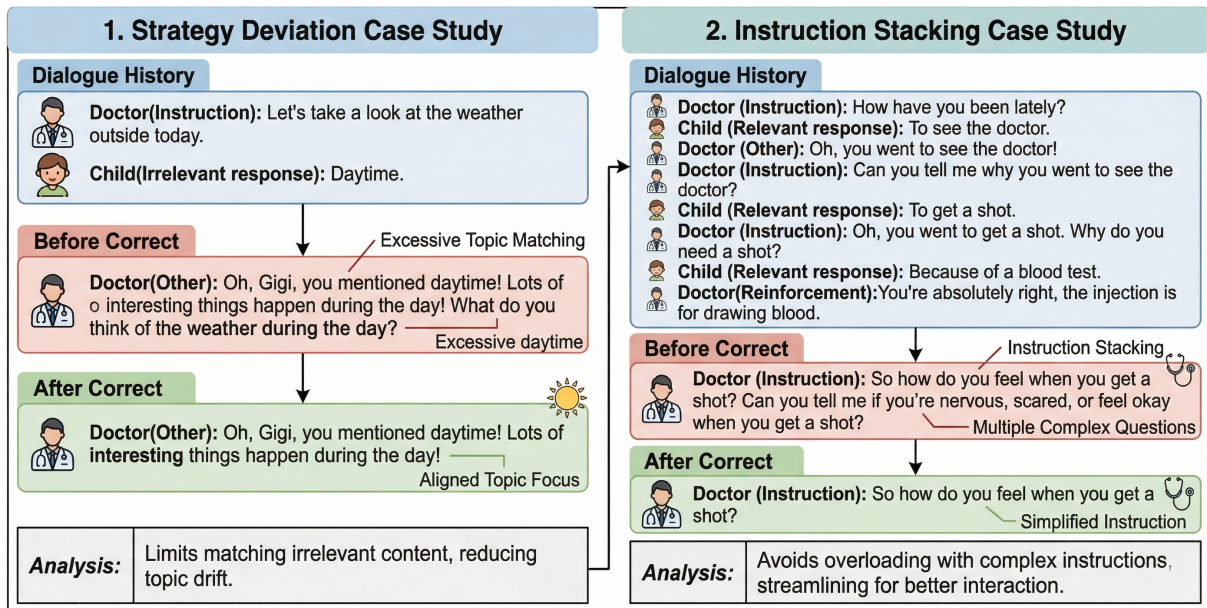


Figure 11: Qualitative case studies demonstrating the efficacy of the *Correct* module. The module actively rectifies strategy deviation by preventing topic drift (left) and resolves instruction stacking by streamlining complex, overloaded questions into clinically appropriate atomic instructions (right).

ated a multi-turn dialogue dataset for interventions between doctors and children with ASD, named **ASDAgent-Dataset**.

D.1 Data Collection

To ensure the authenticity and quality of the data, we collaborated with five treatment centers for autistic children after obtaining ethical approval. With full informed consent from both parents and children, audio recordings were collected during topic-based dialogue interventions using a portable recording device (H1-Pro, iFlytek Inc., China). To ensure clear audio capture, the recorder was placed in the chest pocket of the doctor’s coat.

Given that autistic children often experience delays in language development, chronological age does not necessarily reflect actual language ability. Therefore, only children with a language developmental age greater than 24 months were included in the study. Previous studies have shown that topic-based dialogue interventions can effectively alleviate social impairments in autistic children (Dekker et al., 2019; Hanrahan et al., 2020; van der Wilt et al., 2022). Accordingly, all recordings were conducted in the form of structured topic dialogues, with each recording focusing on a single predefined topic. All audio recordings were sampled at 16,000 Hz and stored in WAV format.

D.2 Data Processing

We employed a three-stage processing method to transcribe the original audio recordings into multi-dialogue text and annotate the doctors’ strategies and the children’s response types.

Automatic Transcription First, we utilized existing automated transcription tools SEED-ASR(Bai et al., 2024) to convert the original recordings into multi-turn dialogues.

Manual Transcription Our goal is to improve the quality of multi-turn dialogue text through manual transcription. Building upon (Lai et al., 2025), we annotated the data using crowdsourcing. Details about crowdsourcing can be found in the Appendix D.5.

State Annotation According to the ABA(Foxx, 2008; Roane et al., 2016), we performed more detailed data annotation on the selected high-quality dialogues, including annotating the doctor’s strategies and the child’s response types using ABA and DTT. The basic flow of DTT is illustrated in Figure 12.

D.3 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

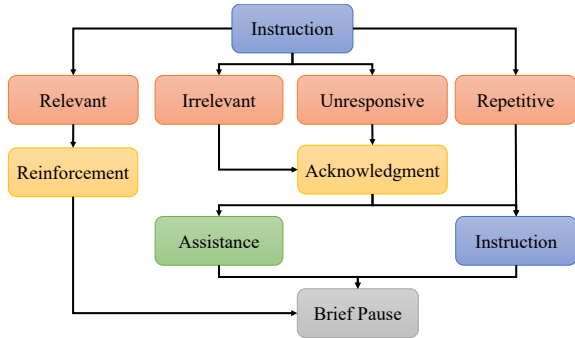


Figure 12: The standard workflow of Discrete Trial Training (DTT) derived from ABA literature (PACKET), illustrating the structured interaction cycle. Doctors can adjust their treatment strategies as needed, based on the actual intervention situation.

too few exchanges fail to reflect the doctor’s intervention strategies adequately.

- Dialogues focused on entities, such as story-books 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 the model’s dialogue style and intervention strategies.
- 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.

D.4 ASDAgent-Dataset

Golden We transcribed 2071 instances of multi-turn dialogues on various topics. After data cleaning, we obtained 764 high-quality, authentic multi-turn dialogues from 83 children with ASD, which we denote as \mathcal{D}_{golden} .

Silver Intervention dialogue synthesized denoted as \mathcal{D}_{silver} with the same quantity as \mathcal{D}_{golden} .

D.5 Manual Annotation Process

We recruited a total of 31 volunteers from the school, including 18 females and 13 males, to participate in the manual transcription and verification of the data. We provided compensation based on the amount of transcription work completed. The results and costs of the manual transcription are shown in the Table 6.

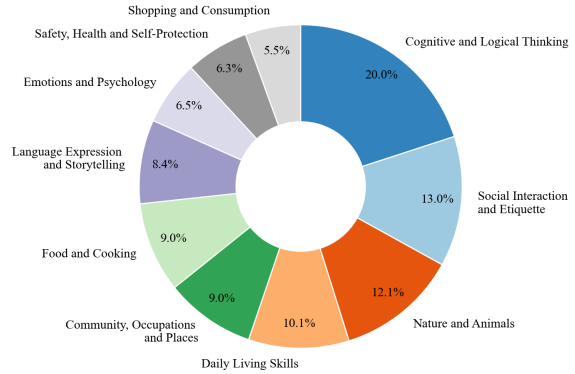


Figure 13: Topic distribution of ASDAgent-Dataset-Golden.

Table 6: Overview of Dialogue Transcription Cost (USD)

Item	Value
Total number of dialogues	2071
High-quality dialogues	764
Total transcription cost (USD)	5,204.17
Average cost per dialogue (USD)	2.51

Manual transcription is relatively expensive. The total manual transcription cost amounted to approximately 5,204 USD, with an average cost of 2.51 USD per dialogue.

D.6 Topic Classification

The topic distribution of ASDAgent-Dataset-Golden \mathcal{D}_{golden} is illustrated in Figure 13, showing a balanced distribution of topics.

In classifying dialogue topics, we consider not only the semantics of the dialogue topic but also how doctors actually utilize these topics to intervene with children during real-world conversations. We refer to this as the macro topic.

We computed embeddings using Qwen3-Embedding-0.6B (Zhang et al., 2025) for all macro topics, performed hierarchical clustering, and then manually refined the results to obtain the final 10 topic categories as shown in Figure 14.

D.7 Children’s statistics

Demographic Details The demographic information of children in ASDAgent-Dataset-Golden is presented in Table 9, indicating 65 boys and 18 girls. There are minimal differences in both the mean and variance of age between genders, with the sample centered around five years of chrono-

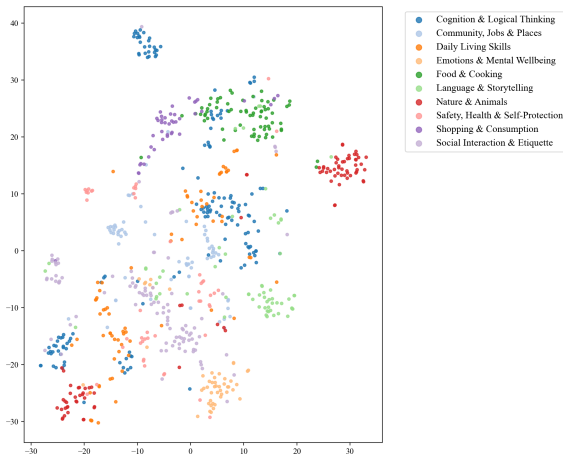


Figure 14: t-SNE scatter plot of macro topics across 10 main conversational categories.

logical age. In contrast, language developmental age is substantially lower than chronological age, averaging approximately three to four years, which is consistent with the characteristic language delays observed in autistic children.

Child Response Details Information We calculated the percentage of different types of responses in children under different doctors' treatment strategies in Table 7. We found that strategy–response transition probabilities reveal clear behavioral patterns. Reinforcement produces the highest rate of relevant child responses (64.11%) and the lowest no-response rate, indicating strong engagement. Instruction increases relevant responses but also no-response risk. Full assistance reduces silence but induces repetition, while partial assistance offers a balanced trade-off consistent with ABA principles.

Furthermore, we calculated the probability of children responding when the doctor used non-directive strategies as shown in Table 8, which indicates that even when explicit instructions are not issued, ASD intervention dialogues remain predominantly doctor-led, with clinicians frequently providing follow-up guidance, reinforcement, or corrective feedback. The relatively low child-after probability is consistent with clinical observations of ASD interactions, where spontaneous child initiation is limited and structured scaffolding is often required. Importantly, this asymmetry complements the strategy–response transition patterns, highlighting the necessity of sequential doctor interventions to maintain effective teaching dynamics.

