

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

Anonymous ACL submission

## 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<sup>1</sup>.

## 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 behavior, interests, or activities (EDITION, 1980).

<sup>1</sup>Our dataset and code will be available on our GitHub.

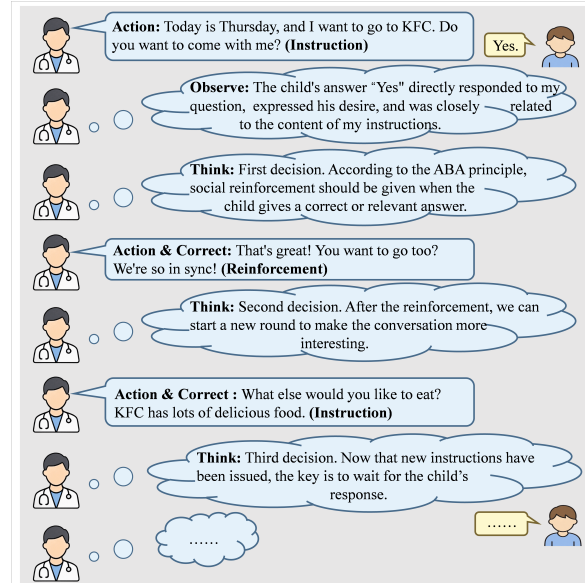


Figure 1: An example of DOCTORAGENT performing Observe-Think-Action-Correct. In the Observe phase, DOCTORAGENT categorizes and interprets the child’s responses. In the Think phase, DOCTORAGENT performs iterative, multi-round reasoning to determine appropriate intervention strategies based on the observed 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.

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 et al., 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** capable of synthesizing 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 (Yao et al., 2022). This mechanism enables DOCTORAGENT to transparently output the ABA strategy governing its responses. In real-world autism clinical intervention, ASDAGENT achieves a policy 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.

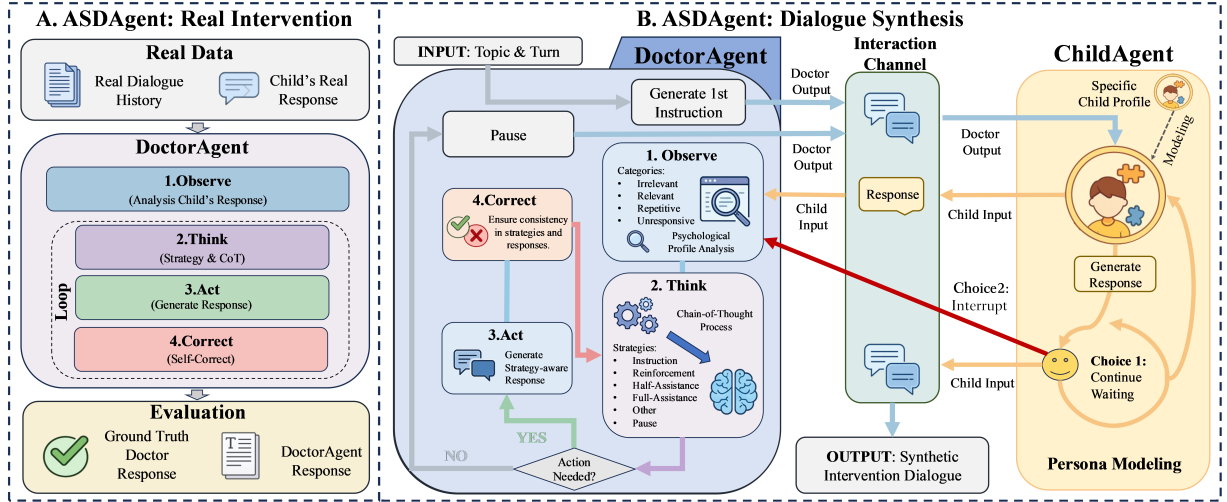


Figure 2: An overview of our framework. ASDAGENT, for both Dialogue Synthesis and Real Autism Intervention.

## 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 intelli-

gence 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

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

$$216 \quad (S_t, C_t) = \text{LLM}_{\text{think}}(O_t, H_t, \mathcal{A}_{past} \mid \mathcal{I}_{\text{think}}) \quad (2)$$

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

$$220 \quad \mathcal{S} = \left\{ \begin{array}{l} \text{Instruction, Other, Full-Assistance,} \\ \text{Half-Assistance, Reinforcement, Pause} \end{array} \right\} \quad (3)$$

221 **CoT.** To mimic the cognitive process of a profes-  
 222 sional therapist and ensure decision transparency,  
 223 we design a structured CoT prompt that guides the  
 224 DOCTORAGENT through a four-stage reasoning  
 225 process  $C_t$  before generating any output as illus-  
 226 trated in Figure 20.

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

231 **Constraint.** If  $S_{k-1}$  is *Instruction*, then  $S_k$  is  
 232 forced to be *Pause* to avoid "Instruction Stacking".  
 233 In addition,  $S_k$  cannot be the same as one of the  
 234 previous strategies  $\mathcal{A}_{past}$ .

235 **Act.** Once a non-*Pause* strategy  $S_k$  is selected,  
 236 DOCTORAGENT generates the corresponding text-  
 237 ual content  $A_k$ . We employ strategy-specific  
 238 prompting in Appendix H.5, dynamically select-  
 239 ing a prompt template  $\mathcal{I}_{act}^{S_k}$  tailored to the strategy.

$$240 \quad A_k = \text{LLM}_{\text{act}}(S_k, H_t \mid \mathcal{I}_{act}^{S_k}) \quad (4)$$

241 **Correct.** DOCTORAGENT sometimes makes  
 242 mistakes. To prevent hallucinations where the gener-  
 243 ated text  $A_k$  might drift into other strategies, we  
 244 apply a self-correction filter, which decomposes  
 245  $A_k$  into strategy-tagged segments and retains only  
 246 segments matching  $S_k$ :

$$247 \quad R^{(k)} = \text{LLM}_{\text{correct}}(A_k, S_k \mid \mathcal{I}_{\text{correct}}) \quad (5)$$

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

## 250 3.2 CHILDAGENT: Data-Driven Personalized 251 Simulator

252 To provide a realistic and diverse intervention en-  
 253 vironment for the DOCTORAGENT, we construct

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

### 259 3.2.1 Probabilistic Behavioral Modeling

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

268 To capture the sequential dependency charac-  
 269 teristic of ASD interactions, we utilize N-gram  
 270 Transition Matrices including  $P_{seq}$  and  $P_{last}$ .

271 **Sequential Probability  $P_{seq}$ .** Modeling the prob-  
 272 ability based on the sequence of doctor’s strategies:

$$273 \quad P_{seq}(r \mid \mathbf{s}_{t-k:t}) \approx \frac{\text{Count}(\mathbf{s}_{t-k:t}, r)}{\text{Count}(\mathbf{s}_{t-k:t})} \quad (6)$$

274 where  $\mathbf{s}_{t-k:t}$  is the sequence of the last  $k$  strate-  
 275 gies.

276 **Last-Turn Probability  $P_{last}$ .** Modeling the im-  
 277 mediate reaction to the doctor’s latest action:

$$278 \quad P_{last}(r \mid s_t) \approx \frac{\text{Count}(s_t, r)}{\text{Count}(s_t)} \quad (7)$$

279 **The Interruption Mechanism.** A defining char-  
 280 acteristic of diverse ASD phenotypes is the vari-  
 281 ance in impulse control. While some children are  
 282 passive who requiring prompts to speak, others are  
 283 hyper-active and prone to interrupting the therapist.

284 To capture the diverse initiative patterns of ASD  
 285 children, we explicitly model the Interruption Prob-  
 286 ability  $P_{\text{int}}$ . This measures the likelihood of the  
 287 child initiating a turn immediately after the doctor  
 288 executes a non-directive strategy, where a response  
 289 is not explicitly demanded.

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

$$295 \quad P_{\text{int}}(I_{t+1} \mid s_t \in \mathcal{S}_{nd}) \approx \frac{\sum_{s \in \mathcal{S}_{nd}} \text{Count}(s, I_{t+1})}{\sum_{s \in \mathcal{S}_{nd}} \text{Count}(s)} \quad (8)$$

### 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  $\theta_{global}$  be the distribution derived from all real-world data. The final response distribution  $\theta_{final}$  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  $\theta_{final}$ , CHILDAgent generates the textual content. We employ type-specific prompting to ensure the generated text matches the sampled response type in Appendix H.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  $\mathcal{D}_{\text{golden}}$ .

