Expert Systems With Applications 298 (2026) 129712

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Language proficiency assessment of autistic children using large language models

Saige Qin [a,b,c], Min Liu [a,c], Tongquan Wei [b], Qiaoyun Liu [a,c,*]

[a] Shanghai Institute of Artificial Intelligence for Education, East China Normal University, Shanghai, 200062, China
[b] School of Computer Science and Technology, East China Normal University, Shanghai, 200062, China
[c] Innovation Lab of Education and Rehabilitation, Faculty of Education, East China Normal University, Shanghai, 200062, China

A R T I C L E I N F O

Keywords:
Large language models
Autism spectrum disorder
Language proficiency assessment
Mullen Scales of Early Learning
eXtreme Gradient Boosting

A B S T R A C T

Language impairment is a common comorbidity in children with autism spectrum disorder (ASD), and language proficiency assessment is a primary method for identifying such impairments. However, traditional assessment tools are often subjective and inefficient, while existing computer-assisted methods are limited by a narrow focus and insufficient use of natural language samples. To address these issues, this study proposes a framework for assessing children's language abilities based on large language models (LLMs). We first preprocess the natural language samples from children and design multiple assessment dimensions and workflows. To enhance the stability of the assessment, we introduce a multi-expert voting mechanism and perform a comparative analysis of various large language models' performance. The experimental results demonstrate a strong correlation between the framework's assessment results and the Mullen Scales of Early Learning (MSEL) verbal developmental quotients, with a Pearson correlation coefficient of 0.8 ($p < 0.001$). Furthermore, the results show that the multi-dimensional evaluation can accurately differentiate between ASD and typically developing (TD) children, achieving a classification accuracy of 0.98. These findings suggest that the proposed framework has significant potential for improving the accuracy of ASD identification.

1. Introduction

Language disorder refers to a condition where a child's language comprehension and expression are significantly below the language level of their peers (Liu, 2019). This disorder is prevalent among children with autism spectrum disorder (ASD) and is the primary reason for the initial medical consultation in over 80 % of children with ASD (Wenqian, 2018). Research indicates that 25–35 % of individuals with ASD exhibit minimal speech production, and even individuals with ASD who have basic verbal communication skills often exhibit syntactic deficits (Boucher, 2012; Rose et al., 2016). These abnormal features of language disorders provide important clues for the screening and diagnosis of children with ASD. Therefore, early detection of language disorders creates favorable conditions for early identification and intervention for children with ASD.

Language proficiency assessment plays a key role as one of the main methods to detect language disorders. Traditional language proficiency assessments rely on clinical assessment scales, questionnaires, and structured interviews. Whereas these methods provide direct assessments of language ability, the process relies on professional analysis, which can be both subjective and time-consuming. In recent years, to enhance the objectivity and efficiency of language proficiency assessment, numerous researchers have begun exploring innovative approaches that incorporate machine learning techniques. For example, the LENA (Language Environment Analysis) system (Gilkerson & Richards, 2008) analyzes children's vocal behaviors using automatic speech processing technologies and is widely used to quantify both language input and output. Other approaches include online multimedia language assessment systems (Li et al., 2025), automated classification methods for language ability that combine automatic speech recognition with neural networks (Gretter et al., 2019), and end-to-end frameworks for scoring expressive language (Gale et al., 2020). In recent years, large language models (LLMs) have emerged as transformative technologies in the assessment of children's language and developmental abilities, demonstrating promising potential. Existing studies have employed LLMs to predict children's language development stages (Feng et al., 2024) and to assess multidimensional abilities such as cognition, social interaction, and emotional development (Yang et al., 2025), highlighting the broad applicability of LLMs in this domain. Despite the progress made in automating language assessments, several challenges persist: (1) most current approaches rely

* Corresponding author.
E-mail addresses: saige_qin@stu.ecnu.edu.cn (S. Qin), mliu@mail.ecnu.edu.cn (M. Liu), tqwei@cs.ecnu.edu.cn (T. Wei), qyliu@spe.ecnu.edu.cn (Q. Liu).

https://doi.org/10.1016/j.eswa.2025.129712
Received 9 January 2025; Received in revised form 26 July 2025; Accepted 10 September 2025
Available online 13 September 2025
0957-4174/© 2025 Elsevier Ltd. All rights reserved, including those for text and data mining, AI training, and similar technologies.

on structured tasks, which often fail to accurately reflect children's language abilities in naturalistic settings; and (2) comprehensive evaluations of broader language skills—including comprehension, expression, communication, and other areas—remain insufficient.

To address these issues, this study proposes an LLM-based framework for assessing language abilities in children with ASD. First, we use speaker identification and automatic speech recognition (ASR) technologies to transcribe audio from child-parent conversations into natural language text. Based on existing research, assessment tools, and diagnostic standards, we design four language evaluation dimensions. A standardized workflow is established to ensure consistency and efficiency, and we create a dedicated prompt to guide the evaluation process with large language models. Next, to enhance accuracy and stability, we introduce a multi-expert voting mechanism. We experimentally compare the performance of different LLMs. A correlation analysis between the model's results and the Mullen Scales of Early Learning (MSEL) Verbal Developmental Quotients (VDQ) shows a strong positive correlation. Additionally, we use the eXtreme Gradient Boosting (XGBoost) classifier to verify how well the multidimensional evaluation distinguishes children with ASD from typically developing (TD) children. The results demonstrate high classification accuracy. The primary contributions of this study are summarized as follows:

- We propose a framework for assessing the language abilities of children with ASD based on LLMs. This framework enables multidimensional evaluation through the analysis of natural language samples, improving efficiency and reducing subjectivity.
- We design four language assessment dimensions that cover key aspects of language development in children with ASD. Additionally, we establish a standardized workflow and create high-quality prompts to to guide LLMs in the evaluation.
- To enhance the stability of the assessment, we introduce a multi-expert voting mechanism to mitigate bias in individual evaluation results. Moreover, by comparing multiple LLMs, we demonstrate a high correlation between the framework's evaluation results and the MSEL. Combined with an XGBoost classification model, the framework accurately distinguishes between children with ASD and TD, further validating its effectiveness.

The remainder of the paper is organized as follows: Section 2 reviews the related work. Section 3 details the proposed framework. Section 4 presents the experimentation on multi-dimensional language proficiency assessment. Section 5 reports the experimentation on child classification. Finally, Section 6 provides the discussion and conclusion.

2. Related work

2.1. Traditional language assessment methods

Traditional language ability assessments typically rely on clinical assessment scales, questionnaires, and structured interviews. The MSEL (Mullen et al., 1995) is one of the standardized tools commonly used to assess children's language and cognitive development. It is suitable for children from birth to 68 months of age and covers multiple developmental domains, including receptive and expressive language. Each domain includes graded descriptions: NA, very low, below average, average, above average, and very high. In this study, the MSEL Verbal Developmental Quotient is used as a core reference standard to validate the effectiveness of the intelligent language assessment approach. In addition to the MSEL, other commonly used tools for assessing children's language development include: the Early Language Milestone Scale (ELMS) (Coplan, 2005), the Peabody Picture Vocabulary Test (PPVT) (Dunn & Dunn, 2019), the Clinical Evaluation of Language Fundamentals (CELF) (Wiig et al., 2013), the Chinese Communicative Development Inventories (CCDI) (Tardif & Fletcher, 2008), and the Language Development Survey (LDS) (Rescorla, 1989). Among them, the ELMS is designed for infants and toddlers aged 0 to 36 months, assessing auditory receptive, visual language, and auditory expressive abilities. The PPVT evaluates receptive vocabulary skills for individuals aged 2.5 years and older. The CELF is applicable to children across different age groups and is designed to provide a comprehensive assessment of their development in multiple dimensions, including language comprehension, expression, and pragmatics. The CCDI are parent-report tools designed to assess language comprehension, expression, and nonverbal communication in infants aged 8 to 30 months, while the LDS relies on parent reports to assess vocabulary and early language development in young children. Although these tools are widely used and valuable for assessing children's language development, they often rely on trained professionals for administration, which introduces challenges such as high subjectivity and low efficiency.

2.2. Intelligent language assessment methods

To overcome the limitations of traditional assessment methods, researchers have begun exploring intelligent approaches to language evaluation. The LENA records children's natural language using wearable devices and employs automatic speech recognition and signal processing techniques to extract key indicators such as adult word count, child vocalization frequency, and conversational turns. Studies have shown that these LENA-derived metrics are moderately and consistently correlated with children's language development levels (Nadwodny et al., 2025; Wang et al., 2020). Furthermore, by leveraging phoneme modeling and age-based regression methods, LENA can reasonably accurately predict expressive language abilities (Richards et al., 2017), providing effective support for language screening and intervention monitoring. However, LENA primarily focuses on quantitative aspects of the language environment and has limited capacity to assess complex language abilities. To address this, Lin et al. developed an online multimedia language assessment system encompassing six types of language tasks under auditory and visual stimuli (Lin et al., 2013). The results demonstrated good psychometric properties, including both reliability and validity. Gretter et al. combined multilingual speech recognition with feedforward neural networks to construct an automated language proficiency classification system, which offers high efficiency and strong scalability (Gretter et al., 2019). Gale et al. proposed an end-to-end framework that integrates adaptive ASR with machine learning to automatically score expressive language (Gale et al., 2020). More recently, Nnamoko et al. applied the Word2Vec model to automatically score the vocabulary subtest of the WASI-II, highlighting the potential of word embedding techniques in the automation of language assessment (Nnamoko et al., 2024).

In recent years, the application of LLMs in medical sciences has expanded rapidly, with research across multiple disciplines confirming their value (Ayers et al., 2023; Caruccio et al., 2024; Kraljevic et al., 2022; Ouyang et al., 2024; Yeo et al., 2023). In the field of child language and developmental assessment, exploration of LLM applications has also gradually commenced. Feng et al. applied LLMs to conversational analysis of children with autism, demonstrating good performance in predicting language developmental stages and providing new insights for clinical auxiliary diagnosis (Feng et al., 2024). Li et al. validated the potential of LLMs in early screening of childhood language disorders and pronunciation error identification based on phoneme-level automatic scoring and feedback (Li et al., 2025). A recent study (2025) indicates that utilizing LLMs to analyze children's self-reported experiences during free play scenarios can effectively assess their multidimensional capabilities in cognitive, motor, social, and emotional domains, highlighting the broad application value of LLMs in child developmental assessment (Yang et al., 2025). Nevertheless, existing studies have yet to systematically evaluate the multidimensional language abilities of children with ASD. To address this gap, this study proposes an LLM-based assessment framework specifically designed for children with ASD, aiming to more comprehensively identify the characteristics of their language impairments and to further explore the potential of LLMs in supporting the identification and intervention of language disorders.

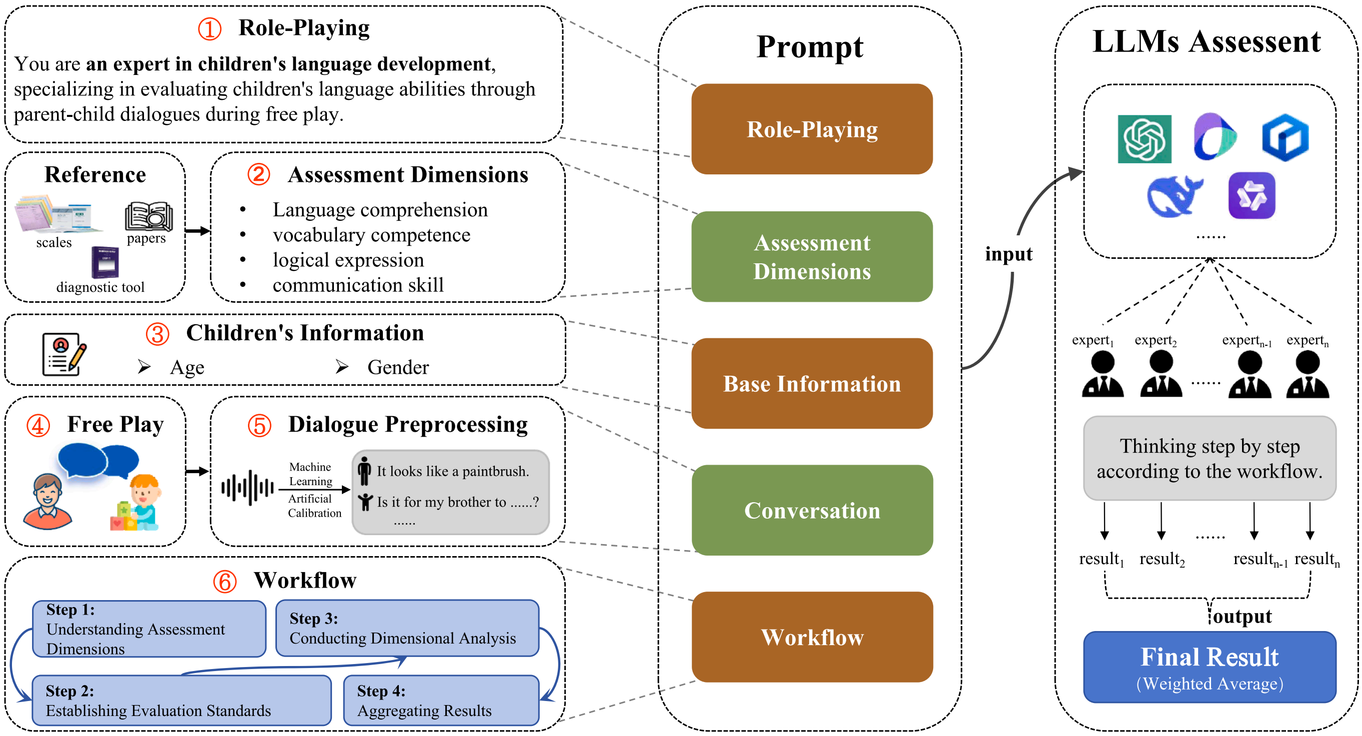

Fig. 1. The framework for language proficiency assessment. The left side illustrates the dataset, preprocessing, and other modules involved in constructing the prompt, including role-playing, evaluation dimensions, and workflow. The central part presents the specific prompts that are designed as part of our evaluation framework, while the right side showcases various large language models (LLMs) and the multi-expert voting mechanism.

3. Method

This chapter provides a detailed explanation of the evaluation framework(as shown in Fig. 1) and its key steps in this study. First, an in-depth analysis of the dataset and its preprocessing methods is presented(see Section 3.1). Next, the specific content of the prompts designed for this study is introduced, along with complete examples (see Section 3.2). Finally, the chapter elaborates on the various LLMs employed and the multi-expert voting mechanism utilized in this research (see Section 3.3).