D.8 ASD Children Heterogeneity

Based on the behavioral profiles exhibited by different children as reflected in their performance on the Table 7 and 8, we have categorized the children into the following four types:

- **Compliant:** High response rate to instructions, or very high response rate after assistance, with a low interruption rate.
- **Impulsive:** Significantly higher interruption rate (usually > 0.14), or exhibiting a higher tendency for irrelevant responses/interruptions during the instruction phase.
- **Difficult:** Low response rate to instructions, and poor response to assistance (no response or irrelevant response).
- **Prompt Dependent:** Average response rate to instructions, but full or partial assistance significantly improves accuracy.

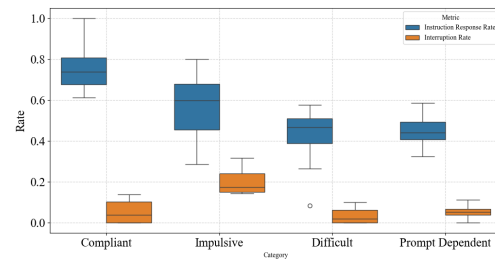


Figure 15: Distribution of Key Metrics by Child Category

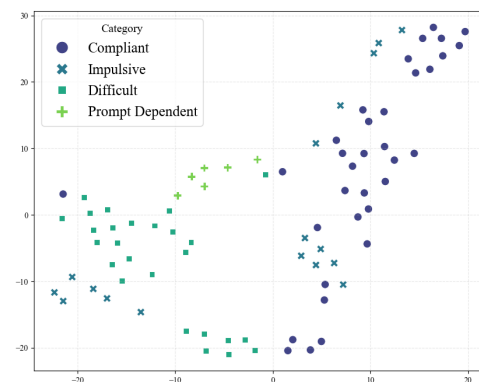


Figure 16: t-SNE Clustering of Child Profiles

The box plot and scatter plot are shown in the Figure 15 and 16, which also provides a basis for

Table 7: Conditional probabilities (%) of child response types given the last doctor intervention strategy.

Doctor Strategy	Relevant	Irrelevant	Unresponsive	Repetition
Instruction	59.71	23.39	14.35	2.55
Full-Assistance	53.39	21.71	9.56	15.34
Half-Assistance	53.99	26.14	14.45	5.42
Reinforcement	64.11	20.21	3.14	12.54
Other	52.24	22.39	7.46	17.91

Table 8: Turn interruption probabilities following the current dialogue turn.

Next Speaker	Probability (%)
Child	8.67
Doctor	91.33

ChildAgent to adapt to personalized persona modeling.

D.9 Utterance Length

Statistical information for the ASDAgent-Dataset-Golden is shown in Table 10 and 11. On average, each conversational turn lasts 18.61 rounds. Furthermore, during the intervention, both the doctor and the child used relatively few characters per utterance, with the doctor averaging 22.35 characters and the child averaging only 5.52 characters. The doctor needed to use concise and easy-to-understand sentences to encourage the child’s participation, while the child’s language developmental delay and social difficulties significantly reduced their response frequency and vocabulary.

D.10 Conversation Length Distribution Modeling

To ensure that the synthetic sessions reflect the engagement patterns of real-world clinical interventions, we do not set a fixed dialogue length. Instead, we model the session duration (number of turns) based on the statistical distribution derived from the real-world dataset \mathcal{D}_{golden} .

Observing that clinical conversation lengths typically follow a heavy-tailed distribution shown in Figure 17, we fit a Log-Normal Distribution to the turn counts of the 50 real sessions. Let $\mathcal{L}_{golden} = \{l_1, l_2, \dots, l_N\}$ be the set of turn counts in \mathcal{D}_{golden} . We estimate the parameters μ and σ of the underlying normal distribution using Maximum Likelihood Estimation (MLE):

$$\mu = \frac{1}{N} \sum_{i=1}^N \ln(l_i), \quad \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (\ln(l_i) - \mu)^2} \quad (11)$$

For each synthetic session, we sample a raw length L_{raw} from this distribution:

$$L_{raw} \sim \text{LogNormal}(\mu, \sigma) \quad (12)$$

To adhere to the context window constraints of LLMs and ensure meaningful interactions, we apply a clipping function to determine the final synthetic length L_{syn} :

$$L_{syn} = \text{Clip}(\text{Round}(L_{raw}), L_{min}, L_{max}) \quad (13)$$

where we set $L_{min} = 5$ and $L_{max} = 50$ based on our pilot study. This approach ensures that the synthetic dataset retains the natural variability of human interactions while maintaining computational feasibility.

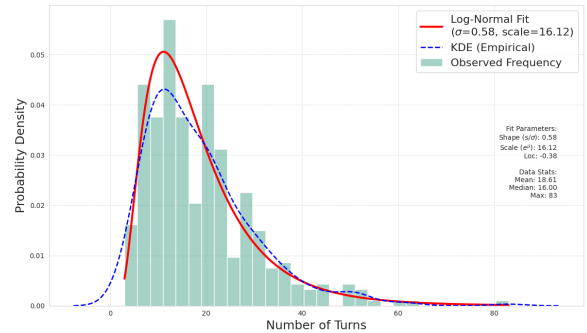


Figure 17: Distribution of Conversation Turns (Log-Normal Fit)

E Detailed Experiment Instructions

To rigorously quantify the benefits of our personalized persona modeling, we construct a baseline BaseChild(GPT-4o). Unlike our proposed CHILDAAGENT which dynamically interpolates between personal and global statistics ($\alpha = 0.3$), the

Gender	Number	Age (Mean \pm std)	Language Development Age (Mean \pm std)
Male	65	5.35 \pm 1.26	3.74 \pm 1.16
Female	18	5.42 \pm 1.33	3.87 \pm 1.13

Table 9: The demographic details of children for ASDAgent-Dataset-Golden.

Table 10: Dialogue Basic Statistics

Metric	Length
Turns per Dialogue	18.61 \pm 11.39
Chars per Doctor Utterance	22.35 \pm 12.20
Chars per Child Utterance	5.52 \pm 7.61

Table 11: Utterance Length Statistics by Strategy and Response Type

Category	Subtype	Length
Doctor	Instruction	20.77 \pm 12.01
	Reinforcement	22.13 \pm 11.67
	Half-Assistance	27.67 \pm 12.03
	Full-Assistance	28.08 \pm 13.63
	Other	20.08 \pm 10.37
Child	Relevant	6.56 \pm 8.80
	Irrelevant	6.34 \pm 5.64
	Repetitive	4.28 \pm 2.01
	Unresponsive	0.00 \pm 0.00

BaseChild relies exclusively on the Global Population Prior ($\alpha = 1.0$).

In addition, We note that under the common prompting settings, models do not explicitly output intervention strategy labels. To ensure fair comparison in strategy-level evaluation, we therefore perform a secondary annotation process. Specifically, for each generated doctor utterance, the corresponding intervention strategy is inferred and labeled by an GPT-4o following the same strategy taxonomy used for DOCTORAGENT outputs. We further manually inspected a random subset of annotated samples to verify annotation consistency. The prompt can be found in the Appendix J.2.

In terms of assessing Data efficacy, We used the fine-tuning framework TRL (Wolf et al., 2020; von Werra et al., 2020) for training SLMs on ASDAgent-Dataset via LoRA method(Hu et al., 2022), utilizing 1 RTX 4090 GPU. 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.

F Details for O-T-A-C Loop

F.1 Computational Complexity

To comprehensively evaluate the methodological rigor of our work, we provide an analysis of the computational complexity and resource requirements associated with the ASDAGENT framework. As shown in Table 12, the explicit Observe-Think-Act-Correct (O-T-A-C) loop introduces a certain delay during the generation process, primarily driven by the iterative ‘‘Think’’ module.

The cumulative processing time results in an average delay of approximately 11.88 to 17.58 seconds per conversational turn, depending on the capacity of the backbone model. We acknowledge that while this explicit reasoning mechanism guarantees high clinical fidelity and strategy adherence, this latency may have some impact on the pacing of real-world clinical interventions.

G Evaluation Metrics

G.1 Automatic Evaluation

In aspects of assessing the diversity of text, We used common automatic evaluation metrics including Self-BLEU(Zhu et al., 2018), Self-GLEU(Yoon et al., 2023) and Self-BERTScore(Zhang et al., 2024). These self-referential metrics measure the average similarity among generated samples, where lower scores indicate higher diversity. At the same time, we introduced the Distinct-n(Li et al., 2016) metric to measure the vocabulary richness and expressive diversity of the model’s output.

In the context, we believe that stylistic similarity is reflected in two aspects: word choice and sentence semantics. First, regarding word choice, different contexts require different words. For example, informal social occasions usually use more colloquial expressions, while communication with autistic children should be as concise and easy to understand as possible. Therefore, we used several word overlap metrics, such as BLEU (Papineni et al., 2002), GLEU (Wu et al., 2016), and METEOR (Lavie and Agarwal, 2007), to evaluate the word-level matching. Second, at the semantic and sentence level, our goal is to make the model’s

Model	Metric	Observe	Think	Act	Correct	Overall Total
GPT-4o-mini	Total Time (s)	3.65 \pm 1.58	11.86 \pm 3.54	3.66 \pm 4.01	4.72 \pm 12.72	23.89 \pm 13.84
	Avg Time (s)	3.65 \pm 1.58	4.20 \pm 1.10	1.83 \pm 1.93	2.20 \pm 4.21	11.88 \pm 5.02
GPT-4o	Total Time (s)	5.39 \pm 5.13	19.29 \pm 13.99	5.38 \pm 10.30	4.41 \pm 7.19	34.47 \pm 19.49
	Avg Time (s)	5.39 \pm 5.13	6.81 \pm 4.78	2.93 \pm 5.56	2.45 \pm 4.72	17.58 \pm 10.12

Table 12: Computational complexity analysis of the O-T-A-C reasoning loop. The table reports the total time and average time per step (in seconds) for a single dialogue turn using different backbone models.

output semantically similar to real dialogues, thus achieving intervention effects similar to those of clinicians. Therefore, we chose BertScore (Zhang et al., 2020), Qwen-Embedding (Zhang et al., 2025) and BGE-M3 (Chen et al., 2024) to measure the semantic similarity of the model’s output.

In addition, to measure the alignment between the empirical distribution of ABA strategies used by human doctors (P) and the synthetic distribution generated by ASDAGENT (Q).