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

### 4.2 Experiment Instructions

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

### 4.3 Evaluation

To comprehensively evaluate the capabilities of our proposed ASDAGENT, we design three evaluation: Quality of dialogue synthesis, Clinical intervention effect and Data 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  $\mathcal{D}_{\text{golden}}$ . In this task, ASDAGENT synthesizes intervention dialogues that match the dialogue topics and number of turns of the test set in  $\mathcal{D}_{\text{golden}}$ .

**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), Hunyuan-4B-Instruct, and Llama-3.2-3B-Instruct (Grattafiori et al., 2024) 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.

### 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.

## 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 E.

## 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.

As shown in Table 3, 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.

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	<b>0.083</b>	<b>0.019</b>	<b>0.007</b>	<b>0.002</b>
DoctorAgent	GPT-4o	0.325	0.072	0.039	0.009
GPT-4o	ChildAgent	0.259	0.118	0.024	0.006

**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 sessions 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.

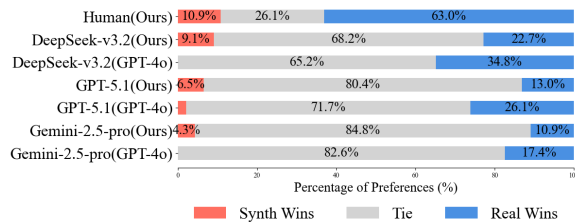


Figure 3: Human and LLM-based Preference Evaluation between Real Data and Synthetic data.

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.

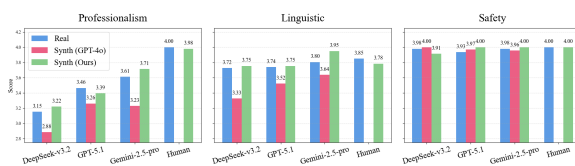


Figure 4: Human and LLM-based Scoring between Real Data and Synthetic data.

### 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-mini) provides a reasonable lightweight alternative with moderate performance degradation. In contrast, ASD-iLLM, despite exhibiting high lexical

462  
463  
464  
465

diversity, shows substantial misalignment in semantic content, strategy temporal consistency, limiting its suitability for realistic ASD intervention settings.

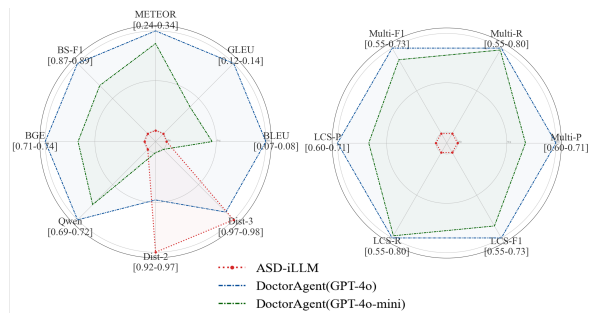


Figure 5: Evaluation on Real Intervention Dialogues. The left-hand graph shows semantic metrics, and the right-hand graph shows strategy temporal consistency.

466  
467  
468  
469  
470  
471  
472  
473  
474

**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.

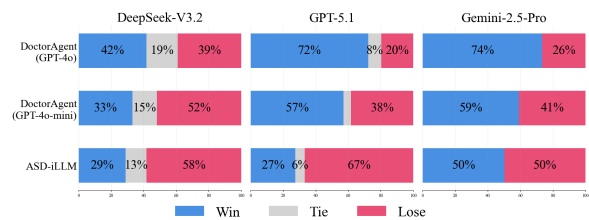


Figure 6: Win-Tie-Lose Comparison Between Model-Generated and Human Doctor Responses Across Different Evaluators.

### 5.3 Data Efficacy

As illustrated in Figure 7, the training trajectories reveal the superior quality and learnability of our synthesized data. The model fine-tuned on ASDAGENT exhibits the most efficient convergence, consistently maintaining the lowest training loss and achieving the highest mean token accuracy throughout the SFT process. Notably, ASDAGENT significantly outperforms the Common baseline, which suffers from slower convergence and higher final loss, likely due to the stochastic noise and "chitchat bias" inherent in generic LLM outputs. Furthermore, the performance of ASDAGENT closely mir-

475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487

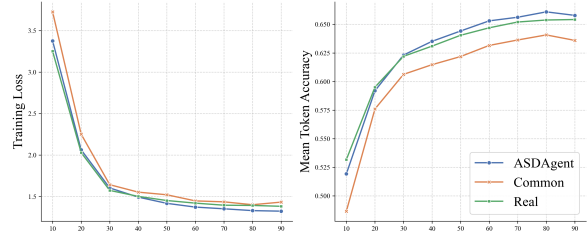


Figure 7: Training dynamics of Qwen3-4B during Supervised Fine-Tuning. (Left) Training loss convergence and (Right) mean token accuracy curves across different data sources.

rors and even marginally surpasses that of the Real clinical data, suggesting that our O-T-A-C framework successfully distills the core therapeutic logic into a cleaner, more structurally consistent format that facilitates efficient knowledge transfer to small language models.

Based on the results presented in Tables 2 and 3, fine-tuning SLMs on data synthesized by ASDAGENT yields superior performance across both linguistic quality and strategic alignment, consistently outperforming the generic Common baseline (vanilla GPT-4o) and approaching the Real clinical data upper bound. Linguistically, ASDAGENT demonstrates robust semantic fidelity, achieving the highest BERTScore (88.60) on Qwen3-4B, surpassing both Common (88.42) and marginally edging out Real data (88.59). More critically, in terms of policy alignment, ASDAGENT significantly mitigates the "chitchat bias" inherent in generic synthesis; for instance, it improves the Multi-F1 of Llama-3.2-3B by over 6% compared to Common, while maintaining a LCS-F1 gap of less than 1% against the Real data on Hunyuan-4B. 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 a privacy-preserving alternative to scarce clinical records.

488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515

## 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

516  
517  
518  
519  
520  
521  
522  
523  
524  
525

Table 2: 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 $\uparrow$
<b>Qwen3-4B-Instruct</b>	Base	8.04	11.24	<b>36.44</b>	87.03	<u>74.71</u>
	Common	12.46	14.82	32.32	88.42	74.14
	ASDAgent	<u>12.59</u>	<u>14.98</u>	33.21	<b>88.60</b>	73.94
	Real	<b>14.13</b>	<b>16.07</b>	<u>34.75</u>	<u>88.59</u>	<b>74.97</b>
<b>Qwen2.5-3B-Instruct</b>	Base	8.88	12.43	<b>33.78</b>	<u>88.18</u>	<b>73.56</b>
	Common	10.64	13.16	31.17	88.17	72.72
	ASDAgent	<u>10.75</u>	<u>13.68</u>	31.57	<b>88.52</b>	73.18
	Real	<b>11.81</b>	<b>14.04</b>	<u>31.71</u>	88.06	<u>73.50</u>
<b>Hunyuan-4B-Instruct</b>	Base	5.91	8.32	30.01	86.58	71.35
	Common	<u>9.97</u>	<u>13.04</u>	<u>30.96</u>	88.11	72.49
	ASDAgent	9.81	12.79	29.77	<b>88.38</b>	<u>72.71</u>
	Real	<b>11.74</b>	<b>14.18</b>	<b>31.80</b>	<u>88.15</u>	<b>73.51</b>
<b>Llama-3.2-3B-Instruct</b>	Base	8.56	11.03	27.42	87.54	72.01
	Common	9.60	12.96	29.24	<b>88.46</b>	72.56
	ASDAgent	<u>10.67</u>	<u>13.10</u>	<u>30.11</u>	<u>88.03</u>	<u>72.67</u>
	Real	<b>11.17</b>	<b>13.45</b>	<b>31.14</b>	87.91	<b>73.40</b>