3.1. Dataset and preprocessing

3.1.1. The dataset

This study was reviewed and approved by the Human Research Protection Committee of East China Normal University (HR706-2022). Prior to the commencement of the study, we obtained written informed consent from all children's parents. The consent form detailed the study's objectives, procedures, and data security measures, ensuring parents were fully informed before agreeing to participate. A total of 79 Mandarin-speaking Chinese children, aged 2 to 6 years, participated in the study, including 29 children with TD, 21 children with developmental disorder(DD), and 29 children with ASD. All children are professionally diagnosed by developmental-behavioral pediatricians. Children in the ASD group receive positive results on both the Childhood Autism Rating Scale (CARS) (Schopler et al., 2010) and the Autism Behavior Checklist (ABC) (DA, 1978). Children in both the DD and TD groups show negative results on the CARS and ABC assessments. The TD group demonstrates age-appropriate developmental outcomes without any indicators of developmental delay or ASD. Additionally, professionals administer the MSEL to all three groups. The demographic information and clinical symptom indicators of participants are presented in Table 1. To further analyze between-group differences in MSEL VDQ, Table 2 is added. Since Table 1 reveals unequal variances in MSEL VDQ across the three groups of children, Welch's ANOVA is employed in Table 2 to ensure accurate between-group difference analysis.

To assess children's language abilities in naturalistic settings, we record 79 parent-child dyads during 15 min free play sessions in a laboratory environment. The total duration of these recordings is approximately 1185 min. To facilitate natural interactions during these sessions,

Table 1
Participants' Demographic and Symptom Variables Across Groups. VDQ: Verbal Developmental Quotient, ADOS-2: Autism Diagnostic Observation Schedule Second Edition, SA: social affect, RRB: restricted or repetitive behavior, M: mean, SD: standard deviation, CV: Coefficient of Variation.

Group	Measure	M ± SD	Variance	CV
TD ($n = 29$)	MSEL (VDQ)	108.12 ± 12.58	158.17	0.12
DD ($n = 21$)	MSEL (VDQ)	50.96 ± 22.58	510.01	0.44
ASD ($n = 29$)	MSEL (VDQ)	35.31 ± 20.97	439.62	0.59
	ABC	52.76 ± 24.95	–	–
	CARS	30.55 ± 6.65	–	–
	ADOS-2 (SA)	16.17 ± 3.32	–	–
	ADOS-2 (RRB)	3.21 ± 1.81	–	–
	ADOS-2 (SA + RRB)	19.38 ± 3.85	–	–

Table 2
Group Comparisons of MSEL VDQ Using Welch's ANOVA. F indicates the magnitude of group differences; p reflects statistical significance ($p < 0.05$ is typically considered significant); η^2 denotes effect size, representing the proportion of total variance explained by group membership.

Comparison	F	p	η^2
ASD vs DD vs TD	146.05	< 0.001	0.75
ASD vs DD	5.94	< 0.05	0.11
ASD vs TD	248.28	< 0.001	0.82
DD vs TD	104.90	< 0.001	0.72

we provide age-appropriate toys for the children, such as small toy cars, cooking sets, bubbles, picture books, balls, and crayons. In our analysis of these audio recordings, we focus on minutes 3 through 13. This selection is made to avoid periods where the child may be silent at the beginning or fatigued towards the end. To ensure the integrity of the dialogues, we adopt a flexible strategy: if key conversations cross the boundaries of this time period, we adjust the starting and ending points of the analysis accordingly.

3.1.2. Dialogue preprocessing

Given that LLMs can efficiently process and generate diverse textual content, data preprocessing is required before evaluation. As shown in Fig. 1⑤, we use the open-source speech recognition model Paraformer-large (Gao et al., 2022) to transcribe the recorded audio, generating corresponding text and timestamps. Additionally, speech segments from both the child and the parent are extracted from the audio as samples. The speaker identification model ERes2Net (Chen et al., 2023a) is then used to determine the speaker for each segment, enabling accurate separation of the child's and parent's speech.

Since children's speech is still in the developmental stage and often exhibits unique phonetic characteristics, the accuracy of automatic transcription is generally low. Frequent transcription errors may affect the accuracy of language ability assessments based on LLMs. To address this issue, we enlisted two experts to manually calibrate the automatically transcribed text to ensure its accuracy. Finally, to protect user privacy, we have implemented strict anonymization measures for sensitive information about children and parents, including their names, home addresses, and phone numbers.

3.2. Prompt design for child language assessment

The prompt design employs Markdown formatting to establish clear hierarchical relationships that enhance model comprehension (OpenAI, 2025). To facilitate systematic reasoning, we implement the Chain-of-Thought (CoT) paradigm, which guides the model through step-by-step analytical processes (Wei et al., 2022). Additionally, we incorporate a role-playing mechanism that enables the LLM to act as an expert in assessing children's language abilities (Chen et al., 2024; Tseng et al., 2024; Tu et al., 2024; White et al., 2023). The system outputs results in JSON format to streamline automated parsing and subsequent processing. The implementation details of these components are elaborated in the subsequent sections.

3.2.1. Role-playing

Research has shown that assigning specific roles to LLMs can significantly enhance their understanding of specialized domains (Kong et al., 2024; Shanahan et al., 2023; Shao et al., 2023; Tseng et al., 2024; Wu et al., 2024). In this study, to improve the application of LLMs in analyzing children's conversational content, we have defined a specific role for the model. As illustrated in Fig. 1①, we position the model as an expert in the field of child language development. This expert specializes in evaluating children's language abilities by analyzing parent-child dialogues in free play scenarios.

3.2.2. Assessment dimensions

As shown in Fig. 1②, this study assesses language abilities across four core dimensions: language comprehension, vocabulary competence, logical expression, and communication skills. These dimensions were developed through a comprehensive process involving systematic literature review, expert consultation, and multi-round collaborative discussions among all authors. The corresponding author is a professor of speech-hearing rehabilitation science, the first author is a doctoral student in intelligent education, the second author is a postdoctoral researcher in speech and hearing rehabilitation science, and the third author is an associate professor in computer science and technology. Table 3 presents the specific scope of each dimension as finalized through this collaborative process. The following provides a detailed explanation of the theoretical foundations and practical basis for each assessment dimension.

Language comprehension: Children with ASD experience significant difficulties in language comprehension, primarily characterized by challenges in understanding semantics, deviating from the conversation topic, and improper use of pronouns (Liu, 2019; Zhou et al., 2023). Research indicates that these children typically understand language at a literal level, struggling to grasp deeper meanings from context (Happé & Frith, 2006). For instance, they often have difficulty with abstract concepts or rhetorical devices, such as metaphors (Su & Naigles, 2020). Additionally, children with ASD frequently provide off-topic answers, misunderstanding the intent of the question or straying from the conversation (Geurts & Embrechts, 2008). They also commonly struggle with distinguishing and using pronouns (e.g., 'you,' 'I,' 'he,' etc.) (Gernsbacher et al., 2016). The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) notes that language deficits in individuals with ASD can range from a complete lack of language to significant impairments in understanding complex semantics and syntax.

Vocabulary competence: This dimension aims to assess the ability of children with ASD to understand, acquire, and use vocabulary. Research has shown that children with ASD often exhibit language stereotypy, such as the repetitive use of the same words or phrases and imitation of parental speech (Shield et al., 2017). In terms of vocabulary acquisition, Wang Bijun et al. found that children with ASD possess a significantly smaller total vocabulary compared to children with TD. They also show marked deficits in the use of verbs, adjectives, and function words (e.g., pronouns and prepositions). Moreover, children with ASD tend to repeatedly use high-frequency words, indicating limited lexical diversity (Wang et al., 2017). Additionally, the DSM-5 notes that individuals with ASD often display stereotyped or repetitive speech patterns, including imitative language and idiosyncratic phrases. Based on these research findings and diagnostic criteria, we have developed a vocabulary assessment dimension specifically for children with ASD.

Logical expression: This dimension focuses on assessing the expression abilities of children with ASD. Comparative studies of language abilities between children with ASD and TD peers have consistently demonstrated that children with ASD show marked deficits in syntactic expression and language organization (Yi et al., 2020). These difficulties are often reflected in the use of contextually inappropriate vocabulary, overly simplistic or incorrect grammatical structures, and disorganized or illogical information flow (Boucher, 2012; Eigsti et al., 2007; Geurts & Embrechts, 2008). The CELF includes subtests related to complex syntax and logical expression, which serve as important references for the development of evaluation criteria in this dimension. Drawing on these findings, this dimension primarily targets language organization and narrative skills, with an emphasis on identifying syntactic impairments and difficulties in coherent expression among children with ASD.

Communication skills: The DSM-5 diagnostic criteria for ASD state that individuals with ASD have significant deficits in social interactions, such as an inability to engage in normal conversations, a reduced interest in sharing, or a lack of response to social interactions. The Autism Diagnostic Interview-Revised (ADI-R) (Rutter et al., 2003) evaluates children's communication skills through a structured interview with parents, providing important reference for assessment dimension design. Research has found that children with ASD often struggle with initiating conversations, responding appropriately, taking turns, repairing, and terminating conversations based on the context (MA et al., 2019; Zhao et al., 2021). Based on these research findings and clinical tools, this assessment focuses on key aspects of language use in social interactions, including communicative initiation, conversational turn-taking, and pragmatic functioning, aiming to comprehensively capture the unique patterns and potential obstacles in social communication for children with ASD.

Table 3
Assessment dimensions and scopes.

Assessment dimensions	Scope of assessment
Language Comprehension	• Whether there are semantic comprehension difficulties (e.g., difficulty understanding literal meanings, grasping contextual meanings, comprehending abstract concepts, or interpreting figurative language and other rhetorical devices); • Whether irrelevant or off-topic responses occur (e.g., topic deviation, misinterpretation of questions); • Whether there is difficulty in differentiating personal pronouns (e.g., confusion in the use of 'you,' 'I,' and 'he/she.')
Vocabulary Competence	• Whether the individual has mastered and can use common or comprehensible vocabulary; • Whether there is repetitive use of the same words, phrases, or short sentences; • Whether there are multiple occurrences of repeating parents' utterances; • Whether novel or invented words or phrases are used in expression.
Logical Expression	• Whether there is disordered use of words (e.g., uttering statements irrelevant to the current context); • Whether difficulties in information organization are present (e.g., lack of coherence between sentences, overly simplistic grammar, errors in grammatical structure, lack of logical consistency and coherence in expression).
Communication Skills	• Whether the individual can initiate, respond to, take turns in, and terminate conversations; • Whether communication difficulties are present (e.g., poor ability to introduce and maintain topics, inability to communicate effectively in specific contexts); • Whether instances of soliloquy occur; • Whether there is frequent topic switching; • Whether a tendency to focus solely on topics of personal interest occurs.

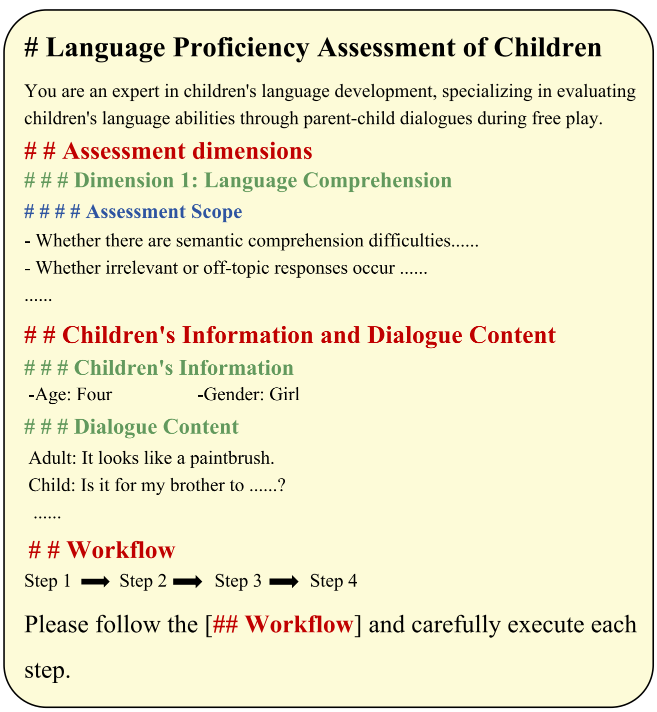

Fig. 2. Prompt example.

3.2.3. Workflow

To ensure that LLMs effectively assess the language abilities of children with ASD, the key focus is on designing a structured workflow to guide the evaluation process. Based on LLM role-playing, assessment dimensions, children's information, and parent-child dialogue content, we develop a workflow that mainly includes the following steps.

- Step 1: Understanding Assessment Dimensions. The LLM thoroughly analyzes and comprehends the four core assessment dimensions and their respective evaluation criteria.
- Step 2: Establishing Evaluation Standards. Based on the understanding gained in Step 1, the LLM incorporates children's information, such as gender and age, to determine tailored evaluation standards for each dimension.
- Step 3: Conducting Dimensional Analysis. Using the evaluation standards defined in Step 2, the LLM analyzes the dialogue content between the child and their parent. It assigns a reasonable grade for the child's language development in each dimension and provides detailed reasoning for the assigned grades.
- Step 4: Aggregating Results. The LLM consolidates the assessment results into a JSON format, using the assessment dimensions as keys and the corresponding grades as values.

The complete prompt, using an example of a child with TD, is shown in Fig. 2.

3.3. LLMs assessment

3.3.1. LLMs

This study selects LLMs based on the following criteria: support for fine-tuning, open-source status, support for Chinese tasks, and commercial availability. Based on these standards, the models selected include ChatGPT (OpenAI, 2024a), Qwen (Yang et al., 2024), DeepSeek (DeepSeek-AI, 2024), ChatGLM (GLM et al., 2024), Yi (I. et al., 2024), ERNIE, and Doubao. Additionally, considering application costs, we also compared versions of some models with different parameter sizes. Table 4 lists the models used in this study along with their key characteristics, including model name, parameter size, open-source status, etc. For models with undisclosed parameter sizes, a"-" is used in the table.

3.3.2. Multi-expert voting mechanism

We employ the multi-expert voting mechanism (as illustrated on the right side of Fig. 1) to enhance the accuracy and robustness of evaluating children's language abilities. Each large language model functions as an expert, independently assessing the performance of children's language samples. The final evaluation result for each child is derived through the weighted aggregation of scores from multiple experts. Specifically, let the outputs from the experts be $y_1, y_2, \ldots, y_n$, with corresponding weights $w_1, w_2, \ldots, w_n$. The final output is computed as the weighted average of these scores:

$$y_{\text{final}} = \sum_{i=1}^{n} w_i \cdot y_i \tag{1}$$

Here, y_i represents the prediction result of the i-th expert, w_i represents the weight of the i-th expert, n is the *number of experts*, $\sum_{i=1}^{n} w_i = 1$, and y_{final} is the final voting result. In this study, we assume that each expert has the same weight.