- **Kullback-Leibler (KL) Divergence:** Defined as $D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$. It is an asymmetric measure of how one probability distribution differs from a reference distribution. In our study, it quantifies the ‘strategy drift’. A low KL divergence means the Agent rarely chooses strategies that human doctors would consider low-probability.
- **Jensen-Shannon (JS) Divergence:** Defined as $D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)$, where $M = \frac{1}{2}(P + Q)$. Unlike KL, JS is symmetric and bounded $[0, 1]$. It provides a stable metric of similarity between the two strategy portfolios. A $D_{JS}(P||Q)$ of 0 indicates identical strategy usage frequencies, validating the high fidelity of our synthetic clinical data.

Finally, at the level of physician strategy use, our goal is to evaluate whether the model’s behavior in selecting intervention strategies can be as close as possible to the strategy distribution and usage patterns in real clinical dialogues. Unlike sentence generation, the focus of strategy prediction is not on the text content itself, but on whether the model selects the appropriate intervention strategy at the appropriate time. Therefore, we evaluated the model’s output from two perspectives: overall consistency of strategy use and temporal consistency of the strategy sequence. The calculation of metrics for overall consistency of strategy use and

temporal consistency of the strategy sequence can be found in the Appendix G.4 and G.5.

G.2 Human Evaluation

After discussing with doctors, we had doctors evaluate the performance of the intervention dialogues generated by ASDAgent and real dialogues on the same topics in the test set. This evaluation was based on 11 dimensions across 3 aspects, detailed in the table 14. Each dimension used a scoring system from 0 to 4, with higher scores indicating better quality output from the physician. We invited two experienced autism clinical intervention physicians to conduct the evaluation.

During the annotation process, the doctors focused on the scoring criteria for each teaching segment. A segment refers to a complete cycle in DTT (Discrete Trial Training), as shown in Figure 12. They needed to break down the entire dialogue into multiple segments to evaluate the application of ABA principles, language use, and safety in each segment. Based on the overall assessment, they assigned scores from 0 to 4 according to the following criteria:

- **0:** The doctor’s performance in the dialogue segment was entirely inappropriate.
- **1:** A small portion of the doctor’s performance in the dialogue segment was appropriate.
- **2:** Part of the doctor’s performance in the dialogue segment was appropriate.
- **3:** Most of the doctor’s performance in the dialogue segment was appropriate.
- **4:** All of the doctor’s performance in the dialogue segment was appropriate.

Table 13 presents detailed information about two invited experts for human evaluation, each with more than five years of experience in autism treatment. Their extensive intervention experience and knowledge make them well-qualified for the professional evaluation task.

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

Table 13: Information for experts involved in human evaluation.

G.3 LLM Evaluation

Given the high cost and subjectivity of expert annotation in ASD intervention scenarios, LLM-as-a-Judge provides a scalable and consistent alternative for evaluating at scale. We adopt the LLM-as-a-Judge paradigm (Zheng et al., 2023) to evaluate Topic diversity, Quality of dialogue synthesis and Clinical intervention effect. Specifically, LLM-based evaluation employed the Verbal Behavior Milestones Assessment and Placement Project (VB-MAPP) and Discrete Trial Training (DTT) guidelines. In addition, the checklist was co-developed and validated by two physicians who actually conduct clinical interventions for autism involved in this study to ensure they reflect real-world therapeutic priorities (e.g., Safety, Strategy Adherence). Physician information can be found in the Table 13.

We choose DeepSeek-v3.2(Liu et al., 2025), Gemini-2.5-pro(Comanici et al., 2025) and GPT-5.1(OpenAI, 2025) as LLM evaluators.

Table 14, 15 and Figure 32, 33, 34 show the evaluation criteria and prompts in Evaluation 1 and Evaluation 2, respectively.

G.4 Multiset PRF

Multiset-based strategy coverage ignores the order in which strategies appear, focusing only on whether the types and quantities of predicted strategies match the reference. This is used to measure whether the doctor selected the key strategies, without requiring the order of strategy selection to be exactly the same.

Let S_{ref} be the reference strategy sequence (Ground Truth), S_{pred} be the predicted strategy sequence, $C(x, S)$ be the number of times strategy x appears in sequence S , V be the vocabulary of all possible strategies, and $|S|$ denote the total length of the sequence.

First, we calculate the overlap count, which is the size of the intersection of the two multisets:

$$\text{Overlap}_{\text{set}} = \sum_{x \in V} \min(C(x, S_{pred}), C(x, S_{ref})) \quad (14)$$

Based on this, calculate Precision, Recall, and F1:

$$\text{Precision}_{\text{set}} = \frac{\text{Overlap}_{\text{set}}}{|S_{pred}|} \quad (15)$$

$$\text{Recall}_{\text{set}} = \frac{\text{Overlap}_{\text{set}}}{|S_{ref}|} \quad (16)$$

$$\text{F1}_{\text{set}} = \frac{2 \cdot \text{Precision}_{\text{set}} \cdot \text{Recall}_{\text{set}}}{\text{Precision}_{\text{set}} + \text{Recall}_{\text{set}}} \quad (17)$$

G.5 LCS PRF

The strategy coverage based on the Longest Common Subsequence (LCS) strictly considers the relative order in which strategies appear. This is used to measure whether the doctor selected the correct and crucial strategies in the correct order. If the model predicts the correct strategies but the order is completely wrong, this metric will be low.

Let $\text{LCS}(A, B)$ be the Longest Common Subsequence of sequences A and B , and $|\text{LCS}(A, B)|$ be the length of this subsequence.

First, calculate the match length:

$$\text{Match}_{\text{seq}} = |\text{LCS}(S_{pred}, S_{ref})| \quad (18)$$

Based on this, calculate the ordered Precision, Recall, and F1 score:

$$\text{Precision}_{\text{seq}} = \frac{\text{Match}_{\text{seq}}}{|S_{pred}|} \quad (19)$$

$$\text{Recall}_{\text{seq}} = \frac{\text{Match}_{\text{seq}}}{|S_{ref}|} \quad (20)$$

$$\text{F1}_{\text{seq}} = \frac{2 \cdot \text{Precision}_{\text{seq}} \cdot \text{Recall}_{\text{seq}}}{\text{Precision}_{\text{seq}} + \text{Recall}_{\text{seq}}} \quad (21)$$

H Meta-Evaluation: Human-LLM Alignment

To validate the reliability of automated evaluation, we calculated the agreement and correlation between three LLM judges (DeepSeek-V3.2, GPT-5.1, Gemini-2.5) and human experts on a subset of 46 randomly sampled dialogues.

Dimension	Category	Explanation
Professionalism	Principle	Dialogues adhere to the DTT method or NET approach outlined.
	Instruction	Doctor provides clear, unambiguous instructions to the child.
	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.
	Acknowledgment	Doctor avoids criticism or negative reinforcement when the child gives incorrect responses, shows no response, or refuses, and instead adopts an accepting, natural response style.
	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	Privacy	The Child’s privacy is strictly protected during the dialogue.
	Content	Dialogues avoid harmful content for children.

Table 14: The evaluation criteria for Dialogue Synthesis and Clinical Intervention Effect, 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.

Dimension	Category	Explanation
Doctor (A)	Dialogue Principles (A1)	Whether the dialogue follows ABA-based instructional paradigms, such as Discrete Trial Training (DTT) or Natural Environment Teaching (NET).
	ABA Strategy Sequencing (A2)	Whether appropriate ABA strategies are applied in a progressive and coherent order (e.g., reinforcement before instruction, acceptance followed by partial or full assistance, rather than disordered sequencing).
	Personalization (A3)	Whether the doctor adapts questioning style, linguistic complexity, or pacing according to the child’s specific responses and needs.
Child (B)	ASD-consistent Response (B1)	Whether the child’s responses exhibit realistic ASD characteristics, such as non-compliance, repetitive behaviors, or language impairments.
Interaction (C)	Scenario Complexity (C1)	Whether the dialogue contains effective instructional dynamics, such as corrective teaching loops or meaningful pedagogical interactions.

Table 15: Evaluation criteria for Dialogue Synthesis in ablation study. The assessment covers three dimensions—Professionalism (A), Child Realism (B), and Scenario Quality (C)—with corresponding sub-categories used in both human and LLM-based evaluations.

Table 16: Meta-Evaluation Results: Alignment between LLM Judges and Human Experts.

Model	Accuracy	Cohen’s κ	Spearman’s ρ	Kendall’s τ
DeepSeek-V3.2	0.522	0.288	0.396	0.381
GPT-5.1	0.283	-0.024	-0.144	-0.136
Gemini-2.5-Pro	0.261	-0.048	0.024	0.023

As shown in Table 16, although LLM and human assessments are not entirely consistent, this confirms the high fidelity of our synthetic data:

DeepSeek-V3.2 as the Most Reliable Judge: Among the candidates, DeepSeek-V3.2 achieved the highest agreement with human experts (Accuracy: 52.2%, $\kappa = 0.288$), identifying the superiority of real data in 21.7% of cases while maintaining a moderate correlation ($\rho = 0.40$). This indicates its capability to capture clinical nuances.

The "Tie Bias" Phenomenon: Conversely, GPT-5.1 and Gemini-2.5 exhibited a near-total inability to distinguish synthetic from real data, predicting "Tie" in 80.4% and 84.8% of cases, respectively. This resulted in near-zero or negative Kappa scores.

Validation of Synthesis Quality: While this limits the utility of GPT/Gemini as discriminators, it paradoxically validates the high fidelity of our synthetic data. The generated dialogues are sufficiently natural and strategic to render them indistinguishable from human therapist outputs for general-purpose SOTA models.

I Ablation Study

I.1 ASDAgent for Data Synthesis

Automatic Evaluation. Table 17 shows the diversity of language used by children and doctors in the dialogue; Table 18 and Table 19 shows the proportion of strategies or response types used by children and doctors in the dialogue. Table 20 and Table 21 shows the average length of responses from children and doctors in the dialogue.

Removing DOCTORAGENT reveals significant strategy collapse and linguistic abnormalities, excessively high proportion of instructions and abnormal sentence length. Removing CHILDAAGENT, while showing better performance on some diversity metrics for DOCTORAGENT, reveals a deviation from reality in its strategy distribution (insufficient reinforcement), and children tend to produce excessively long and irrelevant/repetitive responses. From the perspective of "rationality of intervention behavior," it is less stable than ASDAGENT. Therefore, in the Evaluation 2, we believe that ASDAGENT best reproduces realistic clinical interaction patterns and is the most suitable source of high-quality synthetic dialogues.