Table 3: 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 $\uparrow$
<b>Qwen3-4B-Instruct</b>	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	<u>65.68</u>	<u>70.86</u>	<u>66.04</u>	<u>65.68</u>	<u>70.86</u>	<u>66.04</u>
	Real	<b>67.68</b>	<b>74.75</b>	<b>69.01</b>	<b>67.68</b>	<b>74.75</b>	<b>69.01</b>
<b>Qwen2.5-3B-Instruct</b>	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	<u>65.85</u>	<u>71.24</u>	<u>66.20</u>	<u>65.85</u>	<u>71.24</u>	<u>66.20</u>
	Real	<b>67.27</b>	<b>74.48</b>	<b>68.59</b>	<b>67.27</b>	<b>74.48</b>	<b>68.59</b>
<b>Hunyuan-4B-Instruct</b>	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	<b>66.64</b>	<u>71.46</u>	<u>66.66</u>	<b>66.64</b>	<u>71.46</u>	<u>66.66</u>
	Real	<u>66.47</u>	<b>72.92</b>	<b>67.52</b>	<u>66.47</u>	<b>72.92</b>	<b>67.52</b>
<b>Llama-3.2-3B-Instruct</b>	Base	57.23	58.77	55.88	57.23	58.77	55.88
	Common	59.29	67.21	61.23	59.29	67.21	61.23
	ASDAgent	<u>66.60</u>	<u>72.59</u>	<u>67.51</u>	<u>66.60</u>	<u>72.59</u>	<u>67.51</u>
	Real	<b>67.88</b>	<b>74.18</b>	<b>68.63</b>	<b>67.88</b>	<b>74.18</b>	<b>68.63</b>

a DOCTORAGENT that operationalizes rigorous ABA procedures via an explicit O-T-A-C reasoning loop, coupled with a probabilistic CHILDAgent 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

555	engagement and emotional regulation in children	not possess the legal or ethical authority to act as	605
556	with ASD. By relying solely on text, ASDAgent	an autonomous therapist. Therefore, ASDAgent	606
557	abstracts away these non-verbal signals, potentially	should be utilized strictly as a Clinical Decision	607
558	limiting its ability to detect subtle behavioral trig-	Support System (CDSS) or a training simulator.	608
559	gers or reinforce non-verbal communication mile-	Any application in a real intervention loop must ad-	609
560	stones.	here to a "Human-in-the-Loop" framework, where	610
561	<b>Simplification of Longitudinal Dynamics.</b>	professional therapists review all AI-generated sug-	611
562	While our ChildAgent simulates session-level be-	gestions to ensure safety, efficacy, and ethical com-	612
563	haviors (e.g., turn-taking, impulsivity), it does not	pliance.	613
564	yet fully model the long-term developmental trajec-		
565	tory of a child. In real therapy, a child's skills and		
566	interests evolve over months or years.		
	<b>Ethical Considerations</b>	<b>References</b>	614
567		Zara Abrams. 2025. Using generic ai chatbots for men-	615
568	<b>Data Privacy and Protection.</b> The protection of	tal health support: A dangerous trend. <i>American</i>	616
569	participant privacy is paramount, particularly given	<i>Psychological Association</i> .	617
570	the sensitive nature of clinical data involving chil-	Donald M Baer, Montrose M Wolf, and Todd R Risley.	618
571	children with ASD. Throughout the dataset construc-	1968. Some current dimensions of applied behav-	619
572	tion process, we implemented a rigorous, multi-	ior analysis. <i>Journal of applied behavior analysis</i> ,	620
573	layered de-identification protocol. This involved	1(1):91.	621
574	an initial pass of automated PII (Personally Identi-	Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo	622
575	fiable Information) scrubbing, followed by manual	Chen, Chuang Ding, Linhao Dong, Qianqian Dong,	623
576	verification to ensure the complete removal or ob-	Yujiao Du, Kepan Gao, and 1 others. 2024. Seed-	624
577	fuscation of sensitive attributes, including names,	asr: Understanding diverse speech and contexts	625
578	locations, and institutional references. Our dataset	with llm-based speech recognition. <i>arXiv preprint</i>	626
579	is released strictly for non-commercial research	<i>arXiv:2407.04675</i> .	627
580	purposes under a license that prohibits any attempt	Ariane VS Buescher, Zuleyha Cidav, Martin Knapp, and	628
581	to re-identify individuals.	David S Mandell. 2014. Costs of autism spectrum	629
582	<b>Ethics of Synthetic Data Generation.</b> We ac-	disorders in the united kingdom and the united states.	630
583	knowledge the ethical complexities inherent in sim-	<i>JAMA pediatrics</i> , 168(8):721–728.	631
584	ulating the behaviors of neurodivergent popula-	Wendy W Chapman, Prakash M Nadkarni, Lynette	632
585	tions. A primary concern is the potential for algo-	Hirschman, Leonard W D'avolio, Guergana K	633
586	rithmic stereotyping, where the generative model	Savova, and Ozlem Uzuner. 2011. Overcoming bar-	634
587	might oversimplify ASD phenotypes into repetitive	riers to nlp for clinical text: the role of shared tasks	635
588	or remaining silent, ignoring the high-functioning	and the need for additional creative solutions.	636
589	or "masking" traits often seen in real scenarios.	Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu	637
590	To mitigate this, our ChildAgent utilizes a prob-	Lian, and Zheng Liu. 2024. Bge m3-embedding:	638
591	abilistic behavioral mechanism rather than fixed,	Multi-lingual, multi-functionality, multi-granularity	639
592	caricature-like personas. However, users must rec-	text embeddings through self-knowledge distillation.	640
593	ognize that these synthetic dialogues are statistical	<i>arXiv preprint arXiv:2402.03216</i> .	641
594	approximations and not substitutes for the lived ex-	Xin-Yu Chen, Yu-Ming Chen, Chin-Po Chen, Bo-Hao	642
595	periences of real children. To ensure transparency	Su, Susan Shur-Fen Gau, and Chi-Chun Lee. 2025.	643
596	and prevent misinformation, all synthesized data	Socialrechnet: A multimodal llm-based framework	644
597	is explicitly watermarked or metadata-tagged to	for assessing social reciprocity in autism spectrum	645
598	distinguish it from authentic clinical records.	disorder. In <i>ICASSP 2025-2025 IEEE International</i>	646
599	<b>Clinical Applicability and Safety Scope.</b>	<i>Conference on Acoustics, Speech and Signal Process-</i>	647
600	While ASDAgent demonstrates high fidelity in sim-	<i>ing (ICASSP)</i> , pages 1–5. IEEE.	648
601	ulating intervention scenarios, we explicitly cau-	Gheorghe Comanici, Eric Bieber, Mike Schaeckermann,	649
602	tion against its immediate deployment in unsuper-	Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Mar-	650
603	vised clinical settings. The system lacks valida-	cel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and	651
604	tion through longitudinal clinical trials and does	1 others. 2025. Gemini 2.5: Pushing the frontier with	652
		advanced reasoning, multimodality, long context, and	653
		next generation agentic capabilities. <i>arXiv preprint</i>	654
		<i>arXiv:2507.06261</i> .	655