Table 4
Characteristics of the LLMs used in this study. The symbol "-" indicates that this particular information has not been disclosed.

Model	Model sizes	Hyperpa-rameter tuning (Y/N)	Open source (Y/N)	Support for Chinese tasks (Y/N)	Commer-cial availability (Y/N)
GPT4o	–	N	N	Y	Y
Doubao-Pro-32k	–	Y	N	Y	Y
ERNIE-Speed-8K	–	Y	N	Y	Y
ERNIE-Lite-8K	–	Y	N	Y	Y
DeepSeek-V2-Chat	236B	Y	Y	Y	Y
Qwen2-72B-Instruct	72B	Y	Y	Y	Y
Qwen2-57B-A14B-Instruct	57B	Y	Y	Y	Y
Yi-1.5-34B-Chat	34B	Y	Y	Y	Y
GLM-4-9B-Chat	9B	Y	Y	Y	Y

4. Experimentation on multi-dimensional language proficiency assessment

This section primarily introduces the children's language proficiency assessment experiment. It first provides a detailed description of the experimental setup (see Section 4.1) and evaluation metrics (see Section 4.2). Then, it presents an in-depth analysis of the performance of different LLMs in understanding and following instructions, as well as analyzing dialogue content, to gain a better understanding of the models' performance in the evaluation task (see Section 4.3). Furthermore, a correlation analysis between the model evaluation results and MSEL verbal developmental quotients is conducted to further validate the effectiveness of this study (see Section 4.4).

4.1. Experimental setup

In the experiment, we focus on and adjust three parameters *top_p*, *temperature*, and *number of experts*. *top_p* and *temperature* are commonly used parameters in text generation tasks to control the randomness and diversity of model outputs. Specifically, the *top_p* (typically ranging from 0 to 1) controls the size of the candidate set considered when selecting the next token. Formally, it can be represented as follows:

$$P_{\text{final}} = \arg\max\left(\sum_{i=1}^{k} p_i\right) \tag{2}$$

Among them, p_i denotes the probability of each candidate outcome, $\sum_{i=1}^{k} p_i \leq top\_p$, and *top_p* represents the selected probability threshold. The *temperature* regulates the degree of randomness in the model's prediction of the next token. Formally, it can be represented as follows:

$$P(w_i|T) = \frac{\exp\left(\frac{\log(P(w_i))}{T}\right)}{\sum_j \exp\left(\frac{\log(P(w_j))}{T}\right)} \tag{3}$$

Where $P(w_i)$ is the original probability of the word w_i, and T is the *temperature*. The function $\exp(\cdot)$ denotes the exponential function, and $\log$ is the logarithm. The symbols w_i and w_j represent the current word being considered and all possible candidate words, respectively.

The *number of experts* is a parameter that we introduce in the multi-expert voting mechanism (see Section 3.3.2 for details). To investigate the impact of these three parameters on the stability of the results, we conduct five comparative experiments on each model. Fig. 3 illustrates the trend in these models' results as the parameters change.

Each subfigure in Fig. 3 represents the results of one model. The horizontal axis indicates the *number of experts* participating in the evaluation, whereas the vertical axis reflects the standard deviation of five assessments, quantifying the dispersion of scores. Fig. 3(a)–(f) display the evaluation results for Qwen2-72B-Instruct, Qwen2-57B-A14B-Instruct, Yi-1.5-34B-Chat, ERNIE-Speed-8K, ERNIE-Lite-8K, and GLM-4-9B-Chat, respectively. Each graph contains 15 lines, each representing a unique combination of *top_p* (values: 0.5, 0.75, 1.0) and *temperature* (values: 0.1, 0.3, 0.5, 0.75, 1.0) parameters for that model. Fig. 3(g)–(i) present the evaluation results for GPT4o, Doubao-Pro-32k, and DeepSeek-V2-Chat, respectively. GPT4o and Doubao-Pro-32k are closed-source models with high costs associated with using their official APIs, whereas DeepSeek-V2-Chat requires substantial computational resources that exceed our deployment capabilities. Consequently, for these three models, this study only tests the trend of model stability with varying numbers of experts under the fixed settings of *top_p* = 1.0 and *temperature* = 1.0.

The results reveal a significant impact of expert count on assessment stability. Across all models, with fixed *top_p* and *temperature*, the standard deviation demonstrates a decreasing trend as the *number of experts* increases. As shown in Fig. 3(g), the GPT4o model, with *top_p* set to 1.0 and *temperature* to 1.0, shows a remarkable 60 % reduction in standard deviation from approximately 7.3 with one expert to 3.0 with six experts. This nonlinear downward trend is consistent across all models, with the most pronounced decrease observed when increasing the *number of experts* from one to four. For example, as shown in Fig. 3(c), the Yi-1.5-34B-Chat model, under the same parameter settings, also demonstrates this trend, with its standard deviation decreasing from approximately 9.7 with one expert to about 4.9 with four experts, after which the rate of decline slows significantly. These findings indicate that while increasing the *number of experts* substantially enhances assessment stability, there is a diminishing marginal effect.

The *top_p* and *temperature* settings have equally significant effects on stability. As shown in Fig. 3(a), taking the Qwen2-72B-Instruct model as an example, when top-p is set to 0.5 and *temperature* to 0.1, the standard deviation for evaluations by six experts is approximately 2.2, whereas it increases to about 3.3 when top-p is 1.0 and *temperature* is 1.0. Under the same conditions, ERNIE-Speed-8K's standard deviation increases from about 1.8 to about 5.8. This tendency is also prevalent in other models. This phenomenon suggests that lower *top_p* and *temperature* values tend to produce more stable assessment results.

As shown in Fig. 3, comparing the results of the four open-source models (Qwen2-72B-Instruct, Qwen2-57B-A14B-Instruct, Yi-1.5-34B-Chat, and GLM-4-9B-Chat) reveals that Qwen2-72B-Instruct exhibits the most stable performance. Even with parameter settings of *top_p* = 1.0 and *temperature* = 1.0, its standard deviation in the 6-expert assessment is only about 3.3. In contrast, under the same conditions, the standard deviations of other models are as follows: Qwen2-57B-A14B-Instruct has a standard deviation of about 5.8, Yi-1.5-34B-Chat's standard deviation is approximately 4.1, and GLM-4-9B-Chat reaches approximately 5.2.

In order to quantify the stability of the assessment, we set a stability threshold based on the scoring mechanism of the model assessment (see Section 4.2 for details). Considering that the minimum score difference between different language proficiency levels is 3.75, we define a standard deviation of less than 3.75 as an acceptable level of stability. The application of this threshold helps us to objectively judge the reliability of the assessment results under different model and parameter settings. Based on the set standard deviation threshold, we identify the models and their parameter settings that satisfy the stability requirement. Specifically, Qwen2-72B-Instruct meets the stability requirement for all parameter combinations when the *number of experts* is four or more. Other models, such as ERNIE-Speed-8K, satisfy the requirement

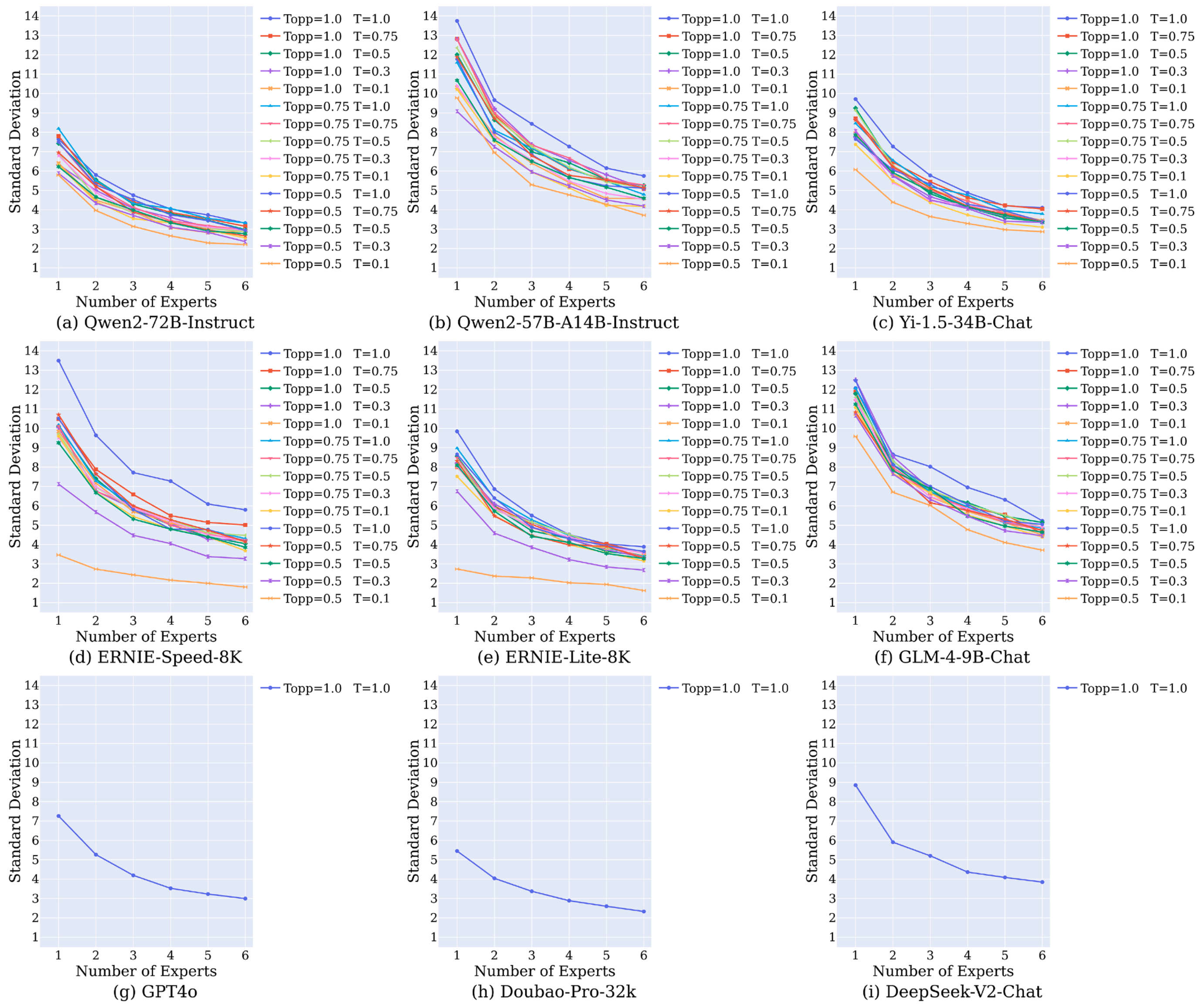

Fig. 3. Performance trends of nine LLMs under different evaluation conditions. The horizontal axis shows the *number of experts* participating in the evaluation, and the vertical axis represents the standard deviation across five assessments. Fig. 3(a)–(f) each contain 15 curves, with each curve representing a unique combination of *top_p* values (0.5, 0.75, 1.0), denoted as Topp, and *temperature* settings (0.1, 0.3, 0.5, 0.75, 1.0), denoted as T. Fig. 3(g)–(i) show only one curve per graph, illustrating the results under fixed parameters of Topp = 1.0 and T = 1.0.

only at lower *top_p* and *temperature* values and with more experts. It is worth noting that although the standard deviation of DeepSeek-V2-Chat decreases as the *number of experts* increases, it never reaches the preset threshold of acceptability within the tested range, and thus its evaluation results are not used in the subsequent experiments in this paper.

4.2. Evaluation metrics

To quantify the results of the LLMs assessment, a numerical mapping method is developed in this study. The method is based on four established assessment dimensions and five proficiency levels, converting level assessments into numerical representations. Considering children's language development, we set the range of the total score for the LLMs language ability assessment to 0-150, which aligns with the observation that the MSEL VDQ of children typically do not exceed 150. Based on the maximum total score of 150, we employ the following method to distribute scores across dimensions and levels: First, the total score is equally distributed among the four assessment dimensions, with a maximum score of 37.5 for each dimension. Second, within each dimension, we use arithmetic progression to assign scores to the five levels as follows: excellent (37.5), good (30), fair (22.5), poor (15), and very poor (7.5). The design of these five levels references the graded descriptions of language comprehension and language expression from the MSEL and is validated through expert review to ensure the scientific rigor and applicability of the scoring criteria.

The workflow of the framework proposed in this paper is clearly defined in Section 3.2.3. However, through the analysis of the experimental results, we find significant differences among various LLMs in terms of performance and adherence to instructions. Some models do not strictly adhere to our predetermined format in returning evaluation grades. Through statistical analysis, we identify three special cases and formulate corresponding treatment strategies:

- Ratings Between Two adjacent Grades: Certain LLMs, following the prompt instructions, provide ratings such as “poor to fair,” “fair to average,” “average to good,” or “good to excellent.” For these evaluations spanning two grades, we assign a score equal to the mean of the two grade values.

- Unevaluable Instances: In some cases, LLMs follow the prompt but responded with "unable to evaluate." For these instances, we assign a score of 0.
- Responses Irrelevant to the Task: In some instances, LLMs completely disregard the prompt instructions, providing content entirely unrelated to the task. In these cases, we implement a more stringent protocol: we repeat the request five times. If the model consistently fails to execute the instructions, we exclude that sample from subsequent evaluation tasks for all models.

To validate the effectiveness of our language assessment framework, we conduct a correlation analysis comparing the model evaluation results with the MSEL verbal developmental quotients. We calculate the Pearson correlation coefficient (r) and the corresponding p-value to quantify the degree of association between these two sets of results and its statistical significance. The Pearson correlation coefficient measures the strength and direction of the linear relationship between two variables, with values ranging from -1 to 1. An r value close to 1 indicates a strong positive linear relationship, whereas an r value close to -1 indicates a strong negative linear relationship. An r value around 0 suggests no linear correlation. The p-value assesses the statistical significance of the observed correlation, with a p-value less than 0.05 typically considered statistically significant.

4.3. The analysis of evaluation results

This section explores the specific performance of LLMs in the language assessment task. To guide LLMs in dialogue analysis and result generation, we clearly outline key information within the prompt, such as evaluation dimensions, scope, and workflow. However, despite providing detailed instructions, we observe significant performance differences across models in executing this task. These differences mainly appear in two areas: 1) understanding and following instructions; 2) analysis of dialogue content.