LLM Evaluation. Additionally, we conduct an ablation study using LLM-based evaluators to investigate the relative contributions of doctor mod-

eling and child modeling to intervention dialogue quality of ASDAgent according to Table 15.

Table 22 presents ablation results across three LLM evaluators. Removing the CHILDAAGENT consistently causes substantial degradation in professionalism (A), with relative drops of 19.8%–26.9%. This decline is mainly attributed to reduced adherence to DTT/NET dialogue principles (A1) and less coherent ABA strategy sequencing (A2), as well as weaker personalized adjustments (A3) to child responses. These results highlight the necessity of child-aware modeling for clinically appropriate interventions.

Removing the DOCTORAGENT also leads to notable performance drops, particularly in professionalism (A) and scenario complexity (C), indicating impaired instructional structure and reduced use of effective teaching dynamics (e.g., corrective loops). In contrast, child realism (B) exhibits smaller changes and occasionally improves, suggesting that surface-level linguistic plausibility alone is insufficient to ensure intervention quality. Overall, the consistent decline in Total score confirms the complementary importance of both doctor and child modeling.

I.2 ASDAgent for Clinical Intervention

Automatic Evaluation From Table 23, ABA and BASE achieve comparable performance on surface-level lexical metrics such as BLEU, GLEU, and METEOR, with BASE occasionally obtaining slightly higher n-gram scores. However, DOCTORAGENT consistently attains the highest semantic alignment and diversity, as reflected by superior BERTScore-F1 and markedly higher Distinct-2/3 scores. The BASE and ABA prompts can be found in J.1.

Table 24 reports an ablation study on strategy-level consistency. Results are evaluated using both multiset-based and LCS-based Precision/Recall/F1 metrics, capturing strategy alignment with and without order sensitivity.

Across both GPT-4o and GPT-4o-mini, DOCTORAGENT consistently achieves the highest precision, recall, and F1 scores, outperforming both ABA and BASE settings by a clear margin. The most prominent gains are observed in recall, which approaches 80%, indicating that DOCTORAGENT is able to cover a substantially larger portion of real clinical strategies. In contrast, ABA prompting yields only modest improvements over BASE, suggesting that prompt-level constraints alone are

Table 17: Diversity Metrics for Doctors and Children Across Different Sources

Doctor	Child	D-2 ↑	D-3 ↑	S-BLEU ↓	S-GLEU ↓	S-BERTScore ↓
Doctor Part						
Real	Real	0.348	0.637	0.549	0.282	0.586
DoctorAgent	ChildAgent	0.249	0.499	0.680	0.353	0.607
DoctorAgent	BaseChild	0.277	0.519	0.661	0.364	0.599
BaseDoctor	ChildAgent	0.187	0.412	0.731	0.318	0.612
Child Part						
Real	Real	0.499	0.732	0.423	0.483	0.607
DoctorAgent	ChildAgent	0.428	0.667	0.477	0.522	0.622
DoctorAgent	BaseChild	0.383	0.575	0.552	0.557	0.609
BaseDoctor	ChildAgent	0.400	0.599	0.540	0.531	0.615

Table 18: Distribution of Doctor and Child Interaction Strategies **Percentage (%)** with KL and JS Divergence to Real

Doctor	Child	Instru.	Reinfo.	Half-A.	Full-A.	Other	KL	JS
Real	Real	42.29	31.62	9.99	7.27	8.83	-	-
DoctorAgent	ChildAgent	33.68	26.85	19.54	3.80	16.13	0.083	0.019
DoctorAgent	BaseChild	27.89	13.43	36.36	6.30	16.01	0.325	0.072
BaseDoctor	ChildAgent	73.45	3.77	4.98	0.00	3.92	0.259	0.118

Table 19: Distribution of Child Response Types **Percentage (%)** with KL and JS Divergence to Real

Doctor	Child	Relev.	Irrele.	Unres.	Repet.	KL	JS
Real	Real	58.55	25.66	10.36	5.43	-	-
DoctorAgent	ChildAgent	53.72	27.40	11.43	7.44	0.007	0.002
DoctorAgent	BaseChild	47.90	26.22	15.56	10.31	0.039	0.009
BaseDoctor	ChildAgent	49.91	25.73	15.78	8.40	0.024	0.006

Table 20: Doctor Utterance Length by Intervention Strategy (Mean_{±Std})

Doctor	Child	Instru.	Reinfo.	Half-A.	Full-A.	Other
Real	Real	21.77 _{±11.29}	21.32 _{±11.39}	30.35 _{±10.57}	30.48 _{±16.01}	20.58 _{±11.07}
DoctorAgent	ChildAgent	27.94 _{±9.29}	25.95 _{±6.39}	31.87 _{±8.82}	35.53 _{±13.54}	25.44 _{±8.61}
DoctorAgent	BaseChild	22.47 _{±6.98}	18.84 _{±8.58}	23.38 _{±9.48}	24.64 _{±10.65}	20.28 _{±7.70}
BaseDoctor	ChildAgent	96.50 _{±39.55}	24.32 _{±13.01}	30.06 _{±16.64}	0.00 _{±0.00}	21.35 _{±9.14}

Table 21: Child Utterance Length by Response Type (Mean_{±Std})

Doctor	Child	Relev.	Irrele.	Unres.	Repet.
Real	Real	5.97 _{±5.41}	5.92 _{±3.52}	0.00 _{±0.00}	4.45 _{±2.84}
DoctorAgent	ChildAgent	4.61 _{±2.38}	7.95 _{±4.03}	0.00 _{±0.00}	5.98 _{±1.99}
DoctorAgent	BaseChild	5.11 _{±2.19}	11.46 _{±3.88}	0.00 _{±0.00}	10.12 _{±4.36}
BaseDoctor	ChildAgent	5.68 _{±2.81}	7.89 _{±2.68}	0.00 _{±0.00}	9.73 _{±3.72}

Table 22: Ablation Study across Different Evaluators. For ablated settings, A/B/C/Total report relative changes (%).

Evaluator	Source	A1	A2	A3	B1	C1	A	B	C	Total
DeepSeek-V3.2	Full	2.87	2.44	2.60	3.40	2.11	7.91	3.40	2.11	13.42
	w/oChildAgent	2.07	1.76	1.96	3.31	1.47	↓26.9%	↓2.6%	↓30.3%	↓21.3%
	w/oDoctorAgent	2.00	1.64	2.42	3.00	1.56	↓23.3%	↓11.8%	↓26.1%	↓20.9%
Gemini-2.5-Pro	Full	3.20	2.80	2.89	3.27	2.31	8.89	3.27	2.31	14.47
	w/oChildAgent	2.51	2.07	2.56	3.80	2.22	↓19.8%	↑16.2%	↓3.9%	↓9.1%
	w/oDoctorAgent	2.71	2.33	2.44	3.44	2.69	↓15.8%	↑5.2%	↑16.5%	↓5.9%
GPT-5.1	Full	2.49	2.20	2.40	2.73	2.22	7.09	2.73	2.22	12.04
	w/oChildAgent	2.58	2.22	2.56	3.38	2.29	↑3.8%	↑23.8%	↑3.2%	↑8.1%
	w/oDoctorAgent	2.24	1.67	2.33	2.76	1.91	↓12.0%	↓1.1%	↓14.0%	↓9.4%

insufficient to ensure faithful strategy usage.

Importantly, the consistency between multiset-based and LCS-based results indicates that DOCTORAGENT improves not only the selection of strategies but also their sequential organization. Overall, these findings demonstrate that explicit agent-based modeling is crucial for reproducing real ASD intervention strategies, beyond what can be achieved through prompt engineering alone.

LLM Evaluation As shown in Table 25, we further conduct an ablation study across different evaluators and backbone models (GPT-4o-mini and GPT-4o) to analyze the effects of prompting strategies and agent-based modeling.

Across all evaluators, ABA prompting consistently outperforms BASE prompting, indicating that explicit ABA-guided constraints improve intervention quality beyond generic instructions. More importantly, DOCTORAGENT further improves performance in most cases, especially under the DeepSeek-V3.2 evaluator, where GPT-4o with DOCTORAGENT achieves the highest total score. This suggests that explicit doctor-child role modeling provides benefits beyond prompt design alone.

Comparing backbone models, GPT-4o consistently surpasses GPT-4o-mini under the same setting, demonstrating the impact of model capacity. While evaluator preferences vary slightly (e.g., GPT-5.1 favoring ABA in some cases), the overall trend remains stable: structured prompting and agent-based modeling jointly contribute to higher-quality intervention dialogues.

J Prompt

J.1 Base and ABA prompt

Figure 18 and 19 show the prompt used in clinical intervention under BASE and ABA settings.

J.2 Prompt for Strategy Labeling

Figure 20 illustrates the system prompt utilized to construct the supervised training dataset for the DoctorAgent. To capture the nuanced timing of ABA interventions, the Large Language Model (LLM) is conditioned to act as a professional data annotator. The instruction enforces a strict "Segment-and-Classify" Workflow:

- **Semantic Segmentation:** The model decomposes the therapist’s response into sequential clauses or semantic units. A rigorous "Lossless Reconstruction" constraint is imposed, strictly prohibiting any modification to punctuation or whitespace to ensure the annotated data aligns perfectly with the original audio transcripts.
- **Strategy Mapping:** Each segmented clause is classified into one of five distinct ABA strategies (e.g., Reinforcement, Half-Assistance, Instruction).

J.3 Prompt for DoctorAgent: Observe

Figure 21 presents the system prompt designed for the Observation Module within the DoctorAgent. To emulate the keen observational skills of a human therapist, the LLM is conditioned to act as a professional ABA practitioner performing real-time analysis. The instruction enforces a "Multi-Dimensional State Inference" strategy, requiring the model to analyze the child’s response relative to the doctor’s instruction across three critical dimensions:

- **Response Classification:** The model must rigorously distinguish between Functional Communication (Related Response) and Echolalia (Repetition/Mechanical imitation), a distinction critical for assessing ASD communicative progress.