656	Vera Dekker, Maaïke H Nauta, Marieke E Timmerman,	Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan	711
657	Erik J Mulder, Lianne van der Veen-Mulders, Bar-	Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,	712
658	bara J van den Hoofdakker, Sjoukje van Warners,	Weizhu Chen, and 1 others. 2022. Lora: Low-rank	713
659	Leonieke JJ Vet, Pieter J Hoekstra, and Annelies	adaptation of large language models. <i>ICLR</i> , 1(2):3.	714
660	de Bildt. 2019. Social skills group training in chil-		
661	children with autism spectrum disorder: a randomized	Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong,	715
662	controlled trial. <i>European child &amp; adolescent psychi-</i>	Zhangyin Feng, Haotian Wang, Qianglong Chen,	716
663	<i>atry</i> , 28:415–424.	Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 oth-	717
		ers. 2025. A survey on hallucination in large lan-	718
664	Chengyun Deng, Shuzhong Lai, Chi Zhou, Mengyi Bao,	guage models: Principles, taxonomy, challenges, and	719
665	Jingwen Yan, Haifeng Li, Lin Yao, and Yueming	open questions. <i>ACM Transactions on Information</i>	720
666	Wang. 2024. Asd-chat: An innovative dialogue inter-	<i>Systems</i> , 43(2):1–55.	721
667	vention system for children with autism based on llm		
668	and vb-mapp. <i>arXiv preprint arXiv:2409.01867</i> .	Aaron Hurst, Adam Lerer, Adam P Goucher, Adam	722
		Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow,	723
669	FIFTH EDITION. 1980. Diagnostic and statistical man-	Akila Welihinda, Alan Hayes, Alec Radford, and 1	724
670	ual of mental disorders. <i>American psychiatric asso-</i>	others. 2024. Gpt-4o system card. <i>arXiv preprint</i>	725
671	<i>ciation</i> , Washington, DC, pages 205–224.	<i>arXiv:2410.21276</i> .	726
672	Annette Estes, Jeffrey Munson, Sally J Rogers, Jes-	Yejin Kim, Chi-Hyun Choi, Selin Cho, Jy-yong Sohn,	727
673	sica Greenson, Jamie Winter, and Geraldine Dawson.	and Byung-Hoon Kim. 2025. Aligning large lan-	728
674	2015. Long-term outcomes of early intervention in	guage models for cognitive behavioral therapy: a	729
675	6-year-old children with autism spectrum disorder.	proof-of-concept study. <i>Frontiers in Psychiatry</i> ,	730
676	<i>Journal of the American Academy of Child &amp; Adoles-</i>	16:1583739.	731
677	<i>cent Psychiatry</i> , 54(7):580–587.		
		Shuzhong Lai, Chenxi Li, Junhong Lai, Yucun Zhong,	732
678	Yi Feng, Mingyang Song, Jiaqi Wang, Zhuang Chen,	Chenyu Yan, Xiang Li, Haifeng Li, Gang Pan, Lin	733
679	Guanqun Bi, Minlie Huang, Liping Jing, and Jian	Yao, and Yueming Wang. 2025. Asd-illm: An in-	734
680	Yu. 2025. Ss-gen: A social story generation frame-	tervention large language model for autistic children	735
681	work with large language models. In <i>Proceedings</i>	based on real clinical dialogue intervention dataset.	736
682	<i>of the AAAI Conference on Artificial Intelligence</i> ,	In <i>Findings of the Association for Computational</i>	737
683	volume 39, pages 1300–1308.	<i>Linguistics: EMNLP 2025</i> , pages 8058–8079.	738
684	Richard M Foxx. 2008. Applied behavior analysis treat-	Alon Lavie and Abhaya Agarwal. 2007. <b>METEOR: An</b>	739
685	ment of autism: The state of the art. <i>Child and adoles-</i>	<b>automatic metric for MT evaluation with high levels</b>	740
686	<i>cent psychiatric clinics of North America</i> , 17(4):821–	<b>of correlation with human judgments</b> . In <i>Proceed-</i>	741
687	834.	<i>ings of the Second Workshop on Statistical Machine</i>	742
		<i>Translation</i> , pages 228–231, Prague, Czech Republic.	743
688	Elizabeth A Fuller and Ann P Kaiser. 2020. The effects	Association for Computational Linguistics.	744
689	of early intervention on social communication out-		
690	comes for children with autism spectrum disorder: A	Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao,	745
691	meta-analysis. <i>Journal of autism and developmental</i>	and Bill Dolan. 2016. <b>A diversity-promoting ob-</b>	746
692	<i>disorders</i> , 50(5):1683–1700.	<b>jective function for neural conversation models</b> . In	747
		<i>Proceedings of the 2016 Conference of the North</i>	748
693	Ethan Goh, Robert Gallo, Jason Hom, Eric Strong,	<i>American Chapter of the Association for Computa-</i>	749
694	Yingjie Weng, Hannah Kerman, Joséphine A Cool,	<i>tional Linguistics: Human Language Technologies</i> ,	750
695	Zahir Kanjee, Andrew S Parsons, Neera Ahuja, and 1	pages 110–119, San Diego, California. Association	751
696	others. 2024. Large language model influence on	for Computational Linguistics.	752
697	diagnostic reasoning: a randomized clinical trial.		
698	<i>JAMA network open</i> , 7(10):e2440969–e2440969.	Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingx-	753
		uan Wang, Bingzheng Xu, Bochao Wu, Bowei	754
699	A. Grattafiori and 1 others. 2024. The llama 3 herd of	Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025.	755
700	models. <i>arXiv preprint arXiv:2407.21783</i> . <a href="https://arxiv.org/abs/2407.21783">https://arxiv.org/abs/2407.21783</a> .	Deepseek-v3. 2: Pushing the frontier of open large	756
701		language models. <i>arXiv preprint arXiv:2512.02556</i> .	757
702	Joschka Haltaufderheide and Robert Ranisch. 2024.	Michael V Lombardo, Meng-Chuan Lai, and Simon	758
703	The ethics of chatgpt in medicine and healthcare:	Baron-Cohen. 2019. Big data approaches to de-	759
704	a systematic review on large language models (llms).	composing heterogeneity across the autism spectrum.	760
705	<i>NPJ digital medicine</i> , 7(1):183.	<i>Molecular psychiatry</i> , 24(10):1435–1450.	761
706	R Hanrahan, E Smith, H Johnson, A Constantin, and	O Ivar Lovaas. 1987. Behavioral treatment and nor-	762
707	M Brosnan. 2020. A pilot randomised control trial of	mal educational and intellectual functioning in young	763
708	digitally-mediated social stories for children on the	autistic children. <i>Journal of consulting and clinical</i>	764
709	autism spectrum. <i>Journal of autism and developmen-</i>	<i>psychology</i> , 55(1):3.	765
710	<i>tal disorders</i> , 50:4243–4257.		

766	Aishik Mandal, Tanmoy Chakraborty, and Iryna Gurevych. 2025. Towards privacy-aware mental health ai models: Advances, challenges, and opportunities. <i>arXiv preprint arXiv:2502.00451</i> .	819
767		820
768		821
769		822
770	Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. <i>arXiv preprint arXiv:2303.13375</i> .	823
771		824
772		825
773		826
774	US Department of Health, Human Services, and 1 others. 2005. Other requirements relating to uses and disclosures of protected health information.	827
775		828
776		829
777	OpenAI. 2025. Gpt-5.1. <a href="https://platform.openai.com/docs/models">https://platform.openai.com/docs/models</a> .	830
778		831
779	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In <i>Proceedings of the 40th annual meeting of the Association for Computational Linguistics</i> , pages 311–318.	832
780		833
781		834
782		835
783		836
784	Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2023. Discovering language model behaviors with model-written evaluations. In <i>Findings of the association for computational linguistics: ACL 2023</i> , pages 13387–13434.	837
785		838
786		839
787		840
788		841
789		842
790		843
791	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). <i>Cochrane database of systematic reviews</i> , (10).	844
792		845
793		846
794		847
795		848
796	Henry S Roane, Wayne W Fisher, and James E Carr. 2016. Applied behavior analysis as treatment for autism spectrum disorder. <i>The Journal of pediatrics</i> , 175:27–32.	849
797		850
798		851
799		852
800	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. <i>JMIR Mental Health</i> , 12(1):e69709.	853
801		854
802		855
803		856
804		857
805	Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. <i>arXiv preprint arXiv:2310.13548</i> .	858
806		859
807		860
808		861
809		862
810		863
811	Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. <i>Nature</i> , 620(7972):172–180.	864
812		865
813		866
814		867
815		868
816	Tristram Smith. 2001. Discrete trial training in the treatment of autism. <i>Focus on autism and other developmental disabilities</i> , 16(2):86–92.	869
817		870
818		871
		872
		873
	Mark L Sundberg. 2008. <i>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</i> . 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 <i>Findings of the Association for Computational Linguistics: ACL 2024</i> , pages 599–621.	
	Qwen Team. 2025. <i>Qwen3 technical report</i> . Preprint, arXiv:2505.09388.	
	U.S. Department of Health and Human Services. 2025. Methods for de-identification of phi. <a href="https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html">https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html</a> .	
	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. <i>Learning and Instruction</i> , 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. <i>Clinical psychology review</i> , 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. <a href="https://github.com/huggingface/trl">https://github.com/huggingface/trl</a> .	
	Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. 2025a. Capabilities of gpt-5 on multimodal medical reasoning. <i>arXiv preprint arXiv:2508.08224</i> .	
	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? <i>arXiv preprint arXiv:2502.11211</i> .	
	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. <i>Advances in neural information processing systems</i> , 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. <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations</i> .	