4.3.1. Understanding and following instructions

A comprehensive analysis of the output results from LLMs reveals significant disparities in their ability to understand instructions. Specifically, Doubao-Pro-32k, ERNIE-Speed-8K, ERNIE-Lite-8K, GLM-4-9B-Chat, Yi-1.5-34B-Chat, and Qwen2-57B-A14B-Instruct show varying limitations in comprehending the provided prompt. The prompt encompasses multiple evaluation dimensions and detailed assessment scope, requiring the LLMs to conduct reasoned analysis and generate evaluation results based on a thorough understanding of these elements. However, these models demonstrate deficiencies in grasping critical information. For instance, in understanding the evaluation dimension of vocabulary competence, models such as ERNIE-Speed-8K, ERNIE-Lite-8K, GLM-4-9B-Chat, Yi-1.5-34B-Chat, and Qwen2-57B-A14B-Instruct erroneously interpret repeated phrase usage, echoing the parent's words, or word fabrication as indicators of stronger language skills. This misinterpretation persists despite the prompt explicitly annotating these behaviors as indicative of limited linguistic ability. Doubao-Pro-32k occasionally exhibits similar misunderstandings in understanding instructions. In contrast, GPT4o and Qwen2-72B-Instruct demonstrate a better comprehension of the prompt.

In terms of following instructions, ERNIE-Speed-8K and ERNIE-Lite-8K show notable deficiencies. The prompt explicitly requires the final output to include JSON-formatted rating results; however, these models frequently produce ambiguous cross-level evaluations (e.g., "average to good") or inaccurate assessment outcomes. They occasionally fail to generate the requisite JSON-formatted rating results. Aside from these models, all others perform well in following instructions.

4.3.2. Analysis of dialogue content

In the natural language samples utilized in this study, some children's linguistic output is relatively limited, presenting a significant challenge to the analytical capabilities of LLMs. When evaluating these children's language abilities, models such as Qwen2-57B-A14B-Instruct, ERNIE-Speed-8K, ERNIE-Lite-8K, GLM-4-9B-Chat, and Yi-1.5-34B-Chat exhibit varying degrees of limitations, often providing overly optimistic assessments (Figs. A.1–A.4 present a set of comparative examples of assessment results for children with ASD who fit the above-described situations). For instance, in cases where children only occasionally produce vocalizations such as "hmm," "ah," or "oh" during communication, these models assume that the children understand the parent's instructions and questions. In terms of logical expression, these models even assess the children's expressions as having a certain degree of coherence. Additionally, in assessing vocabulary competence, the models mistakenly conclude that the children have acquired a basic level of action-related vocabulary and words related to daily objects. In contrast, GPT4o, Qwen2-72B-Instruct, and Doubao-Pro-32k demonstrate superior objectivity and accuracy in their analyses. Concerning language comprehension, these models accurately acknowledge that while the child responds to parental queries, the responses are extremely limited and simple. They also recognize that for more complex statements, the child may not respond due to comprehension difficulties. In terms of logical expression, these models accurately highlight the child's simple expressions, the lack of inter-sentential coherence, and insufficient information organization skills. These findings underscore the significant disparities in the capabilities of various LLMs when analyzing complex language samples, particularly those involving limited linguistic output from children.

To further objectively analyze the models' evaluation capabilities, this study designs and implements a quantitative comparative analysis based on expert annotations. We invite a postdoctoral researcher specializing in speech and hearing rehabilitation sciences, a doctoral student with pediatric practice qualifications, and a doctoral student focusing on special education research for children with autism to independently annotate the language samples from all 79 children. Through discussion, consensus is reached to establish expert reference results as the comparative benchmark. Subsequently, we compare each model's evaluation outputs on the same samples against these reference results on a dimension-by-dimension basis, calculating the rating deviations for each evaluation dimension. The relevant statistical results and analysis are provided in Appendix B.

4.4. The analysis of relevance results

Fig. 4 illustrates the correlation analysis between the results of the multiple LLMs assessment and the MSEL verbal development quotients. The horizontal axis represents the standard deviation, whereas the vertical axis shows the Pearson correlation coefficient's r-value. All models' parameter settings meet the stability requirements (standard deviation < 3.75).

The results demonstrate a clear hierarchical trend in the correlation between the assessment outcomes of different models and the MSEL verbal development quotients. GPT4o, Qwen2-72B-Instruct, and Doubao-Pro-32k demonstrate exceptional performance, with their Pearson correlation coefficients predominantly distribute in the range of 0.7 to 0.8 ($p < 0.001$), exhibiting significant positive correlations with MSEL results. Notably, the GPT4o model stands out with a remarkable correlation coefficient of 0.8 ($p < 0.001$), indicating the strongest positive correlation. This is closely followed by the open-source model Qwen2-72B-Instruct, achieving a correlation coefficient of 0.74 ($p < 0.001$), which also demonstrates a strong positive correlation. These strong positive correlations not only highlight the exceptional performance of GPT4o and Qwen2-72B-Instruct in understanding instructions, following directives, and analyzing dialogue content, but also validate the effectiveness of the assessment framework we designed.

In contrast, the performance of ERNIE-Speed-8K, ERNIE-Lite-8K, GLM-4-9B-Chat, and Yi-1.5-34B-Chat models are comparatively inferior. These models exhibit stronger variability in their correlations,

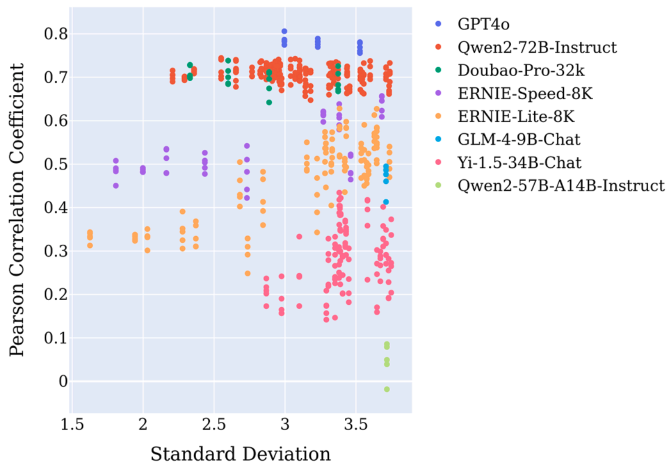

Fig. 4. Relationship between LLMs assessment results and MSEL verbal development quotients: Pearson correlation analysis. The horizontal axis represents the standard deviation, whereas the vertical axis shows the Pearson correlation coefficient's *r*-value. The results of different models are represented by different colors.

primarily distribute within the range of 0.1 to 0.5 (p<0.001), demonstrating weaker correlations. These results are closely related to the models' limitations in understanding instructions, following directives, and analyzing dialogue content.

By comparing the results of Qwen2-72B-Instruct with Qwen2-57B-A14B-Instruct, we observe significant performance differences between models within the same series. Specifically, Qwen2-72B-Instruct demonstrates notably superior performance compared to Qwen2-57B-A14B-Instruct. The results are consistent with our previous observation of the excellent performance of Qwen2-72B-Instruct in analyzing conversations. In Table 5, we have listed the key information for each model at its peak correlation performance, including parameter settings, correlation results, and their corresponding significance levels.

5. Experimentation on child classification

This section primarily presents the experiment on classifying children based on multidimensional language proficiency assessment results. First, the experimental setup used in the study is described in detail (see Section 5.1). Next, the evaluation metrics for the classification experiments are introduced (see Section 5.2). Finally, both three-class and two-class classification experiments are conducted on children with ASD, TD, and DD (see Section 5.3), followed by a comparative analysis of the potential application of multidimensional language ability in identifying children with ASD.

5.1. Experimental setup

Correlation analyses show a significant positive correlation between the model assessment results and the MSEL verbal development quotients, validating the feasibility of our assessment framework. To further test the ability of this language assessment framework to discriminate between different categories of children, we use the language assessment results of the LLMs as features to categorize the children. The experimental features consist of numerical scores for each assessment dimension, along with the overall score, as produced by the LLMs' language ability evaluations. For this purpose, we employ the eXtreme Gradient Boosting (XGBoost) classifier. In the classification experiments, we use the assessment results of three models (GPT4o, Doubao-Pro-32k, and Qwen2-72B-Instruct) that have a correlation coefficient of 0.7 or higher with the MSEL measure. We optimize the XGBoost classifier using a grid search to find the optimal combination of parameters. The optimal parameter settings are detailed in Table 6. Among these parameters, $Colsample\_Bytree$ represents the proportion of features used by each tree. $Learning\_Rate$ is the learning rate, which controls the step size of each iteration. Max_Depth defines the maximum depth of each tree, while $N\_Estimators$ indicates the total number of decision trees in the model. Additionally, $Subsample$ specifies the proportion of observations used to train each tree.

5.2. Evaluation metrics

In this experiment, we use the cross-validation method to evaluate the model's performance. Given the limited sample size, we choose the Leave-One-Out cross-validation strategy. Specifically, in each iteration, a single subject is used as the test set, whereas the remaining subjects form the training set. We employ a comprehensive set of evaluation metrics to assess model performance, including Accuracy, Precision, Recall, Macro-F1, the Receiver Operating Characteristic (ROC) curve, and the Area Under the Curve (AUC). Accuracy reflects the proportion of correctly classified children. Precision measures the proportion of predicted positive samples that are actually positive. Recall (also known as Sensitivity), which is equivalent to the true positive rate, indicates the proportion of actual positive samples that are correctly predicted as positive. Macro-F1 is the unweighted average of F1-scores across all classes, where the F1-score is the harmonic mean of precision and recall. Additionally, the ROC curve visualizes the classifier's performance by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different thresholds, whereas the AUC quantifies the area under the ROC curve, offering a comprehensive measure of the classifier's discriminative ability. These key metrics are calculated as shown in (4) through (7).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{4}$$

$$Precision = \frac{TP}{TP + FP}. \tag{5}$$

$$Recall = \frac{TP}{TP + FN}. \tag{6}$$

$$\text{Macro-F1} = \frac{1}{n}\sum_{i=1}^{n} F1_i, \quad \text{where } F1_i = \frac{2 \cdot Precision_i \cdot Recall_i}{Precision_i + Recall_i} \tag{7}$$

Where *TP* (True Positives) and *TN* (True Negatives) denote the number of correctly predicted positive and negative samples, respectively.

Table 5
The optimal parameters and results of each model. In the *r* column, the underlined and bold results indicate a correlation greater than 0.7, demonstrating a significant strong positive correlation.

Model	*Top_p*	*Temperat-ure*	*Number of experts*	*r*	*p*
GPT4o	1.0	1.0	6	**0.80**	<0.001
Doubao-Pro-32k	1.0	1.0	6	**0.73**	<0.001
Qwen2-72B-Instruct	0.75	0.1	6	**0.74**	<0.001
Qwen2-57B-A14B-Instruct	0.5	0.1	6	0.09	<0.001
ERNIE-Speed-8K	0.75	0.1	6	0.66	<0.001
ERNIE-Lite-8K	0.75	1.0	6	0.63	<0.001
GLM-4-9B-Chat	0.5	0.1	6	0.50	<0.001
Yi-1.5-34B-Chat	0.5	0.75	6	0.43	<0.001

Table 6
Parameter settings of the XGBoost classifier in the child classification experiment.

Model	Task	*Colsample_Bytre*	*Learning_Rate*	*Max_Depth*	*N_Estimators*	*Subsample*
GPT4o	ASD vs DD vs TD	0.6	0.05	5	20	0.6
	ASD vs TD	0.6	0.01	3	10	1.0
	DD vs TD	0.6	0.01	3	10	0.6
	ASD vs DD	0.6	0.01	5	10	0.6
Doubao-Pro- 32k	ASD vs DD vs TD	0.6	0.01	5	10	0.6
	ASD vs TD	0.8	0.1	3	50	1.0
	DD vs TD	0.8	0.01	5	20	1.0
	ASD vs DD	0.6	0.1	5	20	0.6
Qwen2-72B- Instruct	ASD vs DD vs TD	1.0	0.01	5	50	0.6
	ASD vs TD	0.6	0.05	5	50	1.0
	DD vs TD	0.6	0.1	5	50	0.6
	ASD vs DD	0.6	0.01	7	10	0.8

Table 7
Three-class classification results for children with ASD, DD and TD using the XGBoost model. Bold and underlined values represent the best performance.

Model	Accuracy	Macro-F1	Precision			Recall		
			ASD	DD	TD	ASD	DD	TD
GPT4o	0.75	0.73	0.73	0.7	0.78	0.59	**0.67**	**0.97**
Doubao-Pro-32k	0.72	0.67	0.73	**0.86**	0.69	0.83	0.29	0.93
Qwen2-72B-Instruct	**0.82**	**0.80**	**0.76**	**0.86**	**0.87**	**0.9**	0.57	0.93

Table 8
Binary classification of children with ASD and TD using the XGBoost model. Bold and underlined values indicate the highest performance.

Model	Accuracy	Macro-F1	Precision		Recall	
			ASD	TD	ASD	TD
GPT4o	0.91	0.92	0.96	0.88	0.86	0.97
Doubao-Pro-32k	**0.98**	**0.98**	**1**	0.97	0.97	**1**
Qwen2-72B-Instruct	**0.98**	**0.98**	0.97	**1**	**1**	0.97

Table 9
Binary classification of children with DD and TD using the XGBoost model. Bold and underlined values indicate the highest performance.

Model	Accuracy	Macro-F1	Precision		Recall	
			DD	TD	DD	TD
GPT4o	0.94	0.94	**0.95**	0.93	0.9	**0.97**
Doubao-Pro-32k	0.92	0.92	**0.95**	0.9	0.86	**0.97**
Qwen2-72B-Instruct	**0.96**	**0.96**	**0.95**	**0.97**	**0.95**	**0.97**

Table 10
Binary classification of children with ASD and DD using the XGBoost model. Bold and underlined values indicate the highest performance.

Model	Accuracy	Macro-F1	Precision		Recall	
			ASD	DD	ASD	DD
GPT4o	0.78	0.77	0.8	0.75	0.83	**0.71**
Doubao-Pro-32k	0.82	0.79	0.76	**1**	**1**	0.57
Qwen2-72B-Instruct	**0.84**	**0.83**	**0.82**	0.88	0.93	**0.71**

FP (False Positives) denotes the number of negative samples incorrectly categorized as positive, and *FN* (False Negatives) denotes the number of positive samples incorrectly categorized as negative. n denotes the total number of classes. $F1_i$ is the F1 score for the i-th class, defined as the harmonic mean of precision and recall. $Precision_i$ represents the precision of the i-th class, and $Recall_i$ denotes the recall of the i-th class.