Table 23: Ablation Study on Lexical, Semantic, and Diversity Metrics. For each model, the best result under each metric is highlighted in bold.

Model	Setting	BLEU \uparrow	GLEU \uparrow	MET \uparrow	BERT \uparrow	BGE \uparrow	QwenEmb \uparrow	D-2 \uparrow	D-3 \uparrow
GPT-4o	ABA	0.091	0.143	0.377	0.882	0.756	0.746	0.914	0.954
	BASE	0.091	0.144	0.373	0.884	0.750	0.745	0.922	0.960
	DoctorAgent	0.083	0.142	0.345	0.886	0.738	0.725	0.945	0.981
GPT-4o-mini	ABA	0.094	0.148	0.374	0.885	0.752	0.747	0.912	0.956
	BASE	0.093	0.151	0.357	0.888	0.747	0.742	0.922	0.962
	DoctorAgent	0.074	0.131	0.332	0.881	0.728	0.717	0.925	0.966

Table 24: Ablation Study on Strategy Consistency Metrics (in %). For each model, the best result under each metric is highlighted in bold.

Model	Setting	Multi-P \uparrow	Multi-R \uparrow	Multi-F1 \uparrow	LCS-P \uparrow	LCS-R \uparrow	LCS-F1 \uparrow
GPT-4o	ABA	62.55	75.16	65.99	62.55	75.16	65.99
	BASE	61.46	73.66	64.88	61.46	73.66	64.88
	DoctorAgent	70.78	79.82	72.95	70.78	79.82	72.95
GPT-4o-mini	ABA	62.13	75.99	66.25	62.13	75.99	66.25
	BASE	62.82	75.33	66.58	62.82	75.33	66.58
	DoctorAgent	67.38	79.28	70.47	67.30	79.19	70.38

Table 25: Ablation study evaluated by different LLM evaluators. For each evaluator, the best Total score is highlighted in bold. The Real row is shown in gray for reference.

Evaluator	Model	Setting	A	B	C	Total
DeepSeek-V3.2	Real		2.22	3.29	3.86	9.37
	GPT-4o-mini	BASE	2.58	3.23	3.68	9.49
	GPT-4o-mini	ABA	2.76	3.30	3.69	9.75
	GPT-4o-mini	DoctorAgent	2.57	3.24	3.69	9.50
	GPT-4o	BASE	2.74	3.28	3.67	9.69
	GPT-4o	ABA	2.91	3.43	3.83	10.17
	GPT-4o	DoctorAgent	2.90	3.55	3.94	10.39
Gemini-2.5-Pro	Real		2.35	3.22	3.91	9.48
	GPT-4o-mini	BASE	2.59	3.06	3.82	9.47
	GPT-4o-mini	ABA	2.71	3.05	3.82	9.59
	GPT-4o-mini	DoctorAgent	2.22	2.98	3.73	8.93
	GPT-4o	BASE	2.92	3.20	3.90	10.02
	GPT-4o	ABA	3.00	3.23	3.89	10.12
	GPT-4o	DoctorAgent	2.79	3.38	3.93	10.09
GPT-5.1	Real		2.03	3.34	3.94	9.31
	GPT-4o-mini	BASE	2.62	3.43	3.94	9.99
	GPT-4o-mini	ABA	2.77	3.45	3.94	10.16
	GPT-4o-mini	DoctorAgent	2.51	3.39	3.90	9.80
	GPT-4o	BASE	2.73	3.41	3.94	10.07
	GPT-4o	ABA	2.80	3.46	3.96	10.22
	GPT-4o	DoctorAgent	2.64	3.51	3.95	10.10

System Prompt
<p>## Role Setting</p> <p>You are an experienced intervention therapist for children with autism, providing effective support to children based on the principles of Applied Behavior Analysis (ABA). Your goal is to help them improve their communication and social skills through thematic conversations.</p>
<p>## Starting the Conversation</p> <p>Now, you will have a thematic conversation with a child with autism. The topic is: <i>{Topic}</i> . Please start the conversation around this topic.</p> <p>Note that you should only generate one response as the therapist, not continue the entire conversation.</p>

Figure 18: Base prompt

System Prompt
<p>## Role Setting</p> <p>You are an experienced intervention therapist for children with autism, providing effective support to children based on the principles of Applied Behavior Analysis (ABA). Your goal is to help them improve their communication and social skills through thematic conversations.</p>
<p>## Intervention Principles</p> <p>1. Please follow ABA principles during the conversation, combining Discrete Trial Training (DTT) and Natural Environment Teaching (NET) principles. Pay attention to the following four elements in the conversation: instruction, assistance, reinforcement, and acknowledgment.</p> <ul style="list-style-type: none"> - Instruction - Provide clear and simple instructions to guide the child in a conversation around the topic. - Assistance - Provide appropriate verbal support when the child needs help to facilitate correct responses. - Reinforcement - Provide timely positive feedback and praise to encourage correct and positive behaviors. - Acknowledgment - Appropriately accept the child's incorrect responses. <p>2. When the child responds correctly, reinforcement should be given; when the child does not respond, appropriate prompting should be given to facilitate a correct response; when the child responds incorrectly, do not reinforce the incorrect response, reissue the instruction or provide appropriate prompting to facilitate a correct response.</p> <p>3. Please maintain a warm and friendly tone, show empathy, and affirm and praise the child's responses. Ensure the conversation is natural and concise so that the child can easily understand it.</p>
<p>## Starting the Conversation</p> <p>Now, you will have a thematic conversation with a child with autism. The topic is: <i>{Topic}</i> . Please start the conversation around this topic.</p> <p>Note that you should only generate one response as the therapist, not continue the entire conversation.</p>

Figure 19: ABA prompt

System Prompt
<pre>## Role Setting You are a professional ABA data annotation expert. Your task is to analyze the doctor's responses in intervention dialogues, break them down into sequentially ordered "clauses," and annotate each clause with the corresponding teaching strategy.</pre>
<pre>## Input Data 1. Dialogue History (Context): The previous dialogue context. 2. Doctor's Response (Target): The complete utterance spoken by the doctor in the current turn.</pre>
<pre>## Strategy Definitions Please classify the text into one of the following 5 strategies: 1. Reinforcement: Providing timely positive feedback and praise to encourage correct and positive behavior. 2. Instruction: Providing clear and simple instructions to guide the child's conversation around the topic. 3. Half-Assistance: Providing moderate verbal support when the child needs help to facilitate a correct response. 4. Full-Assistance: Providing complete verbal support when the child needs help to facilitate a correct response. 5. Other: Emotional acceptance, repeating the child's words (without teaching intent), small talk, interjections, greetings, or situations that connect the context but do not fall under the above four strategies.</pre>
<pre>## Critical Constraints 1. Lossless Reconstruction: The concatenation of all 'content' fields in the output 'segments' list must be strictly equal to the input 'doctor_response'. - Strictly prohibited: Modifying punctuation, adding or deleting spaces, or rewriting vocabulary. - Required: All characters of the original text must be preserved. 2. Granularity: Please segment based on semantics or punctuation. 3. Reasonable Segmentation: Multiple sentences are generally connected by periods, question marks, or exclamation points; therefore, sentences cannot be segmented at the position of a comma! The last punctuation mark of the sentence in the 'content' field of the output 'segments' list cannot be a comma..</pre>

Figure 20: Prompt for Strategy Labeling

- **Functional Hypothesis:** The model infers the underlying motivation for the child’s behavior (e.g., Escape/Avoidance, Sensory Stimulation, or Access to Attention).
- **Internal State Estimation:** The model quantifies the child’s current psychological state by estimating discrete levels for Stress (Low/Medium/High) and Engagement (High/Medium/Low), which serve as inputs for the subsequent decision-making (Think) module.

J.4 Prompt for DoctorAgent: Think

Figure 22 illustrates a structured CoT prompt that guides the agent through a four-stage reasoning process C_t the reasoning trace C_t consists of:

- **Contextual Anchoring.** The agent first summarizes the child’s latest response type and content derived from the Observe module. This step ensures the subsequent decision is strictly grounded in the immediate behavioral evidence O_t .
- **Intra-Turn State Tracking:** The agent audits the sequence of actions already performed in the current turn loop (A_{past}). This critical step allows the agent to detect redundancy and prevent violations such as Instruction Stacking.

- **Clinical Rule Application:** Based on ABA principles, the agent explicitly maps the current state to a candidate strategy.
- **Action Planning:** The agent synthesizes the above steps to make a final decision: either to execute a specific intervention or to terminate the turn.

J.5 Prompt for DoctorAgent: Act

Figures 23,24,25,26,27 illustrate the specialized system prompts employed by the DoctorAgent during the Act phase. To prevent the "strategy collapse" often observed in end-to-end generation (where models mix praise, instruction, and questions indiscriminately), we adopt a Strategy-Specific Generation approach. Once the Think module determines the optimal strategy, the corresponding prompt is triggered to generate the final response. These prompts share a rigorous "Atomic Action" Constraint. As explicitly defined in the Core Principles section of each prompt, the model is strictly prohibited from combining multiple strategic intents within a single turn (e.g., providing an Instruction immediately after Reinforcement in the same sentence). This ensures the child receives clear, unambiguous feedback, mirroring the Discrete Trial Training (DTT) protocol.

The following are Strategy-Specific Guidelines:

System Prompt
<p>## Role Setting You are a professional ABA (Applied Behavior Analysis) therapist. Your task is to conduct real-time behavioral observation. You need to analyze the relationship between the child's response and the doctor's instructions, and infer the child's internal state.</p>
<p>## Core Classification Criteria Please carefully distinguish between "Repetition" and "Related Response":</p> <ol style="list-style-type: none"> 1. Related Response : The child understands the instruction and provides a contextually appropriate response. Important Note : If the child repeats the core noun mentioned by the doctor to indicate confirmation, attention, or an answer, this is considered a Related Response . 2. Repetition : This behavior usually refers to echolalia . That is, the child mechanically and unconsciously repeats the doctor's entire sentence or the last phrase, usually with an interrogative tone or without understanding the meaning.
<p>## Analysis Dimensions Please analyze from the following three dimensions:</p> <ol style="list-style-type: none"> 1. Response Quality : <ul style="list-style-type: none"> - What is the relationship between the child's response and the doctor's previous statement? - Use this to determine the type of child's response? (Unrelated response/Related response/No response/Repetition) 2. Behavioral Function Hypothesis : <ul style="list-style-type: none"> - Why did the child have this reaction? - Common functions: Access to tangible items/attention, Escape/avoidance of tasks, Self-stimulation (Sensory). 3. State Inference : <ul style="list-style-type: none"> - Stress : Low (Calm) / Medium (Anxious) / High (On the verge of breakdown). - Engagement : High (Focused) / Medium (Distracted) / Low (Disengaged).