874	Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, and 1 others. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. <i>arXiv preprint arXiv:1609.08144</i> .	930
875		931
876		932
877		933
878		
879		
880	An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. <i>arXiv preprint arXiv:2407.10671</i> .	
881		
882		
883		
884		
885		
886		
887	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. <i>Advances in neural information processing systems</i> , 36:11809–11822.	
888		
889		
890		
891		
892	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 <i>The eleventh international conference on learning representations</i> .	
893		
894		
895		
896		
897	Seungwon Yoon and 1 others. 2023. Towards standardizing korean grammatical error correction. In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)</i> .	
898		
899		
900		
901	Jinan Zeidan, Eric Fombonne, Julie Scolah, Alaa Ibrahim, Maureen S Durkin, Shekhar Saxena, Afqah Yusuf, Andy Shih, and Mayada Elsabbagh. 2022. Global prevalence of autism: A systematic review update. <i>Autism research</i> , 15(5):778–790.	
902		
903		
904		
905		
906	T. Zhang and 1 others. 2024. Improving diversity of commonsense generation by large language models. In <i>Findings of the Association for Computational Linguistics: EMNLP 2024</i> .	
907		
908		
909		
910	Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with bert. In <i>International Conference on Learning Representations</i> .	
911		
912		
913		
914	Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. <i>arXiv preprint arXiv:2506.05176</i> .	
915		
916		
917		
918		
919		
920	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. <i>Psychiatric Services</i> , 71(4):385–388.	
921		
922		
923		
924	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. <i>Advances in neural information processing systems</i> , 36:46595–46623.	
925		
926		
927		
928		
929		
	Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. <i>arXiv preprint arXiv:1802.01886</i> .	
	<b>A Case Study</b>	934
	<b>A.1 Case Study in Dialogue Synthesis</b>	935
	Figure 8 shows examples of real and synthetic dialogues on the same topic and with the same number of rounds.	936
		937
		938
	<b>A.2 Case Study in Real Autism Intervention</b>	939
	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.	940
		941
		942
		943
	<b>B ABA Strategy and Response Type</b>	944
	<b>B.1 ABA Strategy</b>	945
	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: <i>Instruction, Reinforcement, Half-Assistance, Full-Assistance, Other</i> and <i>Pause</i> . The child’s response types are categorized as: <i>Irrelevant, Relevant, Repetitive</i> , and <i>Unresponsive</i> .	946
		947
		948
		949
		950
		951
		952
		953
		954
		955
	<b>Instruction</b> 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.	956
		957
		958
		959
		960
	<b>Reinforcement</b> 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 socialization through verbal praise and empathy.	961
		962
		963
		964
		965
		966
		967
		968
		969
		970
	<b>Assistance</b> 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	971
		972
		973
		974
		975
		976
		977

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

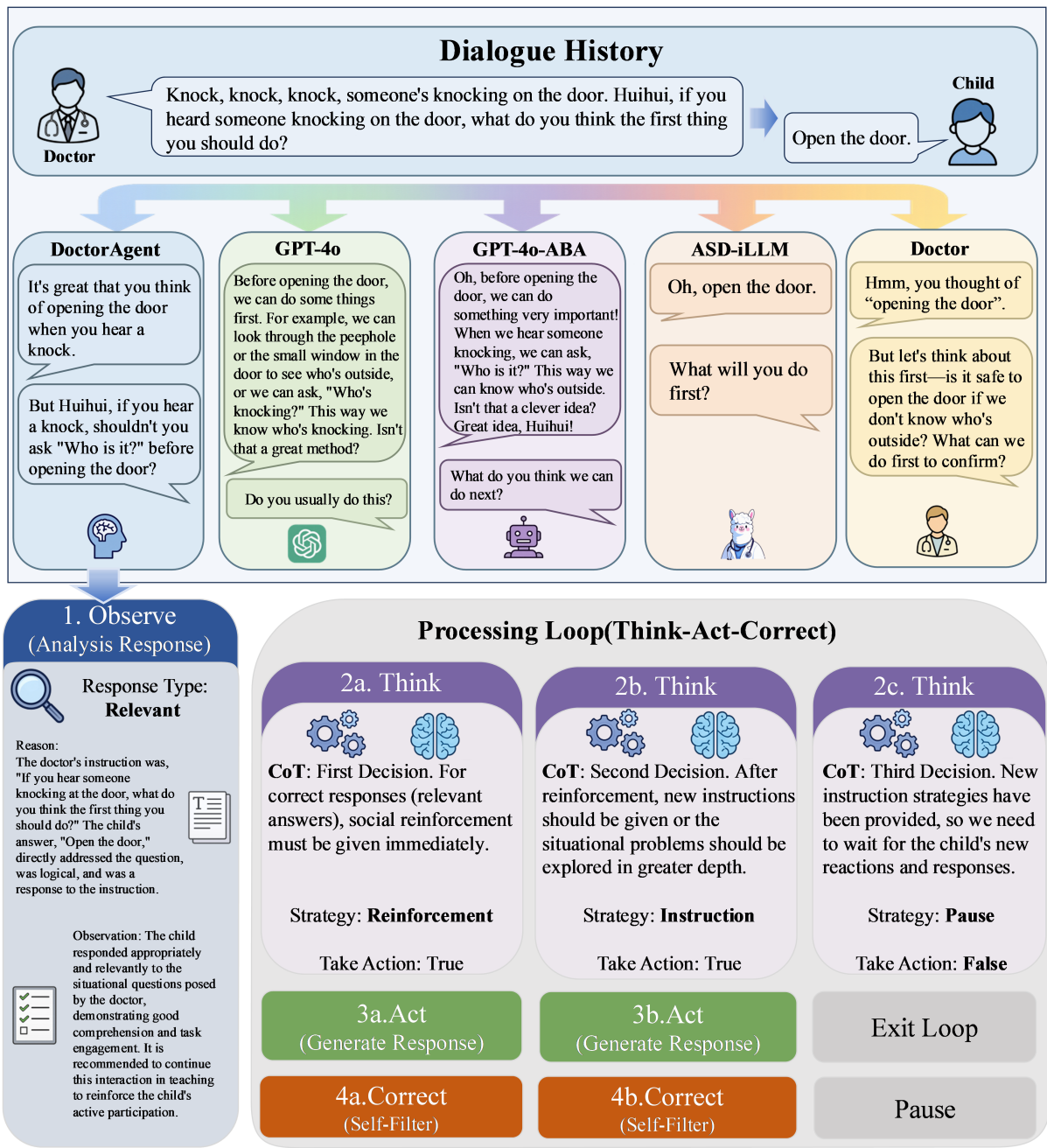


Figure 9: Case Study in Real Autism Intervention. The diagram above illustrates the intervention responses of DOCTORAGENT, the real Doctor, and other models based on a realistic intervention dialogue. The diagram below shows how DOCTORAGENT completes the O-A-T-C process.

978 avoid causing feelings of failure in the child or  
979 creating dependence on the prompts. In thematic  
980 conversation intervention, Assistance usually takes  
981 the form of verbal prompts, such as rephrasing  
982 questions, breaking down questions, or providing  
983 hints to the answer.

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

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

992 **Full-Assistance**, on the other hand, involves  
993 the therapist directly providing clear demonstra-  
994 tions or complete answers when the child can-  
995 not understand the instructions or shows no re-  
996 sponse, guiding the child to imitate or repeat the  
997 correct response. By flexibly using partial and Full-  
998 Assistance at different stages, therapists can ensure  
999 the success rate of the intervention while gradually  
1000 improving the child’s independent response ability.

1001 **Pause** refers to the brief interval between each  
1002 trial, allowing the child time to reflect on and inter-  
1003 nalize their response and the reinforcement.