5.3. The analysis of classification results

Table 7 presents the cross-validation results for XGBoost in the three-class classification task, while Tables 8–10 present the results for the binary classification task. The features used in the experiments are the numerical scores for each dimension of the LLMs' language ability assessment, as well as the total scores. Figs. 5 and 6 depict the ROC curves for the three-class classification and binary classification tasks, respectively.

5.3.1. Classification of ASD, DD, and TD

As shown in Table 7, although the accuracies when using the assessment results of the different models as features for classification are different, they all exceed 0.70. Among them, the classification accuracy based on the assessment results of the Qwen2-72B-Instruct is the highest, at 0.82. More importantly, considering the imbalanced nature of the dataset (ASD = 29, DD = 21, TD = 29), the Macro-F1 scores provide a more reliable evaluation metric. Qwen2-72B-Instruct also achieves the highest Macro-F1 score of 0.80, confirming its superior performance. This indicates that the three groups of children display distinct language abilities during free-play scenarios.

Based on Qwen2-72B-Instruct's assessment results, the classification precision and recall are the highest in most cases. Specifically, the precision (0.76) and recall (0.9) for children with ASD are higher than those achieved by the other two models. For children with DD and TD, GPT4o's assessment results yield the best recall (DD recall: 0.67; TD recall: 0.97). As shown in the 'Recall' column, the recall for children with ASD and DD varies significantly across the three experiments, whereas the recall for children with TD remains relatively stable. Fig. 5 presents the ROC curves and AUC values for the classification tasks of 'ASD vs not ASD,' 'DD vs not DD,' and 'TD vs not TD'. Among these, the overall AUC for the 'TD vs not TD' classification is the highest (all exceeding 0.95). These results indicate a significant difference in language abilities between children with ASD or DD and children with TD. It is easiest to distinguish between children with TD and those with developmental differences based on multidimensional language assessment results.

5.3.2. Classification of ASD and TD

Our proposed multidimensional language abilities demonstrate significant effectiveness in distinguishing between ASD and TD. Table 8 presents the accuracy, Macro-F1, recall, and precision for the ASD and TD classification. The results show that the classification accuracy for ASD and TD exceeds 0.9 when using the assessment outcomes from Doubao-Pro-32k, GPT4o, and Qwen2-72B-Instruct models as features. Among these, the classifications based on Doubao-Pro-32k and Qwen2-72B-Instruct assessment results perform the best, with both accuracies and Macro-F1 scores reaching 0.98. Doubao-Pro-32k achieves perfect precision (1.0) for ASD and perfect recall (1.0) for TD, while Qwen2-72B-Instruct achieves perfect precision (1.0) for TD and perfect recall (1.0) for ASD. Fig. 6(a) presents the ROC curves for 'ASD vs TD,' with all curves exhibiting a steep upward trend, indicating high true positive rates and low false positive rates. This further confirms the

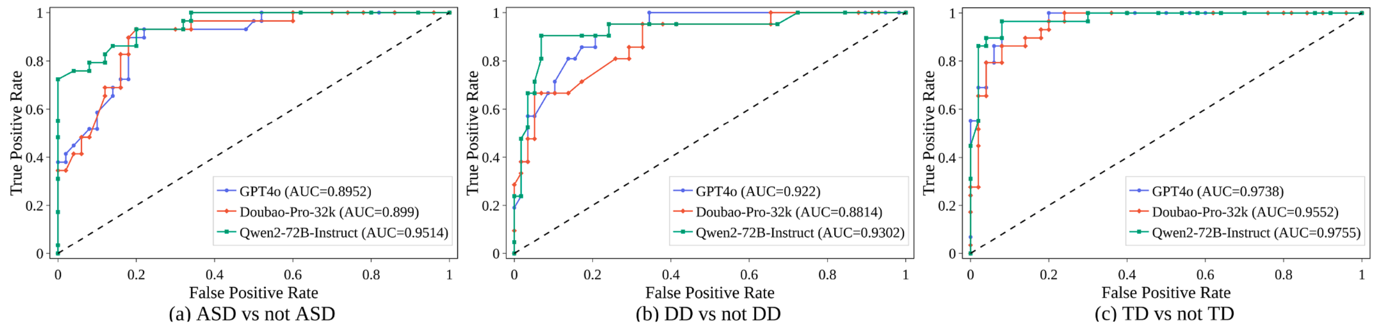

Fig. 5. ROC curves for ASD versus non-ASD, DD versus non-DD, and TD versus non-TD, in the context of a three-class classification.

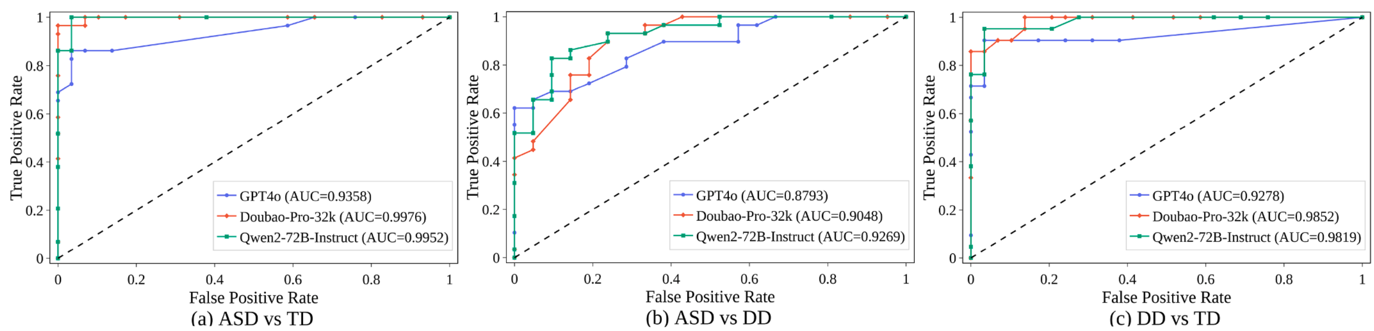

Fig. 6. ROC curves for ASD versus TD, DD versus TD, and ASD versus DD, based on three separate binary classifications.

significant differences in multidimensional language abilities between children with ASD and TD. These language ability features may offer valuable support for the diagnosis of ASD.

5.3.3. Classification of DD and TD

Table 9 presents the classification results of XGBoost based on language assessments from different models for distinguishing between children with DD and TD. The results show that all three models achieve high performance, with Macro-F1 scores ranging from 0.92 to 0.96. Given the slight class imbalance in the dataset (DD = 21, TD = 29), Macro-F1 provides a more reliable evaluation metric than accuracy alone. Among these, Qwen2-72B-Instruct performs the best, with an accuracy of 0.96, a Macro-F1 score of 0.96, precision (DD = 0.95, TD = 0.97), and recall (DD = 0.95, TD = 0.97). Notably, the classification results for 'DD vs TD' are similar to those for 'ASD vs TD', though slightly lower. In Fig. 6, the ROC curves for the 'DD vs TD' classification task all exhibit steep upward trends, with AUC values exceeding 0.9, indicating that the model achieves high accuracy in distinguishing between children with DD and typically developing children. Compared to the 'ASD vs DD' task, this classification task shows steeper curves and superior performance, indicating that the linguistic feature differences between children with DD and typically developing children are more significant than those between children with ASD and children with DD.

5.3.4. Classification of ASD and DD

Compared to the classification tasks of 'ASD vs TD' and 'DD vs TD', the classification of 'ASD vs DD' shows relatively poorer performance. Table 10 indicates that the highest accuracy for this classification task is 0.84, achieved by Qwen2-72B-Instruct with a corresponding Macro-F1 score of 0.83. In terms of recall, the recall rate for children with DD is significantly lower than that for children with ASD. Specifically, when using the assessment results from Doubao-Pro-32k, the recall rate for children with DD is only 0.57. In Fig. 6, the ROC curve for 'ASD vs DD' is closer to the bottom-right corner compared to the 'ASD vs TD' and 'DD vs TD' curves, indicating relatively weaker classification performance. These results are largely consistent with the group differences revealed by Welch's ANOVA test results in Table 2. Table 2 compares the performance of three groups of children on MSEL VDQ, showing that 'ASD vs TD' and 'DD vs TD' yielded *F*-values of 248.28 ($p < 0.001$) and 104.90 ($p < 0.001$), respectively, demonstrating highly significant differences. The corresponding effect sizes (η^2) were 0.82 and 0.72, respectively, both exceeding 0.14, indicating large effects and substantial between-group differences. In contrast, the ASD vs DD comparison showed an *F*-value of 5.94 ($p < 0.05$), indicating a significant difference, but with an effect size (η^2) of only 0.11, falling between 0.06 and 0.14, representing a medium effect and suggesting more limited between-group differences compared to the other comparisons. This finding provides important explanatory evidence for the relatively lower discriminative performance of 'ASD vs DD' in the classification experiments, suggesting that ASD and DD children exhibit greater similarities in language abilities compared to TD children, making discrimination between these two clinical groups more challenging.

6. Discussion and conclusion

This section first examines the role of natural language samples in assessing children's language abilities. It then introduces the advantages and potential of LLMs, explores the adaptability of the proposed assessment framework, and outlines future research directions.

Natural language samples more closely reflect children's actual language performance. Compared to standardized tests, natural language samples can more accurately capture children's language abilities and developmental levels during natural interactions. These samples provide multidimensional insights into language development, such as language comprehension, expressive ability, syntactic complexity, and lexical diversity. Using natural language samples to assess the language abilities of children with autism offers multiple advantages. Firstly, this approach is more convenient, enabling data collection in natural settings, which minimizes the need for their cooperation. Secondly, natural language samples are also suitable for younger children who are unable to independently complete assessments, thereby avoiding difficulties they may encounter in standardized testing. Additionally, by directly analyzing children's conversational performance, this method mitigates potential subjective bias from parental evaluations, providing a more natural and objective means of assessing children's language development.

LLMs have demonstrated significant potential in the medical field. A 2023 study (Kung et al., 2023) showed that ChatGPT, without specialized training, was able to pass or come close to passing the United States Medical Licensing Examination (USMLE). In recent years, with advancements in technology, the capabilities of LLMs have grown significantly stronger (OpenAI, 2024b), and an increasing number of studies have confirmed their value in medical applications (Benoit, 2023; Chen et al., 2023b; Singhal et al., 2023; Xue et al., 2023; Yang et al., 2022). This highlights the extensive medical knowledge LLMs possess and their potential to support medical education. Therefore, LLMs can leverage their accumulated medical knowledge and language comprehension skills to provide robust support in analyzing and evaluating children's language capabilities. However, the application of LLMs in pediatric language assessment raises several critical ethical considerations that must be addressed. First, obtaining informed consent from children's guardians is essential, and studies must undergo appropriate institutional review board approval. Additionally, strict data de-identification protocols must be implemented to remove all personally identifiable information, ensuring compliance with data privacy regulations. When utilizing LLM APIs, researchers must verify whether data collection for model optimization can be disabled, or obtain explicit assurance from providers that user data will not be used for model training or storage. It is imperative to emphasize that AI-powered tools should serve solely as supplementary assessment instruments. Results generated by LLMs must be integrated with professional clinical expertise and comprehensive evaluation to mitigate risks associated with over-dependence on automated systems. As technology advances, the costs of training and deploying LLMs are decreasing. In this study, we tested various open-source and proprietary LLMs, offering users a comprehensive comparison. Users can make more informed decisions based on their needs and budgets, considering factors such as performance, cost, and privacy.

Our assessment framework demonstrates promising potential for general applicability. Although the subjects involved in this study are Mandarin-speaking children, the framework may have the potential to be extended for use with children from other linguistic backgrounds. Theoretically, by adjusting the assessment dimensions, we could optimize for the specific features of different languages in terms of grammar, vocabulary, and pragmatics. Furthermore, the application of this framework may not be limited to children with ASD. With appropriate modifications, we hypothesize that the framework has the potential to be extended to children with other types of language disorders. Additionally, the flexibility of this framework suggests that it may not only be applicable to natural language samples from parent-child free play but could also be expanded to language data in other contexts. However, these potential applications and extensions require further empirical research for validation. Future studies could explore these possibilities to determine the actual applicability of the framework across different languages, disorder types, and contexts.

Overall, this study proposes a more natural, objective, and intelligent approach to assessing the language abilities of children with autism. Building on this foundation, future research could proceed in several directions: integrating multimodal large language models that incorporate speech and behavioral features to improve assessment accuracy; expanding the scope of data collection to include more diverse linguistic and cultural contexts to enhance the model's generalizability; systematically investigating the effects of different prompt variations on assessment outcomes to improve methodological robustness; and incorporating longitudinal data collection mechanisms to enable dynamic tracking of children's language development.

CRediT authorship contribution statement

Saige Qin: Conceptualization, Data curation, Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing; **Min Liu:** Conceptualization, Data curation, Formal analysis, Visualization, Writing – original draft; **Tongquan Wei:** Conceptualization, Data curation, Formal analysis, Supervision; **Qiaoyun Liu:** Formal analysis, Data curation, Writing – review & editing, Supervision.

Data availability

The data that has been used is confidential.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Funding: This work is supported by the Major Program of Humanities and Social Sciences of the Shanghai Education Commission (2023SKZD07); the State Language Commission's "14th Five-Year Plan" Scientific Research Project (YB145-21); the Ministry of Education's Humanities and Social Science project (23YJC740035); the Doctoral Research Innovation Fund Project of East China Normal University (2024BSJJ08); and the China Postdoctoral Science Foundation (2023M731104).

Appendix A. Examples of language proficiency assessment results

Appendix B. Quantitative comparison of bias characteristics in language ability assessment models

Fig. B.1(a)–(d) present the distribution of rating deviations between each model's assessments and expert annotations across four language evaluation dimensions. Fig. B.2 further displays the evaluation deviations for the overall language proficiency dimension. The horizontal axis represents the rating deviation of model evaluations relative to expert evaluations, ranging from -4 to 4 based on the five-point rating scale employed in this study; the vertical axis indicates the percentage of samples corresponding to each deviation value, reflecting the distribution of models across different deviation levels.