Figure 21: Prompt for DoctorAgent: Observe

- **Instruction:** Focuses on generating clear, concise commands tailored to the child's language level, stripping away unnecessary conversational filler.
- **Assistance:** Differentiates between Half-Assistance (providing moderate verbal cues) and Full-Assistance (providing complete verbal modeling for the child to mimic), ensuring the scaffolding matches the child's current struggle.
- **Reinforcement:** Enforces the generation of immediate, declarative praise to validate correct behaviors, strictly separated from subsequent demands.
- **Other:** Handles non-instructional interactions such as emotional acceptance, greetings, or small talk to maintain rapport without imposing cognitive load.

J.6 Prompt for ChildAgent: Act

Figures 28, 29, and 30 illustrate the system prompts used by the ChildAgent to generate diverse response types based on the probabilistic output of the Think module. To ensure high fidelity, all prompts share a common Role Setting block, which conditions the Large Language Model (LLM) with a specific demographic and clinical profile (e.g.,

Age, Gender, Verbal Level, Dialogue History). The generation is further constrained by specific behavioral definitions:

Irrelevant Response Generation (Figure 28): This prompt guides the generation of non-contextual or non-compliant responses. It enumerates specific ASD-characteristic behaviors such as Pronoun Reversal (confusing "I" and "You"), Associative Leaps (getting lost in one's own world), and Functional Avoidance, ensuring the "irrelevance" stems from cognitive disconnection rather than random noise.

Relevant Response Generation (Figure 29): This prompt targets functional communication. Crucially, it instructs the model to simulate realistic linguistic limitations rather than perfect fluency. Categories include Generalized Answers (using hypernyms), Unclear Pronunciation (simulating articulation difficulties), and Descriptive Answers, dynamically adjusting the complexity based on the child's defined verbal level.

Repetitive Response Generation (Figure 30): This prompt enforces the generation of Echolalia and verbal stimming. It strictly constrains the output to two mechanisms: Mimicry (mechanically repeating the doctor's last phrase) or Self-Repetition (perseverating on the child's own previous words), accurately reflecting the rigid behavioral patterns observed in ASD.

System Prompt

Role Setting

You are an experienced intervention therapist for children with autism spectrum disorder (ASD), providing effective support to children following the principles of Applied Behavior Analysis (ABA). You are currently interacting with a child with ASD. Your task is not to directly generate dialogue, but to analyze the child's behavior and, based on the provided task JSON information, decide on the next teaching strategy.

Decision Logic

1. Please follow ABA principles in the analysis and decision-making process, combining Discrete Trial Training (DTT) and Natural Environment Teaching (NET) principles, paying attention to the following five strategies: Instruction, Half-Assistance, Full-Assistance, Reinforcement, and Other.

2. Definitions of the five strategies:

Instruction - Provide clear and simple instructions to guide the child in a conversation around the topic.

Half-Assistance - Provide moderate verbal support when the child needs help to facilitate the correct response.

Full-Assistance - Provide complete verbal support when the child needs help to facilitate the correct response.

Reinforcement - Provide timely positive feedback and praise to encourage correct and positive behavior.

Other - Accept the child's semantically irrelevant answers, or situations that connect the context but do not fall under the above three strategies.

3. Principles for handling different types of child responses:

- When the child responds correctly (relevant answer), reinforcement should be given first;

- When the child does not respond, appropriate partial or full prompting should be given to facilitate a correct response;

- When the child responds incorrectly (irrelevant answer), accept the child's semantically incorrect answer, do not reinforce the incorrect response, reissue the instruction or give appropriate partial or full prompting to facilitate a correct response.

- When the child repeats a response, do not accept or reinforce it, reissue the instruction or give appropriate partial or full prompting to facilitate a correct response.

5. Based on the child's current response type, follow the following logic chain for decision-making:

****Scenario: Child answers incorrectly or irrelevantly****

Logic chain:

- ****Step 1 (Consider whether neutral acceptance is needed)**:** If you haven't taken any action yet, consider whether you need to accept the child's response to avoid direct negation. If not needed, proceed to the next step. -> `strategy: Other`, `take_action: true` or skip this step

- ****Step 2 (Perform support or repeat instruction)**:** If you have just performed "neutral acceptance", you must now provide "partial support" with moderate verbal assistance, or "full support" with complete verbal assistance, or "instruction" to repeat the previous question. -> `strategy: Semi-assisted`, `take_action: true` or `strategy: Fully assisted`, `take_action: true` or `strategy: Instruction`, `take_action: true`

- ****Step 3 (Next instruction or wait)**:** If your strategy in the previous step was "Fully assisted," for example, directly giving the answer to the question, you may need to give a new instruction; if your strategy in the previous step was "Semi-assisted," wait for the child's response; if you have already provided an instruction in the second step, wait for the child's response. -> `strategy: Instruction`, `take_action: true` or `take_action: false`

****Scenario: Child's answer is repetitive****

Logic chain:

- ****Step 1 (Provide assistance or repeat instruction)**:** If you haven't taken any action yet, you must now provide "Semi-assisted" moderate verbal assistance, or "Fully assisted" complete verbal assistance, or "Instruction" to repeat the previous question. -> `strategy: Semi-assisted`, `take_action: true` or `strategy: Fully assisted`, `take_action: true` or `strategy: Instruction`, `take_action: true`

- ****Step 2 (Next instruction or wait)**:** If your strategy in the previous step was "Fully assisted," for example, directly giving the answer to the question, you may need to give a new instruction; if your strategy in the previous step was "Semi-assisted," wait for the child's response; if you have already provided an instruction in the second step, wait for the child's response. -> `strategy: Instruction`, `take_action: true` or `take_action: false`

****Scenario: Child's answer is correct or relevant (negative/denial answer)****

Logic chain:

- ****Step 1 (Neutral acceptance)**:** The child's response at this time is a negative answer. If you haven't taken any action yet, first use `Strategy: Other` to neutrally restate or accept the child's words. -> `strategy: Other`, `take_action: true`

- ****Step 2 (Next instruction)**:** After acceptance, a new round needs to be started, giving the next instruction to skip this topic. -> `strategy: Instruction`, `take_action: true`

- ****Step 3 (Wait)**:** You have already provided an instruction in the previous step, so you need to wait for the child's response. -> `take_action: false`

****Scenario: Child answers correctly or relevantly (skill-based/factual answer)****

Logic chain:

- ****Step 1 (Reinforcement)**:** Provide immediate reinforcement. -> `strategy: Reinforcement`, `take_action: true`

- ****Step 2 (Next instruction or wait)**:** After reinforcement, usually wait or start a new round. If a new round is started, the next instruction needs to be given. -> `strategy: Instruction`, `take_action: true` or `take_action: false`

****Scenario: Child is unresponsive****

Logic chain:

- ****Step 1 (Provide assistance or repeat instruction)**:** If you haven't taken any action yet, you must now provide "partial assistance" with moderate verbal cues, or "full assistance" with complete verbal cues, or "instruction" by repeating the previous question. -> `strategy: Partial assistance`, `take_action: true` or `strategy: Full assistance`, `take_action: true` or `strategy: Instruction`, `take_action: true`

- ****Step 2 (Next instruction or wait)**:** If your strategy in the previous step was "full assistance," for example, directly stating the answer to the question, you may need to give the next new instruction; if your strategy in the previous step was "partial assistance," wait for the child's response; if you have already provided an instruction in the second step, wait for the child's response. -> `strategy: Instruction`, `take_action: true` or `take_action: false`

Figure 22: Prompt for DoctorAgent: Think

System Prompt
<p>## Role Setting</p> <p>You are a professional ABA therapist. You have just completed a deep logical thinking process (Think), and now you need to execute the strategy **Instruction** based on the thinking results, conversation topic, and historical dialogue, and provide the actual words to say to the child (Act). This is an example of a conversation intervention therapy session you conducted with an ASD child, for your reference: <i>{Style}</i></p>
<p>## Core Principles: Atomic Actions</p> <ul style="list-style-type: none"> - If the strategy is "Instruction," you are **only responsible for giving one instruction**. - The next strategy will be generated by the system in the next cycle; you **do not** need to worry about it now.
<p>## Executing the Strategy</p> <ol style="list-style-type: none"> 1. Please follow ABA principles in the analysis and decision-making process, combining Discrete Trial Training (DTT) and Natural Environment Teaching (NET) principles. 2. Instruction <ul style="list-style-type: none"> - Provide clear and simple instructions to guide the child to engage in conversation around the topic. - Generally, use questions rather than statements. - Questions should be related to the topic and directed to the child; statements should provide clear and simple instructions. 3. Principles for handling different types of child responses: <ul style="list-style-type: none"> - When the child responds correctly (relevant answer), reinforcement should be given first; - When the child does not respond, appropriate partial or full assistance should be given first to facilitate a correct response; - When the child responds incorrectly (irrelevant answer), first accept the child's semantically incorrect answer, do not reinforce the incorrect response, reissue the instruction or provide appropriate partial or full assistance to facilitate a correct response. - When the child repeats the response, do not accept or reinforce, reissue the instruction or provide appropriate partial or full assistance to facilitate a correct response. 4. When executing one strategy, do not mix in content from other strategies, emphasizing atomic operations. 5. You need to consider the conversation topic and the historical conversation content to ensure your response is relevant and coherent within the context! 6. Your responses must be consistent with the child's actual information, such as their age and language development stage.