## 1004 B.2 Child Response Type

1005 **Relevant** responses refer to children’s answers that  
1006 semantically or functionally match the instructions  
1007 or questions given by the doctor, indicating that the  
1008 child understands the current topic and can respond  
1009 appropriately;

1010 **Irrelevant** responses refer to children’s answers  
1011 that have no clear connection to the current instruc-  
1012 tions or topic, possibly reflecting attention shifts,  
1013 comprehension difficulties, or language organiza-  
1014 tion problems;

1015 **Repetitive** responses refer to children simply  
1016 repeating the doctor’s words or their own previous  
1017 expressions without providing new information or  
1018 independent responses, usually reflecting imitative  
1019 behavior or limitations in response strategies;

1020 **Unresponsive** responses refers to the child not  
1021 giving any verbal or non-verbal response within a  
1022 reasonable waiting time, which may be related to  
1023 comprehension difficulties, avoidance behavior, or  
1024 emotional state.

## C Details for ASDAgent-Dataset 1025

1026 Currently, there are no publicly available datasets  
1027 for ASD dialogue intervention. Therefore, we cre-  
1028 ated a multi-turn dialogue dataset for interventions  
1029 between doctors and children with ASD, named  
1030 **ASDAgent-Dataset**.

### C.1 Data Collection 1031

1032 To ensure the authenticity and quality of the data,  
1033 we collaborated with five treatment centers for  
1034 autistic children after obtaining ethical approval.  
1035 With full informed consent from both parents and  
1036 children, audio recordings were collected during  
1037 topic-based dialogue interventions using a portable  
1038 recording device (H1-Pro, iFlytek Inc., China). To  
1039 ensure clear audio capture, the recorder was placed  
1040 in the chest pocket of the doctor’s coat.

1041 Given that autistic children often experience de-  
1042 lays in language development, chronological age  
1043 does not necessarily reflect actual language ability.  
1044 Therefore, only children with a language develop-  
1045 mental age greater than 24 months were included in  
1046 the study. Previous studies have shown that topic-  
1047 based dialogue interventions can effectively allevi-  
1048 ate social impairments in autistic children (Dekker  
1049 et al., 2019; Hanrahan et al., 2020; van der Wilt  
1050 et al., 2022). Accordingly, all recordings were con-  
1051 ducted in the form of structured topic dialogues,  
1052 with each recording focusing on a single predefined  
1053 topic. All audio recordings were sampled at 16,000  
1054 Hz and stored in WAV format.

### C.2 Data Processing 1055

1056 We employed a three-stage processing method to  
1057 transcribe the original audio recordings into multi-  
1058 dialogue text and annotate the doctors’ strategies  
1059 and the children’s response types.

1060 **Automatic Transcription** First, we utilized  
1061 existing automated transcription tools SEED-  
1062 ASR(Bai et al., 2024) to convert the original record-  
1063 ings into multi-turn dialogues.

1064 **Manual Transcription** Our goal is to improve  
1065 the quality of multi-turn dialogue text through man-  
1066 ual transcription. Building upon (Lai et al., 2025),  
1067 we annotated the data using crowdsourcing. Details  
1068 about crowdsourcing can be found in the Appendix  
1069 C.5.

1070 **State Annotation** According to the ABA(Foxx,  
1071 2008; Roane et al., 2016), we performed more de-  
1072 tailed data annotation on the selected high-quality  
1073 dialogues, including annotating the doctor’s strate-

1074  
1075  
1076

gies and the child’s response types using ABA and DTT. The basic flow of DTT is illustrated in Figure 10.

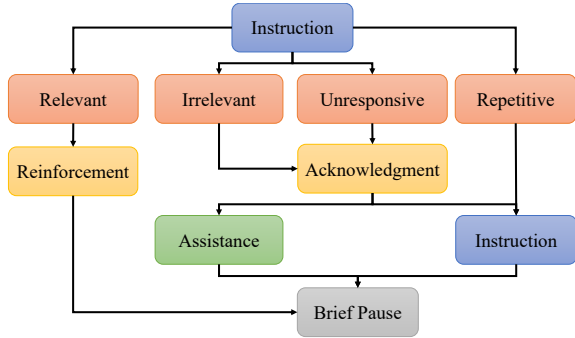


Figure 10: The workflow of DTT from ABA. Doctors can adjust their treatment strategies as needed, based on the actual intervention situation.

1077

### C.3 Data Cleaning

1078  
1079  
1080

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

1081  
1082  
1083  
1084

- We removed multi-turn dialogue texts with fewer than five exchanges. Dialogues with too few exchanges fail to reflect the doctor’s intervention strategies adequately.

1085  
1086  
1087  
1088  
1089  
1090  
1091

- Dialogues focused on entities, such as storybooks or toys, were removed. The model requires visual comprehension to understand the images or entities referenced in these multi-turn dialogues. Currently, our focus is on the model’s dialogue style and intervention strategies.

1092  
1093  
1094  
1095  
1096  
1097

- 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.

1098

### C.4 ASDAgent-Dataset

1099  
1100  
1101  
1102  
1103

**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  $\mathcal{D}_{golden}$ .

1104  
1105

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

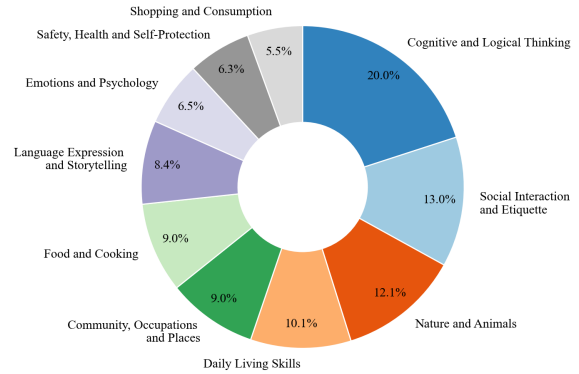


Figure 11: Topic distribution of ASDAgent-Dataset-Golden.

### C.5 Crowdsourcing

1106

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 4.

1107  
1108  
1109  
1110  
1111  
1112  
1113

Table 4: 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.

1114  
1115  
1116  
1117

### C.6 Topic Classification

1118

The topic distribution of ASDAgent-Dataset-Golden  $\mathcal{D}_{golden}$  is illustrated in Figure 11, showing a balanced distribution of topics.

1119  
1120  
1121

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.

1122  
1123  
1124  
1125  
1126

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

1127  
1128  
1129  
1130

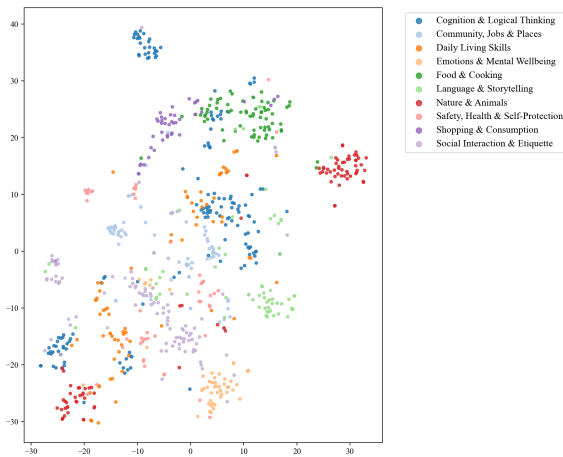


Figure 12: t-SNE scatter plot of macro topics across 10 main conversational categories.

to obtain the final 10 topic categories as shown in Figure 12.

### C.7 Children’s statistics

**Demographic Details** The demographic information of children in ASDAgent-Dataset-Golden is presented in Table 7, 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 chronological 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 5. 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 6, 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.

### C.8 ASD Children Heterogeneity

Based on the behavioral profiles exhibited by different children as reflected in their performance on the Table 5 and 6, 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.

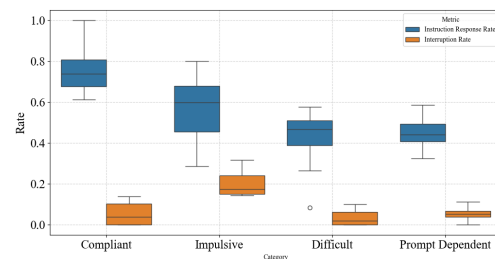


Figure 13: Distribution of Key Metrics by Child Category

The box plot and scatter plot are shown in the Figure 13 and 14, which also provides a basis for ChildAgent to adapt to personalized persona modeling.