The results reveal significant differences in model performance on language evaluation tasks. GPT4o demonstrates the highest consistency across all evaluation dimensions (the four sub-dimensions and overall language proficiency), achieving perfect match rates (Deviation = 0) ranging from 62 % to 71 %. Specifically, in the language comprehension dimension (as illustrated in Fig. B.1(a)), GPT4o attains the highest perfect match rate of 71 % among all evaluated models. Moreover, GPT4o maintains coverage rates of 96 %–100 % for deviations within one rating level ($|\text{Deviation}| \leq 1$) across all evaluation dimensions, demonstrating evaluation performance significantly superior to that of other models.

In contrast, Qwen2-57B-A14B exhibits an overestimation tendency, with mean deviations ranging from +0.65 to +0.9, the highest among all models. The deviation reaches its peak of +0.9 in the language comprehension dimension, and this model is the only one to produce deviations spanning four rating levels, with positive deviation samples accounting for 53 %–63 % of all cases. These characteristics result in perfect match rates of merely 25 % to 34 %, less than half of GPT4o's performance. Furthermore, its coverage rate for deviations within one level ($|\text{Deviation}| \leq 1$) ranges from only 70 % to 79 %, the lowest among all evaluated models.

Doubao-Pro-32k is the only model that demonstrates negative deviations across all dimensions, with mean deviations ranging from

第一步: 理解[## 评估维度]中的内容
Step 1: Comprehend the content presented under the section [## Assessment Dimensions]

1. **语言理解能力**
1. ** Language Comprehension**
- 该维度评估儿童对语言的理解能力，包括字面意义、上下文意义、抽象概念和比喻的解析，主要考察是否存在语义理解困难、答非所问及人称代词分辨问题。
- This dimension assesses children's ability to comprehend language, including literal meanings, contextual meanings, abstract concepts, and metaphors. It focuses on identifying difficulties in semantic comprehension, giving irrelevant responses, and distinguishing personal pronouns.

2. ** 词汇掌握能力**
2. ** Vocabulary Competence**
- 该维度反映儿童词汇使用的丰富性、恰当性及创造性，关注是否词汇贫乏、重复使用词语、模仿家长语言或创造新词。
- This dimension reflects children's ability to use vocabulary with richness, appropriateness, and creativity. It examines whether their vocabulary is limited, whether they frequently repeat the same words, mimic parental speech, or invent new words.

3. **逻辑表达能力**
3. **Logical Expression**
- 该维度衡量儿童表达思想的条理性和逻辑性，重点关注词语使用混乱和信息组织困难的问题。
- This dimension evaluates children's ability to express their thoughts clearly and logically, emphasizing issues such as disorganized word usage and difficulties in structuring information.

4. **沟通交流能力**
4. **Communication Skills**
- 该维度评估儿童在对话中的主动性和连贯性，包括发起、回应、轮替、终止对话的能力，以及维持话题关联性的表现。
- This dimension assesses children's initiative and coherence in conversations, including their ability to initiate, respond to, alternate, and terminate conversations, as well as to maintain topic relevance and focus.

第二步: 基于【第一步】的理解，给出该年龄的儿童在每个维度上的评估标准
Step 2: Based on the understanding from [Step 1], provide the assessment standards for children of this age in each dimension.

1. **语言理解能力**
1. **Language Comprehension**
- 应能理解较为简单的字面意义和上下文意义，尝试理解简单的比喻及抽象概念。
- Children should be able to understand both simple literal and contextual meanings, and they should also attempt to interpret basic metaphors and abstract concepts.
- 非常少发生答非所问的情况，人称代词通常能分辨清楚。
- Irrelevant responses are very rare, and personal pronouns are usually distinguished accurately.

2. **词汇掌握能力**
2. **Vocabulary Competence**
- 应有一定的词汇量，能够使用丰富且合适的词汇表达。
- Children should have a certain vocabulary size and be able to use a rich and appropriate range of words for expression.
- 偶尔会有重复词语、短语，或者家长的话语内容的情况。不频繁使用新异或编造的词语。
- There may occasionally be repetition of words, phrases, or the content of the parent's speech. Novel or invented words are used infrequently.

3. **逻辑表达能力**
3. **Logical Expression**
- 应能基本连贯地表达意见和故事，信息组织较为清晰。
- Children should be able to express opinions and tell stories in a generally coherent manner, with relatively clear organization of information.
- 语法偶尔有错误，但有基本的逻辑和连贯性。
- Occasional grammatical errors are acceptable, but there should be basic logic and coherence in their expression.

4. **沟通交流能力**
4. **Communication Skills**
- 应能够主动发起和回应对话，有一定的轮替和终止对话的能力。
- Children should be able to actively initiate and respond to conversations, with some ability to alternate turns and appropriately end conversations.
- 能根据语境进行有效交际，较少自言自语，能够较好地维持话题。
- They should be able to communicate effectively based on context, rarely talk to themselves, and generally maintain topic relevance.

第三步: 结合儿童和家长的对话内容进行分析理解，给出合理语言等级评估
Step 3: Analyze the child-parent conversation and provide a reasonable language proficiency assessment.

1. **语言理解能力**
1. **Language Comprehension**
- 孩子能理解家长的许多问题并做出回应，如对于榴莲味道和假装吃食物的互动。
- The child can understand parents' questions and respond, such as discussing durian's taste or pretending to eat.
- 孩子的回答大部分是切合主题的，比如给家长解释"这是玩具"，"我帮你切一切"等。
- The child's responses are mostly on-topic, such as explaining to the parent, "This is a toy," or "I'll cut it for you."
- 没有明显的人称代词混淆问题。
- There is no obvious issue with pronoun confusion.
- 评级: 良好
- Rating: Good

2. **词汇掌握能力**
2. **Vocabulary Competence**
- 孩子有一定的词汇量，能使用如"榴莲"，"玩具"，"蟹黄"等词汇。
- The child has a decent vocabulary and can use words such as "durian," "toy," and "crab roe."
- 孩子有一些重复家长话语的现象，如"妈妈，这不用放不下"，但整体词汇使用较丰富。
- The child occasionally repeats the parent's phrases, such as "Mom, this doesn't need to be put down," but overall, the vocabulary is relatively rich.
- 没有频繁使用新异或编造词语。
- The child does not frequently use neologisms or invented words.
- 评级: 良好
- Rating: Good

3. **逻辑表达能力**
3. **Logical Expression**
- 孩子在描述做饭、吃东西、玩积木等环节相对逻辑清晰，如在游戏中描述切香蕉和煮玫瑰茶等事情。
- The child demonstrates relatively clear logic when describing activities such as cooking, eating, and playing with blocks, for example, describing tasks like cutting bananas and making rose tea during play.
- 孩子能够基本连贯地进行表达，有少量语法错误，但总体表达较为清晰。
- The child is able to express themselves fairly coherently, with a few grammatical errors, but overall, the communication is clear.
- 评级：良好
- Rating: Good

4. **沟通交流能力**
4. **Communication Skills**
- 孩子能主动参与对话，回应家长，并能开始和转移话题。
- The child actively engages in conversations, responds to the parent, and is able to initiate and shift topics.
- 孩子能够在互动过程中保持对话的连贯性，如继续谈论煮饭、生鲜食物等。
- The child is able to maintain the coherence of the conversation during interactions, such as continuing to talk about cooking or fresh food.
- 没有自言自语和频繁更换话题的现象，且能跟随家长的引导。
- There is no self-talk or frequent topic switching, and the child is able to follow the parent's lead.
- 评级: 优秀
- Rating: Excellent

第四步: 使用评估维度和等级评估结果以JSON格式汇总评估结果
Step 4: Summarize the evaluation results using the assessment dimensions and levels in JSON format.

```json
{
 "语言理解能力": "良好",
 "Language Comprehension": "Good",
 "词汇掌握能力": "良好",
 "Vocabulary Competence": "Good",
 "逻辑表达能力": "良好",
 "Logical Expression": "Good",
 "沟通交流能力": "优秀"
 "Communication Skills": "Excellent"
}
```

Fig. A.1. GPT4o Assessment Results Example: 61-month-old TD Child.

第一步: 理解[# # 评估维度]中的内容
Step 1: Comprehend the content presented under the section [# # Assessment Dimensions]

1. **语言理解能力**
1. ** Language Comprehension**
- 该维度关注儿童理解语言的深度和准确度。涵盖的范围包括理解字面意义和上下文意义、抽象概念以及修辞手法。此外，还考察儿童是否能正确回答问题和分辨人称代词。
- This dimension focuses on the depth and accuracy of children's language comprehension, including literal and contextual meanings, abstract concepts, and rhetorical devices. It also assesses children's ability to correctly answer questions and distinguish personal pronouns..

2. ** 词汇掌握能力**
2. ** Vocabulary Competence**
- 该维度关注儿童是否掌握并使用丰富的词汇，避免重复使用同样的词语和短句。此外，还考察儿童是否会重复家长的话，或使用新造的词汇。
- This dimension focuses on whether children have mastered and use a diverse vocabulary, avoiding repetition of words and phrases. It also assesses whether children repeat their parents' words or use newly coined terms.

3. **逻辑表达能力**
3. **Logical Expression**
- 该维度关注儿童是否能够连贯、有逻辑地组织语言。评估的范围包括语法正确性、句子的联系，以及表达的连贯性和逻辑性。
- This dimension focuses on whether children can organize language coherently and logically. It assesses grammatical accuracy, sentence cohesion, and the overall coherence and logic of their expression.

4. **沟通交流能力**
4. **Communication Skills**
- 该维度关注儿童在对话中的主动性和交互性。涵盖的范围包括主动发起或回应对话、保持和转换话题的能力、以及避免自言自语和频繁更换话题。
- This dimension focuses on children's initiative and interactivity in conversation. It includes the ability to initiate or respond to dialogue, maintain and switch topics, and avoid talking to themselves or frequently changing topics.

第二步: 基于【第一步】的理解，给出该年龄的儿童在每个维度上的评估标准
Step 2: Based on the understanding from [Step 1], provide the assessment standards for children of this age in each dimension.

1. **语言理解能力**
1. **Language Comprehension**
- 能理解大多数字面和上下文的意思，能理解简单的抽象概念和比喻。
- Can understand most literal and contextual meanings, as well as basic abstract concepts and metaphors.
- 能准确回答问题，主题不偏离，准确使用和分辨一般的人称代词。
- Can answer questions accurately, stay on topic, and correctly use and distinguish general personal pronouns.

2. **词汇掌握能力**
2. **Vocabulary Competence**
- 掌握并能使用丰富的常见词汇，很少或基本不反复使用相同词语、短句。
- Master and effectively use a wide range of common vocabulary, with little to no repetition of the same words or phrases.
- 不经常重复家长的话。不使用编造的词汇或短语。
- Do not frequently repeat the parent's words and avoid using invented vocabulary or phrases.

3. **逻辑表达能力**
3. **Logical Expression**
- 词语使用准确，不混乱。
- The use of vocabulary is precise and clear.
- 语句有连贯性，语法结构较完整。
- The sentences are coherent, and the grammatical structure is complete.
- 信息组织明确，表达具有逻辑性。
- The information is well-organized, and the expression is logical.

4. **沟通交流能力**
4. **Communication Skills**
- 能主动发起、回应、轮替、终止对话。
- Can initiate, respond, alternate, and terminate conversations.
- 沟通较为顺畅，不存在明显困难。很少自言自语。
- Communication is smooth, with no noticeable difficulties. Self-talk is infrequent.
- 通常不会频繁更换话题。
- The child typically does not frequently switch topics.

第三步: 结合儿童和家长的对话内容进行分析理解，给出合理语言等级评估
Step 3: Analyze the child-parent conversation and provide a reasonable language proficiency assessment.

1. **语言理解能力**
1. **Language Comprehension**
- 在对话中，孩子回应“恩”次数较多，没有明显的答非所问，也没有主动表达对复杂内容的理解。
- In the conversation, the child responds with "hmm" frequently, without giving irrelevant answers or actively expressing an understanding of complex topics.
- 理解家长对具体行为的指示（如“用右手画”、“插在空的地方”），但未主动表达对抽象内容的理解。代词未大量出现，无法评估代词分辨能力。
- The child understands specific instructions (e.g., "draw with your right hand," "insert it into the empty space"), but does not express abstract understanding. Pronouns are infrequent, making it hard to assess pronoun differentiation ability.
综合分析：孩子能够理解基本的指令和对话内容，但缺乏主动进行复杂表达的能力，表现出其抽象理解能力的不足。
Comprehensive analysis: The child can understand basic instructions and conversational content, but lacks the ability to proactively express complex ideas, indicating a deficiency in abstract understanding.
- 评级: 一般
- Rating: Fair

2. **词汇掌握能力**
2. **Vocabulary Competence**
- 孩子回应中多次重复“恩”，词汇使用相对单一。
- The child repeatedly responds with "hmm," and their vocabulary usage is relatively limited.
- 孩子没有表现出使用新异、编造词汇的情况。没有重复家长话语的现象。
- The child did not exhibit the use of novel or invented words, nor did they repeat the parent's words.
综合分析：孩子的词汇使用较单一，基本只用“恩”应答，未展现出丰富词汇掌握能力。
Comprehensive analysis: The child's vocabulary usage is quite limited, mainly responding with "hmm," without showing a rich vocabulary
- 评级: 差
- Rating: Poor

3. **逻辑表达能力**
3. **Logical Expression**
- 孩子语言再现较少，大部分回应为单字“恩”，未能有效评估语句连贯性和信息组织；
- The child’s language reproduction is minimal, with most responses being the single word "hmm," making it difficult to effectively assess sentence coherence and information organization.
综合分析：由于孩子的语言输出较少，难以全面分析。
Comprehensive analysis: Due to the child's limited language output, a comprehensive analysis is difficult to conduct.
- 评级：一般
- Rating: Fair

4. **沟通交流能力**
4. **Communication Skills**
- 孩子主动发起、回应对话不多，存在一定沟通被动性和单一性。
- The child initiates and responds to conversation infrequently, demonstrating a certain level of passivity and lack of variety in communication.
- 未出现频繁更换话题，自言自语或只聚焦自己话题的情况。
- There is no frequent topic switching, self-talk, or sole focus on personal topics.
综合分析：孩子主要以“恩”回应对话，主动发起、轮替对话能力差，未表现主动引导话题的能力。
Comprehensive analysis: The child primarily responds with "hmm," has poor ability to initiate or take turns in conversation, and does not demonstrate the ability to actively guide the topic.
- 评级: 差
- Rating: Poor

第四步: 使用评估维度和等级评估结果以JSON格式汇总评估结果
Step 4: Summarize the evaluation results using the assessment dimensions and levels in JSON format.

```
```json
{
 "语言理解能力": "一般",
 "Language Comprehension": "Fair",
 "词汇掌握能力": "差",
 "Vocabulary Competence": "Poor",
 "逻辑表达能力": "一般",
 "Logical Expression": "Fair",
 "沟通交流能力": "差"
 "Communication Skills": "Poor"
}
```
```

Fig. A.2. GPT4o Assessment Results Example: 66-month-old ASD Child.