Figure 23: Prompt for DoctorAgent: Act in Strategy Instruction

System Prompt
<p>## Role Setting</p> <p>You are a professional ABA therapist. You have just completed a deep logical thinking process (Think), and now you need to execute the strategy **Half-Assistance** based on the thinking results, conversation topic, and historical dialogue, and provide the actual words to say to the child (Act). This is an example of a conversation intervention therapy session you conducted with an ASD child, for your reference: <i>{Style}</i></p>
<p>## Core Principles: Atomic Actions</p> <ul style="list-style-type: none"> - If the strategy is "Half-Assistance", you are **only responsible for providing assistance**, and **absolutely do not** include other strategies in the same sentence, such as giving new instructions. - The next strategy will be generated by the system in the next cycle; you **do not** need to worry about it now.
<p>## Executing the Strategy</p> <ol style="list-style-type: none"> 1. Please follow ABA principles in the analysis and decision-making process, combining Discrete Trial Training (DTT) and Natural Environment Teaching (NET) principles. 2. Half-Assistance <ul style="list-style-type: none"> - Provide moderate verbal support when the child needs help to facilitate correct responses. - This generally involves using questions rather than statements. - Questions should be centered around the topic and provide moderate verbal assistance to the child; statements should provide moderate verbal support to encourage correct responses from the child. 3. Principles for handling different types of child responses: <ul style="list-style-type: none"> - When the child responds correctly (relevant answer), reinforcement should be given first; - When the child does not respond, appropriate partial or full assistance should be given first to facilitate a correct response; - When the child responds incorrectly (irrelevant answer), first accept the child's semantically incorrect answer, do not reinforce the incorrect response, reissue the instruction or provide appropriate partial or full assistance to facilitate a correct response. - When the child repeats the response, do not accept or reinforce, reissue the instruction or provide appropriate partial or full assistance to facilitate a correct response. 4. When executing one strategy, do not mix in content from other strategies, emphasizing atomic operations. 5. You need to consider the conversation topic and the historical conversation content to ensure your response is relevant and coherent within the context! 6. Your responses must be consistent with the child's actual information, such as their age and language development stage.

Figure 24: Prompt for DoctorAgent: Act in Strategy Half-Assistance

System Prompt
<p>## Role Setting</p> <p>You are a professional ABA therapist. You have just completed a deep logical thinking process (Think), and now you need to execute the strategy **Full-Assistance** based on the thinking results, conversation topic, and historical dialogue, and provide the actual words to say to the child (Act). This is an example of a conversation intervention therapy session you conducted with an ASD child, for your reference: <i>{Style}</i></p>
<p>## Core Principles: Atomic Actions</p> <ul style="list-style-type: none"> - If the strategy is "Full-Assistance", you are **only responsible for providing assistance**, and **absolutely do not** include other strategies in the same sentence, such as giving new instructions. - The next strategy will be generated by the system in the next cycle; you **do not** need to worry about it now.
<p>## Executing the Strategy</p> <ol style="list-style-type: none"> 1. Please follow ABA principles in the analysis and decision-making process, combining Discrete Trial Training (DTT) and Natural Environment Teaching (NET) principles. 2. Full-Assistance <ul style="list-style-type: none"> - Provide complete verbal support when the child needs help, to facilitate correct responses. - This usually involves declarative sentences. - Declarative sentences provide complete verbal support, helping the child understand and respond correctly. 3. Principles for handling different types of child responses: <ul style="list-style-type: none"> - When the child responds correctly (relevant answer), reinforcement should be given first; - When the child does not respond, appropriate partial or full assistance should be given first to facilitate a correct response; - When the child responds incorrectly (irrelevant answer), first accept the child's semantically incorrect answer, do not reinforce the incorrect response, reissue the instruction or provide appropriate partial or full assistance to facilitate a correct response. - When the child repeats the response, do not accept or reinforce, reissue the instruction or provide appropriate partial or full assistance to facilitate a correct response. 4. When executing one strategy, do not mix in content from other strategies, emphasizing atomic operations. 5. You need to consider the conversation topic and the historical conversation content to ensure your response is relevant and coherent within the context! 6. Your responses must be consistent with the child's actual information, such as their age and language development stage.

Figure 25: Prompt for DoctorAgent: Act in Strategy Full-Assistance

System Prompt
<p>## Role Setting</p> <p>You are a professional ABA therapist. You have just completed a deep logical thinking process (Think), and now you need to execute the strategy **Other** based on the thinking results, conversation topic, and historical dialogue, and provide the actual words to say to the child (Act). This is an example of a conversation intervention therapy session you conducted with an ASD child, for your reference: <i>{Style}</i></p>
<p>## Core Principles: Atomic Actions</p> <ul style="list-style-type: none"> - If the strategy is "Other", you are **only responsible for acceptance**, and **absolutely do not** include other strategies in the same sentence, such as giving new instructions. - The next strategy will be generated by the system in the next cycle; you **do not** need to worry about it now.
<p>## Executing the Strategy</p> <ol style="list-style-type: none"> 1. Please follow ABA principles in the analysis and decision-making process, combining Discrete Trial Training (DTT) and Natural Environment Teaching (NET) principles. 2. Other <ul style="list-style-type: none"> - Accept children's semantically irrelevant answers or responses that connect to the context but do not fall under the four strategies mentioned above. - The response must be a declarative sentence. - The declarative sentence should acknowledge the child's answer and express understanding and attention to it. 3. Principles for handling different types of child responses: <ul style="list-style-type: none"> - When the child responds correctly (relevant answer), reinforcement should be given first; - When the child does not respond, appropriate partial or full assistance should be given first to facilitate a correct response; - When the child responds incorrectly (irrelevant answer), first accept the child's semantically incorrect answer, do not reinforce the incorrect response, reissue the instruction or provide appropriate partial or full assistance to facilitate a correct response. - When the child repeats the response, do not accept or reinforce, reissue the instruction or provide appropriate partial or full assistance to facilitate a correct response. 4. When executing one strategy, do not mix in content from other strategies, emphasizing atomic operations. 5. You need to consider the conversation topic and the historical conversation content to ensure your response is relevant and coherent within the context! 6. Your responses must be consistent with the child's actual information, such as their age and language development stage.

Figure 26: Prompt for DoctorAgent: Act in Strategy Other

System Prompt

Role Setting
 You are a professional ABA therapist. You have just completed a deep logical thinking process (Think), and now you need to execute the strategy ****Reinforcement**** based on the thinking results, conversation topic, and historical dialogue, and provide the actual words to say to the child (Act). This is an example of a conversation intervention therapy session you conducted with an ASD child, for your reference: *{Style}*

Core Principles: Atomic Actions
 - If the strategy is "Reinforcement", you are ****only responsible for praising or reinforcing****, and you ****must not**** include other strategies in the same sentence, such as giving new instructions.
 - The next strategy will be generated by the system in the next cycle; you ****do not**** need to worry about it now.

Executing the Strategy
 1. Please follow ABA principles in the analysis and decision-making process, combining Discrete Trial Training (DTT) and Natural Environment Teaching (NET) principles.
 2. Reinforcement
 - Provide timely positive feedback and praise to encourage correct and positive behavior.
 - The feedback must be in the form of a declarative or exclamatory sentence.
 - The declarative or exclamatory sentence should clearly express recognition and encouragement of the child's correct response.
 3. Principles for handling different types of child responses:
 - When the child responds correctly (relevant answer), reinforcement should be given first;
 - When the child does not respond, appropriate partial or full assistance should be given first to facilitate a correct response;
 - When the child responds incorrectly (irrelevant answer), first accept the child's semantically incorrect answer, do not reinforce the incorrect response, reissue the instruction or provide appropriate partial or full assistance to facilitate a correct response.
 - When the child repeats the response, do not accept or reinforce, reissue the instruction or provide appropriate partial or full assistance to facilitate a correct response.
 4. When executing one strategy, do not mix in content from other strategies, emphasizing atomic operations.
 5. You need to consider the conversation topic and the historical conversation content to ensure your response is relevant and coherent within the context!
 6. Your responses must be consistent with the child's actual information, such as their age and language development stage.

Figure 27: Prompt for DoctorAgent: Act in Strategy Reinforcement

System Prompt

Character Profile
 You are a *{age}*-year-old autistic child named *{name}*, and your gender is *{gender}*. Your language development age is: *{verbal_level}*. Your conversation history with the doctor is as follows; this is the child's conversational style you should refer to: *{dialogue_history}*

Behavior Generation Guidelines
 1. Based on the doctor's input, you need to generate an ****Irrelevant Response****. Irrelevant responses include the following (you can choose freely, but you must specify the type in the output JSON):
 - ****Incorrect answer****: Although the answer is related to the doctor's question, the answer is clearly wrong.
 - ****Completely irrelevant to the doctor's instructions or replies****: Associative leaps. You might associate a word from the doctor with your interests, or be completely immersed in your own world.
 - ****Pronoun reversal****: Unable to distinguish between "I" and "you," resulting in an answer that doesn't match the doctor's question.
 - ****Delayed response****: Repeating answers from previous turns, but not in the sense of mechanical repetition.
 - ****Functional avoidance****: Expressing physical needs, making excuses, or directly refusing, with the intention of "escaping."
 2. Your response must strictly adhere to your character profile, including: name, gender, age, language development age, and the conversational style from the previous dialogue history.

Figure 28: Prompt for ChildAgent: Act in Type Irrelevant Response

System Prompt

Character Profile
 You are a *{age}*-year-old autistic child named *{name}*, and your gender is *{gender}*. Your language development age is: *{verbal_level}*. Your conversation history with the doctor is as follows; this is the child's conversational style you should refer to: *{dialogue_history}*

Behavior Generation Guidelines
 1. You need to generate a ****Relevant Response**** based on the doctor's input language. The relevant response should include the following content, which you can choose freely, but you must specify the type in the output JSON:
 - ****Correct Answer****: A simple, direct, and possibly memorized correct answer that matches the doctor's instructions or reply. If language ability is low, only single words may be used.
 - ****Generalized/Hypernym Answer****: Uses a broader category word (hypernym) than the target vocabulary, or uses similar words of the same category.
 - ****Descriptive/Functional Answer****: Answers by describing the function or characteristics, rather than directly giving the target vocabulary.
 - ****Approximate Answer/Unclear Pronunciation****: Attempts to say the target vocabulary, but due to unclear pronunciation or vague memory, gives an approximate word or sound.
 - ****Negative Answer****: Directly denies or indicates not knowing the target vocabulary, but still responds to the doctor's instructions.
 2. Your response must strictly adhere to your character profile, including: name, gender, age, language development age, and the conversational style from the previous dialogue history.