### C.9 Utterance Length

Statistical information for the ASDAgent-Dataset-Golden is shown in Table 8 and 9. On average, each conversational turn lasts 18.61 rounds. Furthermore, during the intervention, both the doctor

Table 5: 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 6: Turn interruption probabilities following the current dialogue turn.

Next Speaker	Probability (%)
Child	8.67
Doctor	91.33

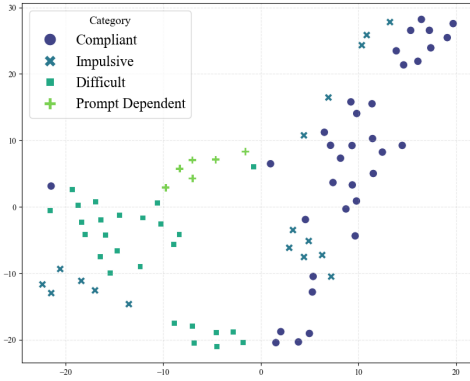


Figure 14: t-SNE Clustering of Child Profiles

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.

### C.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 15, 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  $\mu$  and  $\sigma$  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.

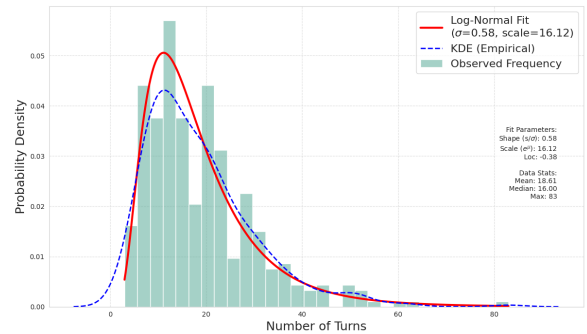


Figure 15: Distribution of Conversation Turns (Log-Normal Fit)

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

Table 8: 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 9: 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

## D 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 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 H.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.

## E Evaluation Metrics

### E.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 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.

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 E.4 and E.5.

## E.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 11. 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 10. 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 10 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.

## E.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. 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 11 and 12 show the evaluation criteria in Evaluation 1 and Evaluation 2.

## E.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:

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

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

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

## E.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

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

Table 10: Information for experts involved in human evaluation.

<b>Dimension</b>	<b>Category</b>	<b>Explanation</b>
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 11: 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.

<b>Dimension</b>	<b>Category</b>	<b>Explanation</b>
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 12: 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.

1390 model predicts the correct strategies but the order  
1391 is completely wrong, this metric will be low.

1392 Let  $LCS(A, B)$  be the Longest Common Subse-  
1393 quence of sequences  $A$  and  $B$ , and  $|LCS(A, B)|$   
1394 be the length of this subsequence.

1395 First, calculate the match length:

$$1396 \text{Match}_{\text{seq}} = |LCS(S_{\text{pred}}, S_{\text{ref}})| \quad (18)$$

1397 Based on this, calculate the ordered Precision,  
1398 Recall, and F1 score:

$$1399 \text{Precision}_{\text{seq}} = \frac{\text{Match}_{\text{seq}}}{|S_{\text{pred}}|} \quad (19)$$

$$1400 \text{Recall}_{\text{seq}} = \frac{\text{Match}_{\text{seq}}}{|S_{\text{ref}}|} \quad (20)$$

$$1401 \text{F1}_{\text{seq}} = \frac{2 \cdot \text{Precision}_{\text{seq}} \cdot \text{Recall}_{\text{seq}}}{\text{Precision}_{\text{seq}} + \text{Recall}_{\text{seq}}} \quad (21)$$

## 1402 F Meta-Evaluation: Human-LLM 1403 Alignment

1404 To validate the reliability of automated evaluation,  
1405 we calculated the agreement and correlation be-  
1406 tween three LLM judges (DeepSeek-V3.2, GPT-  
1407 5.1, Gemini-2.5) and human experts on a subset of  
1408 46 randomly sampled dialogues.

1409 As shown in Table 13, although LLM and hu-  
1410 man assessments are not entirely consistent, this  
1411 confirms the high fidelity of our synthetic data:

1412 DeepSeek-V3.2 as the Most Reliable Judge:  
1413 Among the candidates, DeepSeek-V3.2 achieved  
1414 the highest agreement with human experts (Accu-  
1415 racy: 52.2%,  $\kappa = 0.288$ ), identifying the superior-  
1416 ity of real data in 21.7% of cases while maintaining  
1417 a moderate correlation ( $\rho = 0.40$ ). This indicates  
1418 its capability to capture clinical nuances.

1419 The "Tie Bias" Phenomenon: Conversely, GPT-  
1420 5.1 and Gemini-2.5 exhibited a near-total inability  
1421 to distinguish synthetic from real data, predicting  
1422 "Tie" in 80.4% and 84.8% of cases, respectively.  
1423 This resulted in near-zero or negative Kappa scores.

1424 Validation of Synthesis Quality: While this lim-  
1425 its the utility of GPT/Gemini as discriminators,  
1426 it paradoxically validates the high fidelity of our  
1427 synthetic data. The generated dialogues are suf-  
1428 ficiently natural and strategic to render them in-  
1429 distinguishable from human therapist outputs for  
1430 general-purpose SOTA models.

## 1431 G Ablation Study

### 1432 G.1 ASDAgent for Data Synthesis

1433 **Automatic Evaluation.** Table 14 shows the di-  
1434 versity of language used by children and doctors  
1435 in the dialogue; Table 15 and Table 16 shows the  
1436 proportion of strategies or response types used by  
1437 children and doctors in the dialogue. Table 17 and  
1438 Table 18 shows the average length of responses  
1439 from children and doctors in the dialogue.

1440 Removing DOCTORAGENT reveals significant  
1441 strategy collapse and linguistic abnormalities, ex-  
1442 cessively high proportion of instructions and ab-  
1443 normal sentence length. Removing CHILDAAGENT,  
1444 while showing better performance on some diver-  
1445 sity metrics for DOCTORAGENT, reveals a devia-  
1446 tion from reality in its strategy distribution (insuffi-  
1447 cient reinforcement), and children tend to produce  
1448 excessively long and irrelevant/repetitive responses.  
1449 From the perspective of "rationality of intervention  
1450 behavior," it is less stable than ASDAGENT. There-  
1451 fore, in the Evaluation 2, we believe that ASDA-  
1452 GENT best reproduces realistic clinical interaction  
1453 patterns and is the most suitable source of high-  
1454 quality synthetic dialogues.

1455 **LLM Evaluation.** Additionally, we conduct an  
1456 ablation study using LLM-based evaluators to in-  
1457 vestigate the relative contributions of doctor mod-  
1458 eling and child modeling to intervention dialogue  
1459 quality of ASDAgent according to Table 12.

1460 Table 19 presents ablation results across three  
1461 LLM evaluators. Removing the CHILDAAGENT con-  
1462 sistently causes substantial degradation in profes-  
1463 sionalism (A), with relative drops of 19.8%–26.9%.  
1464 This decline is mainly attributed to reduced adher-  
1465 ence to DTT/NET dialogue principles (A1) and less  
1466 coherent ABA strategy sequencing (A2), as well  
1467 as weaker personalized adjustments (A3) to child  
1468 responses. These results highlight the necessity  
1469 of child-aware modeling for clinically appropriate  
1470 interventions.

1471 Removing the DOCTORAGENT also leads to no-  
1472 table performance drops, particularly in profes-  
1473 sionalism (A) and scenario complexity (C), in-  
1474 dicating impaired instructional structure and re-  
1475 duced use of effective teaching dynamics (e.g., cor-  
1476 rective loops). In contrast, child realism (B) ex-  
1477 hibits smaller changes and occasionally improves,  
1478 suggesting that surface-level linguistic plausibility  
1479 alone is insufficient to ensure intervention quality.  
1480 Overall, the consistent decline in Total score con-

Table 13: Meta-Evaluation Results: Alignment between LLM Judges and Human Experts.