第一步: 理解[## 评估维度]中的内容
Step 1: Comprehend the content presented under the section [## Assessment Dimensions]

1. **语言理解能力**
1. ** Language Comprehension**
- 语义理解困难：儿童是否难以理解字面意义、上下文意义、抽象概念或修辞手法。
- Semantic Comprehension Difficulties: Whether children have difficulty understanding literal meanings, contextual meanings, abstract concepts, or rhetorical devices.
- 答非所问：儿童是否经常主题偏离或误解问题。
- Irrelevant responses: Do children frequently deviate from the topic or misunderstand questions.
- 人称代词分辨不清：儿童是否难以区分"你、我、他"等代词。
- Personal pronoun confusion: Whether children have difficulty distinguishing between pronouns such as 'you,' 'I,' and 'he/she'.

2. ** 词汇掌握能力**
2. ** Vocabulary Competence**
- 词汇：儿童是否能使用丰富的词汇。
- Vocabulary: Whether children are able to use rich vocabulary.
- 重复使用词语和家长话语：儿童是否频繁使用相同的词语、短语或短句，是否经常重复家长的话。
- Repetitive use of words and parental speech: Whether children frequently use the same words, phrases, or short sentences, and whether they often repeat their parents' words.
- 使用新异或编造词汇：儿童是否使用不符合场景的词汇。
- Use of novel or made-up vocabulary: Whether children use words that are inappropriate for the context.

3. **逻辑表达能力**
3. **Logical Expression**
- 词语混乱使用：儿童是否说出与情境不相关的词语。
- Confused word usage: Whether children say words that are unrelated to the situation.
- 信息组织困难：儿童是否表达缺乏逻辑性和连贯性。
- Difficulty organizing information: Whether children's expression lacks logic and coherence.

4. **沟通交流能力**
4. **Communication Skills**
- 主动发起、回应、轮替、终止对话：儿童是否能有效参与对话。
- Initiating, responding, turn-taking, and ending conversations: Whether children can effectively participate in conversations.
- 沟通交流困难：儿童是否难以引导话题或根据语境交际，是否经常自言自语。
- "Communication difficulties: Whether children have trouble guiding topics or communicating according to context, and whether they often talk to themselves.
- 话题：儿童是否频繁改变话题、是否只关注自己感兴趣的话题。
- Topics: Whether children frequently change topics, and whether they only focus on topics that interest them.

第二步: 基于【第一步】的理解，给出该年龄的儿童在每个维度上的评估标准
Step 2: Based on the understanding from [Step 1], provide the assessment standards for children of this age in each dimension.

1. **语言理解能力**
1. **Language Comprehension**
- 能理解抽象概念和修辞手法，较少出现答非所问，能正确使用人称代词。
- Can understand abstract concepts and rhetorical devices, rarely gives irrelevant responses, and can correctly use personal pronouns.

2. **词汇掌握能力**
2. **Vocabulary Competence**
- 能使用丰富的词汇，较少重复使用词语、短语或短句，较少重复家长的话，不使用新异或编造词汇。
- Can use rich vocabulary, rarely repeats words, phrases, or short sentences, seldom repeats parents' words, and does not use novel or made-up vocabulary.

3. **逻辑表达能力**
3. **Logical Expression**
- 能使用词语恰当，表达逻辑性和连贯性较好。
- Can use words appropriately, with good logical and coherent expression.

4. **沟通交流能力**
4. **Communication Skills**
- 能主动发起、回应、轮替、终止对话，能根据语境进行有效交际，较少自言自语，较少频繁更换话题，能关注不同话题。
- Can actively initiate, respond, turn-take, and end conversations, able to communicate effectively according to context, rarely engages in self-talk, seldom frequently changes topics, and can focus on different topics.

第三步: 结合儿童和家长的对话内容进行分析理解，给出合理语言等级评估
Step 3: Analyze the child-parent conversation and provide a reasonable language proficiency assessment.

1. **语言理解能力**
1. **Language Comprehension**
- 儿童能理解家长的指令，如"插进去"、"画一个龙卷风"等，没有出现明显的语义理解困难。
- Children can understand parents' instructions, such as 'insert it' and 'draw a tornado,' without showing obvious semantic comprehension difficulties.
- 儿童能正确回应家长的问题，如"你在看什么啊"，没有出现答非所问的情况。
- Children can correctly respond to parents' questions, such as 'What are you looking at?', without giving irrelevant responses.
- 儿童能正确使用人称代词，如"妈妈一张，宝宝一张"。
- Children can correctly use personal pronouns, such as 'One for mommy, one for baby.'

2. **词汇掌握能力**
2. **Vocabulary Competence**
- 儿童能使用丰富的词汇，如"龙卷风"、"一"、"小脚丫"等。
- Children can use rich vocabulary, such as 'tornado,' 'one,' and 'little feet.'
- 儿童没有频繁使用相同的词语、短语或短句，也没有重复家长的话
- Children do not frequently use the same words, phrases, or short sentences, nor do they repeat their parents' words.
- 儿童没有使用新异或编造的词汇。
- Children do not use novel or made-up vocabulary.

3. **逻辑表达能力**
3. **Logical Expression**
- 儿童能使用词语恰当，如"龙卷风"、"一"等。
- Children can use words appropriately, such as 'tornado' and 'one.'
- 儿童表达逻辑性和连贯性较好，如在拼图过程中能按照家长的指导进行。
- Children's expression shows good logic and coherence, such as being able to follow parents' guidance during puzzle activities.

4. **沟通交流能力**
4. **Communication Skills**
- 儿童能主动发起、回应、轮替、终止对话，如"恩"、"哎"等回应。
- Children can actively initiate, respond, turn-take, and end conversations, such as responses like 'mm-hmm' and 'hey.'
- 儿童能根据语境进行有效交际，如在拼图过程中能理解家长的指导。
- Children can communicate effectively according to context, such as being able to understand parents' guidance during puzzle activities.
- 儿童较少自言自语，没有频繁更换话题，能关注不同话题。
- Children rarely engage in self-talk, do not frequently change topics, and can focus on different topics.

第四步: 使用评估维度和等级评估结果以JSON格式汇总评估结果
Step 4: Summarize the evaluation results using the assessment dimensions and levels in JSON format.

````
```json
{
 "语言理解能力": "优秀",
 "Language Comprehension": "Excellent",
 "词汇掌握能力": "优秀",
 "Vocabulary Competence": "Excellent",
 "逻辑表达能力": "优秀",
 "Logical Expression": "Excellent",
 "沟通交流能力": "优秀"
 "Communication Skills": "Excellent"
}
```
````

Fig. A.3. Qwen2-57B-A14B-Instruct Assessment Results Example: 66-month-old ASD Child.

第一步: 理解[## 评估维度]中的内容
Step 1: Comprehend the content presented under the section [## Assessment Dimensions]

1. **语言理解能力**
1. ** Language Comprehension**
- 关注儿童是否存在理解语义困难，例如：难以理解字面意义、上下文、抽象概念以及修辞手法等；答非所问，人称代词使用不当等。
- Examining whether children exhibit difficulties in understanding semantics, including challenges in comprehending literal meanings, contextual nuances, abstract concepts, and rhetorical devices; giving irrelevant responses to questions, or improper use of personal pronouns.

3. **逻辑表达能力**
3. **Logical Expression**
- 关注儿童在语言表达中的逻辑性，是否出现混乱使用词语，表达是否连贯，语法是否正确等。
- Examining the logicality of children's language expression, including whether they misuse words, whether their expressions are coherent, and whether their grammar is accurate.

2. ** 词汇掌握能力**
2. ** Vocabulary Competence**
- 关注儿童的词汇丰富度，是否反复使用相同词语，是否多次重复家长的话语，是否使用编造的单词或短语表达。
- Examining children's lexical richness, including whether they repeatedly use the same words, frequently repeat their parents' speech, or use invented words or phrases for expression.

4. **沟通交流能力**
4. **Communication Skills**
- 关注儿童是否能够主动发起、回应、轮替、终止对话；是否存在因引导话题或保持话题的能力不足而导致的交流困难，是否自言自语或频繁更换话题等。
- Examining whether children are able to actively initiate, respond to, take turns in, and terminate conversations; whether communication difficulties result from a lack of ability to guide or sustain topics; and whether they engage in self-talk or frequently change topics.

第二步: 基于【第一步】的理解，给出该年龄的儿童在每个维度上的评估标准
Step 2: Based on the understanding from [Step 1], provide the assessment standards for children of this age in each dimension.

1. **语言理解能力**
1. **Language Comprehension**
- 应能理解基本词汇和简单句子，能正确理解抽象概念的简单形式。
- Children should be able to understand basic vocabulary and simple sentences, as well as correctly comprehend simple forms of abstract concepts.
- 对简单的比喻和修辞手法具备初步理解能力，人称代词使用基本正确。
- Children should have a basic understanding of simple metaphors and rhetorical devices, and use personal pronouns correctly.

3. **逻辑表达能力**
3. **Logical Expression**
- 能基本连贯地表达自己的意思。
- Children should be able to express their thoughts coherently.
- 语法结构较为简单但基本准确。
- Children should use relatively simple but mostly accurate grammatical structures.
- 语法错误和表达不连贯的情况不应频繁出现。
- Grammatical errors and incoherent expressions should not occur frequently.

2. **词汇掌握能力**
2. **Vocabulary Competence**
- 具有较丰富的常见词汇，偶尔重复家长的话语内容属正常现象，但不应过于频繁。
- Children should have a relatively rich vocabulary of common words. Occasional repetition of their parents' speech is normal, but it should not happen too frequently.
- 应能正确使用常见词汇和短语，不常出现编造的单词。
- Children should be able to use common vocabulary and phrases correctly, with few instances of invented words.

4. **沟通交流能力**
4. **Communication Skills**
- 能主动发起、回应、轮替对话。
- Children should be able to actively initiate, respond to, and take turns in conversation.
- 能较好地保持话题的一致性。
- Children should be able to maintain topic consistency effectively.
- 偶尔自言自语或切换话题不应影响整体互动。
- Occasional self-talk or topic switching should not disrupt the overall interaction.

第三步: 结合儿童和家长的对话内容进行分析理解，给出合理语言等级评估
Step 3: Analyze the child-parent conversation and provide a reasonable language proficiency assessment.

1. **语言理解能力**
1. **Language Comprehension**
- 孩子能理解家长提问的小宝宝和螃蟹的相关问题，但有时固定在"吃"这一概念上，未完全能灵活理解上下文的变化（如家长多次提示没有牙齿不能吃，孩子仍然坚持说可以吃）。
- The child can understand questions from the parent about a baby and a crab, but sometimes fixates on the concept of "eating" and struggles to fully understand the changes in context (e.g., despite repeated reminders that the baby cannot eat because they have no teeth, the child continues to insist that they can eat).
- 人称代词使用基本正确，如"小宝宝吃饱了，小宝宝开心"。
- The use of personal pronouns is generally correct, such as in the sentence "The baby is full, and the baby is happy."
- 评级: 一般
- Rating: Fair

3. **逻辑表达能力**
3. **Logical Expression**
- 孩子有时表达不连贯和跳跃，如从"吃螃蟹"跳到"紫色好朋友"。
- The child sometimes exhibits incoherent and disjointed expressions, such as jumping from "eating crab" to "purple good friend."
- 在游戏假想情境中，能产生一系列逻辑简洁的对话，但表述上存在连贯性和逻辑性不足。
- In imaginative play scenarios, the child can produce a series of logically simple dialogues, but there are issues with coherence and logical consistency in their expressions.
- 评级：一般
- Rating: Fair

2. **词汇掌握能力**
2. **Vocabulary Competence**
- 孩子能使用较丰富的词汇（例如涉及不同情境的词汇，如"披萨"、"螃蟹"、"葡萄"等）。
- The child is able to use a relatively rich vocabulary, including words related to different contexts, such as "pizza," "crab," and "grapes."
- 存在较多重复家长话语的现象，如多次重复"吃螃蟹""小宝宝"等。
- There is a frequent occurrence of repeating the parent's speech, such as repeating phrases like "eat crab" and "baby."
- 评级: 一般
- Rating: Fair

4. **沟通交流能力**
4. **Communication Skills**
- 孩子能主动发起对话，例如"给我一个披萨吧"。
- The child can actively initiate a conversation, such as saying, "Give me a pizza."
- 存在频繁更改话题和自言自语的现象，如从吃东西到讨论颜色，再到想象游戏。
- There is a frequent occurrence of topic shifting and self-talk, such as moving from talking about food to discussing colors, and then to imaginative play.
- 评级: 一般
- Rating: Fair

第四步: 使用评估维度和等级评估结果以JSON格式汇总评估结果
Step 4: Summarize the evaluation results using the assessment dimensions and levels in JSON format.

```json
{
 "语言理解能力": "一般",
 "Language Comprehension": "Fair",
 "词汇掌握能力": "一般",
 "Vocabulary Competence": "Fair",
 "逻辑表达能力": "一般",
 "Logical Expression": "Fair",
 "沟通交流能力": "一般"
 "Communication Skills": "Fair"
}
```

Fig. A.4. GPT4o Assessment Results Example: 69-month-old DD Child.