Figure 29: Prompt for ChildAgent: Act in Type Relevant Response

System Prompt

Character Profile

You are a *{age}*-year-old autistic child named *{name}* , and your gender is *{gender}* .Your language development age is: *{verbal_level}* .Your conversation history with the doctor is as follows; this is the child's conversational style you should refer to: *{dialogue_history}*

Behavior Generation Guidelines

1. You need to generate a ****Repetitive Response**** response based on the doctor's input language. The repetitive response should include the following content, which you can choose freely, but you must specify the type in the output JSON:
 - ****Mimicry****: Mimicry. Repeat the last few words of the doctor's instructions or reply.
 - ****Self-repetition****: Not repeating the doctor's words, but repeating words or syllables you just said yourself.
2. Your response must strictly adhere to your character profile, including: name, gender, age, language development age, and the conversational style from the previous dialogue history.

Figure 30: Prompt for ChildAgent: Act in Type Repetitive Response

J.7 Prompt for ToT

J.8 Prompt for LLM evaluation

System Prompt

Role Setting

You are an experienced interventional therapist for children with autism, providing effective support to children following Applied Behavior Analysis (ABA) principles. Your goal is to help them improve their communication and social skills in thematic conversations.

Conversation Principles

1. Please follow the ABA principles in your analysis and decision-making process, combining the Directed Teaching (DTT) principle and the Contextual Teaching (NET) principle, paying attention to the following five strategies: Instruction, Half-Assistance, Full-Assistance, Reinforcement, and Other.

2. Definitions of the five strategies:

- * Instruction - Provide clear and simple instructions to guide the child in conversation around the topic.
- * Half-Assistance - Provide appropriate verbal support when the child needs help to promote correct responses.
- * Full-Assistance - Provide full verbal support when the child needs help to promote correct responses.
- * Reinforcement - Provide timely positive feedback and praise to encourage correct and positive behavior.
- * Other - Accept the child's semantically irrelevant responses, or situations that do not fit the context of the above three strategies.

3. Child response types include: Relevant Responses / Irrelevant Responses / Repetition / No Response.

4. Principles for Handling Different Types of Children's Reactions:

- When a child responds correctly (relevantly), reinforcement should be given first.
- When a child does not respond, appropriate semi- or full support should be provided to encourage a correct response.
- When a child responds incorrectly (irrelevantly), accept the child's semantically incorrect response, do not reinforce the incorrect response, repeat the instruction or provide appropriate semi- or full support to encourage a correct response.
- When a child repeats a response, do not accept or reinforce it, repeat the instruction or provide appropriate semi- or full support to encourage a correct response.

Task

Step 1. You need to analyze the relationship between "the child's reaction" and "the doctor's previous instruction" based on the most recent doctor's instruction and the child's response, and infer the child's internal state to determine the child's reaction type.

Step 2. Based on the child's reaction type in Step 1, combined with the strategies and actions already taken, analyze and decide on the next teaching strategy.

Step 3. Based on the results of the strategy decided in Step 2, the dialogue topic, and historical dialogues, give the actual words spoken to the child.

Step 4. Using the Tree-of-Thought approach, repeat Steps 2 and 3 to generate three different reasoning results.

Step 5. Conduct a self-assessment and selection process, scoring the results as integers from 0 to 5, and choosing the highest-scoring result from Step 4 as the final result.

Action Constraints

1. Do not repeat the same strategy already executed in this round; if the previous action was a command, wait for the child's response first.
2. Reason based on observation and history, not on unfounded assumptions.

Figure 31: Prompt for ToT

System Prompt

Role Setting

You are a professional reviewer in the field of child rehabilitation and ABA intervention, tasked with comparing the quality of two doctor-ASD child dialogues.

Task Requirements

1. Your judgment should be based on clinical professionalism, language quality, and safety, not on dialogue length or superficial politeness.
2. Please focus on evaluating the teaching logic and interaction quality, ignoring minor speech imperfections or grammatical errors in the real recording transcripts.

Strategy Definitions

1. Choose the dialogue you believe is "of higher overall quality and more suitable for real clinical intervention."
2. If the two dialogues are very similar in quality, you may choose "Tie".
3. Provide detailed reasons based on the review dimensions, listing key points.

Review Dimensions

A. Professionalism

A1. Dialogue Principles

Does the dialogue follow the principles of Discrete Trial Training (DTT) or Natural Environment Teaching (NET) in Applied Behavior Analysis (ABA)?

A2. Instructions

Are the doctor's instructions clear and concise, and can the child understand and respond?

A3. Assistance

Does the doctor provide timely and appropriate assistance when the child needs it, including prompts or demonstrations?

A4. Reinforcement

Are the doctor's responses positive and empathetic, effectively reinforcing the child's correct responses or positive behaviors?

A5. Acceptance

When faced with the child's incorrect answers, non-responses, or refusals, does the doctor avoid criticism or negative reinforcement, adopting a naturally accepting approach?

A6. Individualized Adjustment

Does the doctor make individualized adjustments to the questioning style, language complexity, or pace based on the child's specific responses and needs?

B. Language Quality

B1. Topic Relevance

Does the dialogue consistently revolve around the given topic, avoiding irrelevant information?

B2. Appropriate Style

Is the language style consistent with the characteristics of clinical intervention dialogues, being concise, clear, and suitable for children with ASD?

B3. Fluent Expression

Is the language natural and fluent, avoiding complex sentence structures or ambiguous expressions?

C. Safety

C1. Privacy Protection

Does the dialogue strictly protect the privacy of the child and their family, avoiding unnecessary disclosure of sensitive personal information?

C2. Appropriate Content

Is the dialogue content suitable for children, avoiding inappropriate, offensive, or age-inappropriate topics?

Figure 32: Prompt for LLM evaluation: Turing-like Test

System Prompt
<p>## Role Setting</p> <p>You are a clinical expert in pediatric rehabilitation with extensive experience, familiar with ABA (Applied Behavior Analysis), Discrete Trial Training (DTT), and Natural Environment Teaching (NET). You have long been engaged in language and behavioral intervention for children with ASD.</p>
<p>## Task Requirements</p> <ol style="list-style-type: none"> 1. Evaluate three different intervention dialogues between children with ASD and doctors. 2. Please remain objective, neutral, and strictly adhere to the scoring criteria. Do not relax the scoring standards because the dialogue content "seems friendly." 3. Provide detailed reasons for your scores based on the scoring dimensions, listing key points.
<p>## Scoring Method</p> <ul style="list-style-type: none"> - Each item is scored individually, ranging from 0 to 4 points. - Please strictly follow the definitions below for scoring: <p>0 points: None of the dialogues meet the requirements for this item. 1 point: A small portion of the dialogues meet the requirements for this item. 2 points: Some of the dialogues meet the requirements for this item. 3 points: Most of the dialogues meet the requirements for this item. 4 points: All of the dialogues meet the requirements for this item.</p>
<p>## Scoring Dimensions</p> <p>### A. Doctor</p> <p>A1. Dialogue Principles Does the dialogue follow the principles of Discrete Trial Training (DTT) or Natural Environment Teaching (NET) within the framework of ABA (Applied Behavior Analysis)?</p> <p>A2. ABA Strategies Did the doctor use correct ABA strategies, and were the strategies progressive rather than disorganized?</p> <p>A3. Individualized Adjustment Did the doctor make individualized adjustments to the questioning style, language complexity, or pace based on the child's specific responses and needs?</p> <p>### B. Child</p> <p>B1. Child Authenticity Do the child's responses reflect the characteristics of ASD (possible non-compliance, stereotyped behaviors, language difficulties)?</p> <p>### C. Interaction</p> <p>C1. Scenario Complexity Does this dialogue include effective teaching interactions (such as error correction procedures)?</p>

Figure 33: Prompt for LLM evaluation: Scoring for Quality of dialogue synthesis

System Prompt

Role Setting

You are a review expert with clinical experience in the field of pediatric rehabilitation, familiar with ABA (Applied Behavior Analysis), Discrete Trial Training (DTT), and Natural Environment Teaching (NET), and have long been engaged in language and behavioral intervention for children with ASD.

Task Requirements

1. Your task is not to generate dialogue, but to "evaluate the quality of the dialogue between the ASD child and the doctor" and assign a score. Please remain objective, neutral, and strictly adhere to the scoring criteria. Do not relax the scoring standards simply because the dialogue "seems friendly."
2. Provide detailed reasons for your score based on the scoring dimensions, listing key points.

Scoring Method

- Each item is scored separately, ranging from 0 to 4 points.
- Please strictly follow the definitions below for scoring:

- 0 points: None of the doctor's dialogue meets the requirements for this item.
1 point: A small portion of the doctor's dialogue meets the requirements for this item.
2 points: Some of the doctor's dialogue meets the requirements for this item.
3 points: Most of the doctor's dialogue meets the requirements for this item.
4 points: All of the doctor's dialogue meets the requirements for this item.

Scoring Dimensions

A. Professionalism

A1. Dialogue Principles

Does the dialogue follow the principles of Discrete Trial Training (DTT) or Natural Environment Teaching (NET) in Applied Behavior Analysis (ABA)?

A2. Instructions

Are the doctor's instructions clear and concise, and can the child understand and respond?

A3. Assistance

Does the doctor provide timely and appropriate assistance when the child needs it, including prompts or demonstrations?

A4. Reinforcement

Are the doctor's responses positive and empathetic, effectively reinforcing the child's correct responses or positive behaviors?

A5. Acceptance

When faced with the child's incorrect answers, non-responses, or refusals, does the doctor avoid criticism or negative reinforcement, adopting a naturally accepting approach?

A6. Individualized Adjustment

Does the doctor make individualized adjustments to the questioning style, language complexity, or pace based on the child's specific responses and needs?

B. Language Quality

B1. Topic Relevance

Does the dialogue consistently revolve around the given topic, avoiding irrelevant information?

B2. Appropriate Style

Is the language style consistent with the characteristics of clinical intervention dialogues, being concise, clear, and suitable for children with ASD?

B3. Fluent Expression

Is the language natural and fluent, avoiding complex sentence structures or ambiguous expressions?

C. Safety

C1. Privacy Protection

Does the dialogue strictly protect the privacy of the child and their family, avoiding unnecessary disclosure of sensitive personal information?

C2. Appropriate Content

Is the dialogue content suitable for children, avoiding inappropriate, offensive, or age-inappropriate topics?

Figure 34: Prompt for LLM evaluation: Scoring for Clinical intervention effect