Model	Accuracy	Cohen’s $\kappa$	Spearman’s $\rho$	Kendall’s $\tau$
<b>DeepSeek-V3.2</b>	<b>0.522</b>	<b>0.288</b>	<b>0.396</b>	<b>0.381</b>
GPT-5.1	0.283	-0.024	-0.144	-0.136
Gemini-2.5-Pro	0.261	-0.048	0.024	0.023
LLM Ensemble	0.283	-0.003	0.039	0.037

Table 14: Diversity Metrics for Doctors and Children Across Different Sources

Doctor	Child	D-2 $\uparrow$	D-3 $\uparrow$	S-BLEU $\downarrow$	S-GLEU $\downarrow$	S-BERTScore $\downarrow$
<b>Doctor Part</b>						
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	<b>0.277</b>	<b>0.519</b>	<b>0.661</b>	0.364	<b>0.599</b>
BaseDoctor	ChildAgent	0.187	0.412	0.731	<b>0.318</b>	0.612
<b>Child Part</b>						
Real	Real	0.499	0.732	0.423	0.483	0.607
DoctorAgent	ChildAgent	<b>0.428</b>	<b>0.667</b>	<b>0.477</b>	<b>0.522</b>	0.622
DoctorAgent	BaseChild	0.383	0.575	0.552	0.557	<b>0.609</b>
BaseDoctor	ChildAgent	0.400	0.599	0.540	0.531	0.615

Table 15: Distribution of Doctor and Child Interaction Strategies (in %) 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	<b>0.083</b>	<b>0.019</b>
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 16: Distribution of Child Response Types (in %) 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	<b>0.007</b>	<b>0.002</b>
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 17: Doctor Utterance Length by Intervention Strategy (Mean $\pm$ Std)

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

Table 18: Child Utterance Length by Response Type (Mean $\pm$ Std)

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

1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530

firm the complementary importance of both doctor and child modeling.

## G.2 ASDAgent for Clinical Intervention

**Automatic Evaluation** From Table 20, 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 H.1.

Table 21 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 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 22, 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.

## H Prompt

### H.1 Base and ABA prompt

Figure 16 and 17 show the prompt used in clinical intervention under BASE and ABA settings.

### H.2 Prompt for Strategy Labeling

Figure 18 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).

### H.3 Prompt for DoctorAgent: Observe

Figure 19 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:

1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575

Table 19: 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%

Table 20: Ablation Study on Lexical, Semantic, and Diversity Metrics. For each model, the best result under each metric is highlighted in bold.

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

Table 21: Ablation Study on Strategy Consistency Metrics (in %). For each model, the best result under each metric is highlighted in bold.

Model	Setting	Multi-P ↑	Multi-R ↑	Multi-F1 ↑	LCS-P ↑	LCS-R ↑	LCS-F1 ↑
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	<b>70.78</b>	<b>79.82</b>	<b>72.95</b>	<b>70.78</b>	<b>79.82</b>	<b>72.95</b>
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	<b>67.38</b>	<b>79.28</b>	<b>70.47</b>	<b>67.30</b>	<b>79.19</b>	<b>70.38</b>

System Prompt
<p><b>## Role Setting</b></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><b>## Starting the Conversation</b></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 16: Base prompt

Table 22: 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	<b>10.39</b>
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	<b>10.12</b>
	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	<b>10.22</b>
	GPT-4o	DoctorAgent	2.64	3.51	3.95	10.10

System Prompt
<p><b>## Role Setting</b></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><b>## Intervention Principles</b></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"> <li>- Instruction - Provide clear and simple instructions to guide the child in a conversation around the topic.</li> <li>- Assistance - Provide appropriate verbal support when the child needs help to facilitate correct responses.</li> <li>- Reinforcement - Provide timely positive feedback and praise to encourage correct and positive behaviors.</li> <li>- Acknowledgment - Appropriately accept the child's incorrect responses.</li> </ul> <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><b>## Starting the Conversation</b></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 17: ABA prompt

<b>System Prompt</b>	
<b>## Role Setting</b>	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.
<b>## Input Data</b>	1. Dialogue History (Context): The previous dialogue context. 2. Doctor's Response (Target): The complete utterance spoken by the doctor in the current turn.
<b>## Strategy Definitions</b>	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.
<b>## Critical Constraints</b>	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..

Figure 18: Prompt for Strategy Labeling

1576	• <b>Response Classification:</b> The model must rigorously distinguish between Functional Communication (Related Response) and Echolalia (Repetition/Mechanical imitation), a distinction critical for assessing ASD communicative progress.	1602
1577		1603
1578		1604
1579		1605
1580		1606
1581		
1582	• <b>Functional Hypothesis:</b> The model infers the underlying motivation for the child's behavior (e.g., Escape/Avoidance, Sensory Stimulation, or Access to Attention).	1607
1583		1608
1584		1609
1585		
1586	• <b>Internal State Estimation:</b> 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.	1610
1587		1611
1588		1612
1589		1613
1590		
1591		
1592	<b>H.4 Prompt for DoctorAgent: Think</b>	1614
1593	Figure 20 illustrates a structured CoT prompt that guides the agent through a four-stage reasoning process $C_t$ the reasoning trace $C_t$ consists of:	1615
1594		1616
1595		1617
1596	• <b>Contextual Anchoring.</b> 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 $O_t$ .	1618
1597		1619
1598		1620
1599		1621
1600		1622
1601		1623
		1624
		1625
		1626
		1627
		1628

<b>System Prompt</b>
<p><b>## Role Setting</b>            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><b>## Core Classification Criteria</b>            Please carefully distinguish between "Repetition" and "Related Response":            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>
<p><b>## Analysis Dimensions</b>            Please analyze from the following three dimensions:            1. Response Quality :            - 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 :            - Why did the child have this reaction?            - Common functions: Access to tangible items/attention, Escape/avoidance of tasks, Self-stimulation (Sensory).            3. State Inference :            - Stress : Low (Calm) / Medium (Anxious) / High (On the verge of breakdown).            - Engagement : High (Focused) / Medium (Distracted) / Low (Disengaged).</p>

Figure 19: Prompt for DoctorAgent: Observe

viding 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:

- **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.

## H.6 Prompt for ChildAgent: Act

Figures 26, 27, and 28 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 26):** 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 27):** 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 28):** 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

## 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 20: Prompt for DoctorAgent: Think

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

Figure 21: Prompt for DoctorAgent: Act in Strategy Instruction

<b>System Prompt</b>
<p><b>## Role Setting</b></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 <b>**Half-Assistance**</b> 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><b>## Core Principles: Atomic Actions</b></p> <ul style="list-style-type: none"> <li>- If the strategy is "Half-Assistance", you are <b>**only responsible for providing assistance**</b>, and <b>**absolutely do not**</b> include other strategies in the same sentence, such as giving new instructions.</li> <li>- The next strategy will be generated by the system in the next cycle; you <b>**do not**</b> need to worry about it now.</li> </ul>
<p><b>## Executing the Strategy</b></p> <ol style="list-style-type: none"> <li>1. Please follow ABA principles in the analysis and decision-making process, combining Discrete Trial Training (DTT) and Natural Environment Teaching (NET) principles.</li> <li>2. Half-Assistance <ul style="list-style-type: none"> <li>- Provide moderate verbal support when the child needs help to facilitate correct responses.</li> <li>- This generally involves using questions rather than statements.</li> <li>- 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.</li> </ul> </li> <li>3. Principles for handling different types of child responses: <ul style="list-style-type: none"> <li>- When the child responds correctly (relevant answer), reinforcement should be given first;</li> <li>- When the child does not respond, appropriate partial or full assistance should be given first to facilitate a correct response;</li> <li>- 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.</li> <li>- 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.</li> </ul> </li> <li>4. When executing one strategy, do not mix in content from other strategies, emphasizing atomic operations.</li> <li>5. You need to consider the conversation topic and the historical conversation content to ensure your response is relevant and coherent within the context!</li> <li>6. Your responses must be consistent with the child's actual information, such as their age and language development stage.</li> </ol>

Figure 22: Prompt for DoctorAgent: Act in Strategy Half-Assistance

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

Figure 23: Prompt for DoctorAgent: Act in Strategy Full-Assistance

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

Figure 24: 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 25: Prompt for DoctorAgent: Act in Strategy Reinforcement

1686

(perseverating on the child's own previous words),

1687

accurately reflecting the rigid behavioral patterns

1688

observed in ASD.

```

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 26: 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 27: 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 28: Prompt for ChildAgent: Act in Type Repetitive Response