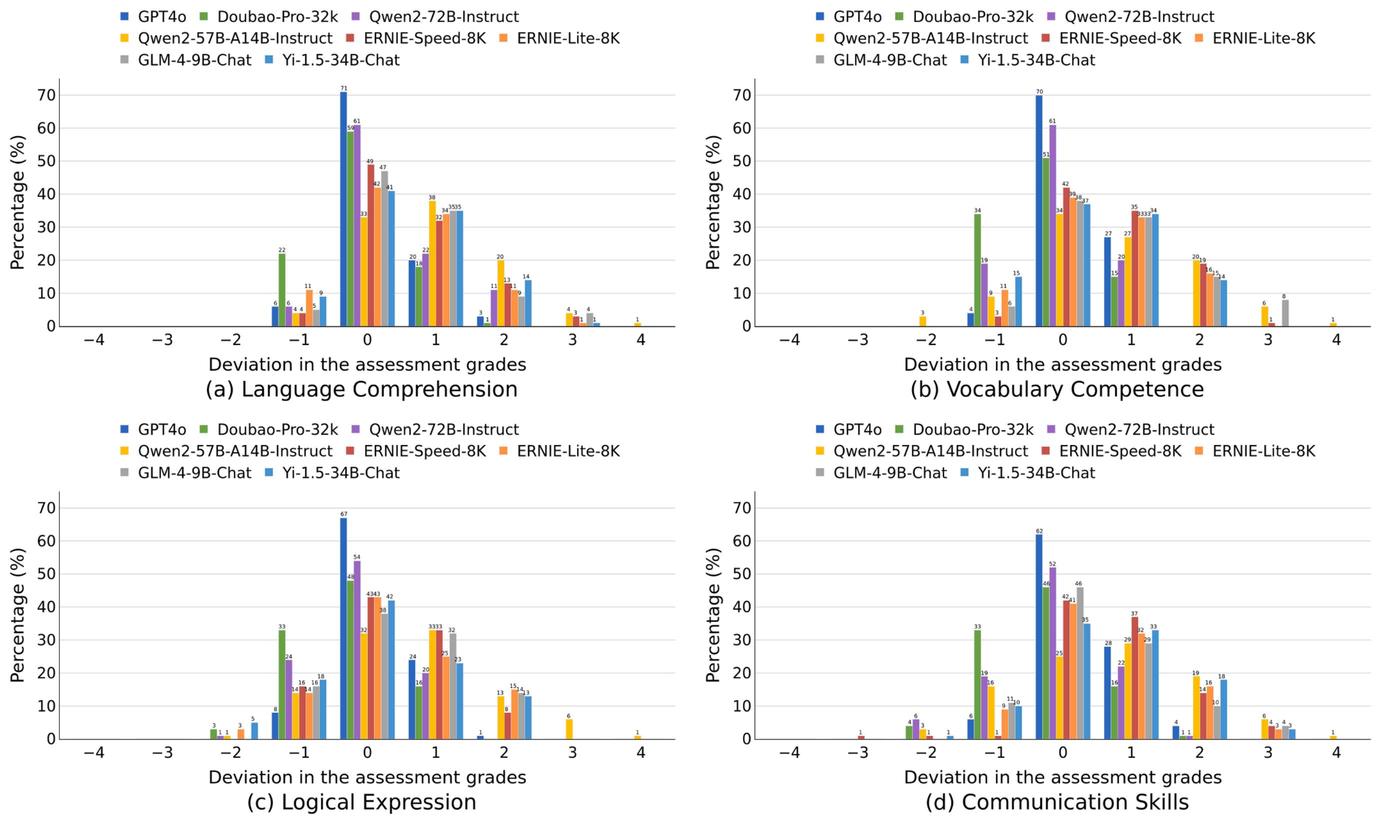

Fig. B.1. Assessment bias of different models in each dimension. The horizontal axis represents the rank bias of model evaluation results relative to expert results, ranging from -4 to 4, determined by the 5-level evaluation system; the vertical axis represents the percentage corresponding to each bias interval.

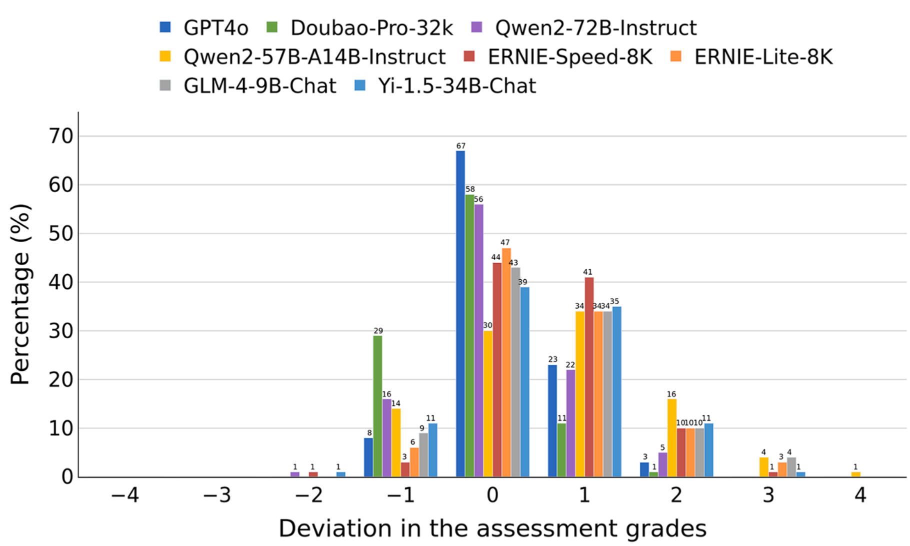

Fig. B.2. Assessment bias of different models in the overall language ability dimension. The horizontal axis represents the rank bias of model evaluation results relative to expert results, ranging from -4 to 4, determined by the 5-level evaluation system; the vertical axis represents the percentage corresponding to each bias interval.

−0.23 to −0.02. Despite this conservative scoring tendency, it maintains acceptable deviation coverage rates of 95 % to 100 % across all dimensions, demonstrating good assessment performance. Qwen2-72B-Instruct shows relatively balanced performance, achieving perfect match rates of 52 %–61 % and exhibiting a deviation of only +0.14 in the overall language proficiency dimension, which represents the smallest absolute deviation among all models. The remaining models, including the ERNIE series, GLM-4-9B-Chat, and Yi-1.5-34B-Chat, all exhibit moderate overestimation tendencies, with mean deviations ranging from +0.21 to +0.81.

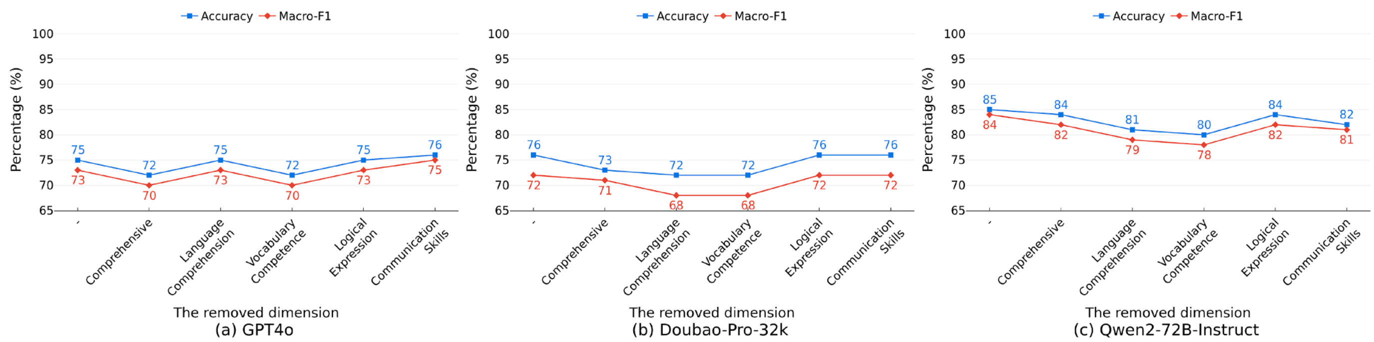

Fig. C.1. Impact of individual language ability dimensions on ASD/DD/TD classification performance using XGBoost.

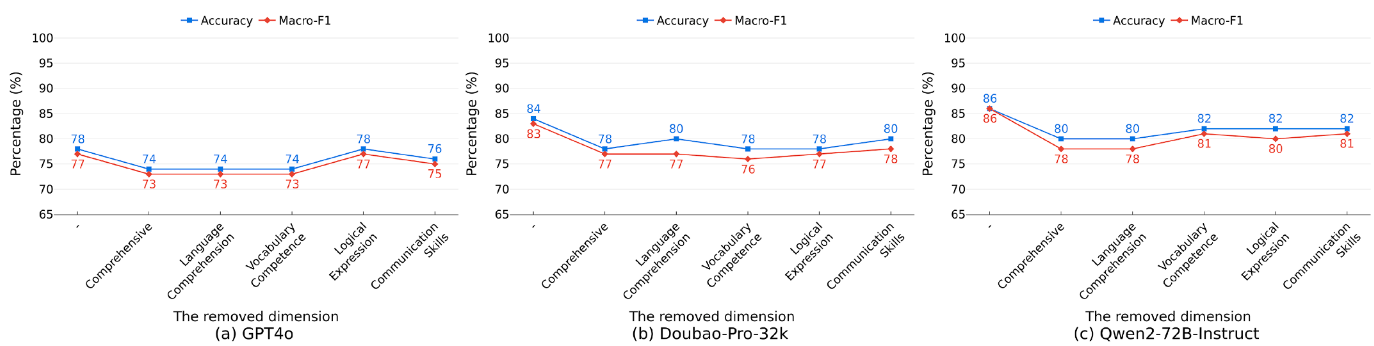

Fig. C.2. Impact of individual language ability dimensions on ASD vs DD classification performance using XGBoost.

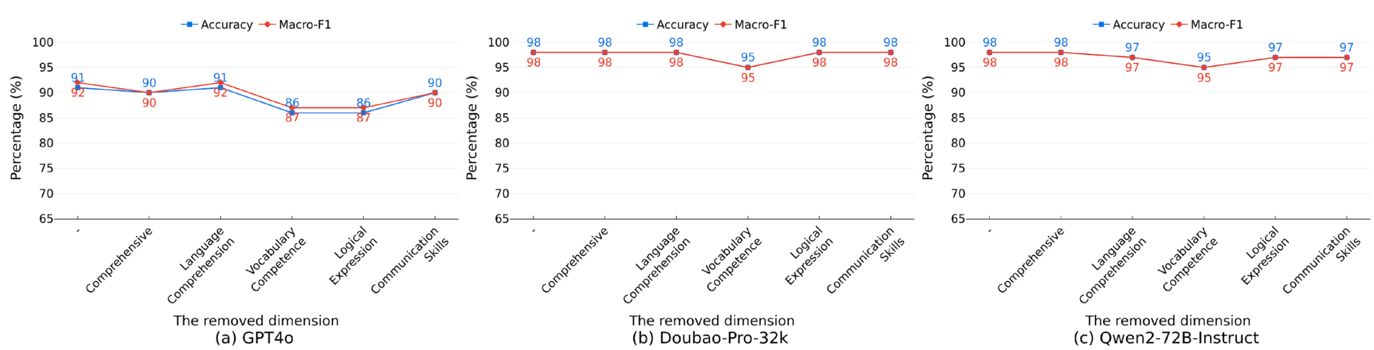

Fig. C.3. Impact of individual language ability dimensions on ASD vs TD classification performance using XGBoost.

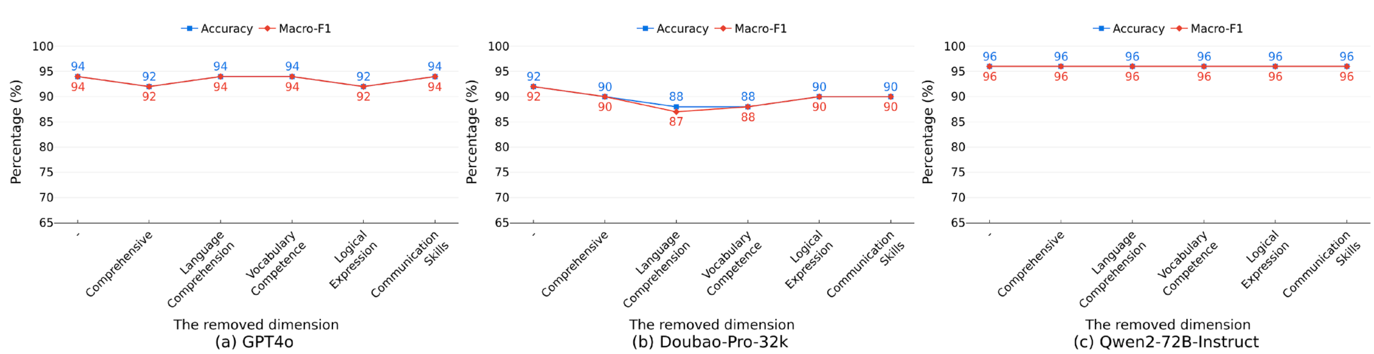

Fig. C.4. Impact of individual language ability dimensions on DD vs TD classification performance using XGBoost.

Appendix C. Impact of individual assessment dimension results on classification performance

Figs. C.1–C.4 show the impact of different language proficiency dimensions on classification task results. The experiments systematically remove features from each dimension for statistical analysis. The x-axis represents six feature configurations: the complete set of four language dimensions plus overall language proficiency, and five configurations with overall language proficiency, language comprehension, vocabulary competence, logical expression, and communication skills individually removed. The y-axis shows model classification performance for each configuration. The two lines represent performance using Accuracy and Macro-F1 evaluation metrics.

The experimental results demonstrate that in the XGBoost classification tasks based on evaluation outcomes from LLMs, the Qwen2-72B-Instruct model performs the best, achieving an average accuracy of 91.25 % across four classification tasks. This significantly outperforms Doubao-Pro-32k (87.5 %) and GPT-4o (84.5 %). Analysis of task difficulty shows that binary classification tasks involving typically developing children (ASD vs. TD, DD vs. TD) generally achieve accuracy rates above 94 %, whereas tasks distinguishing between autism spectrum disorder and developmental delay (ASD vs. DD, ASD vs. DD vs. TD) show relatively lower accuracy (79–82 %). This reflects the similarity in language performance between ASD and DD groups and the complexity of clinical diagnosis.

Comparative analysis of the impact of each ability dimension on classification results:

- Overall Language Ability: Removing this dimension leads to an average decrease of 2.3 % in both Accuracy and Macro-F1 across all tasks, with a more pronounced effect in the complex ASD-DD classification task, where Accuracy drops by up to 6 %. This indicates that this dimension, as a composite measure of overall language ability, holds significant discriminative value.
- Language Comprehension: Its removal causes an average decline of 2.3 % in Accuracy and 2.8 % in Macro-F1. The greatest impact is observed in the ASD vs. DD vs. TD classification, with average decreases of 4.2 % in Accuracy and 6.0 % in Macro-F1, highlighting the critical role of language comprehension in distinguishing complex cognitive impairments.
- Vocabulary Competence: Removing this dimension results in the largest average performance drop among all dimensions, with Accuracy decreasing by 3.4 % and Macro-F1 by 3.7 %. As a fundamental indicator of language development, it exhibits stable discriminative power and plays a crucial role across all tasks.
- Logical Expression: Its removal leads to an average decrease of 1.8 % in Accuracy and 2.0 % in Macro-F1.
- Communication Skills: Removing this dimension results in an average decline of 1.3 % in Accuracy and 1.5 % in Macro-F1.

It is noteworthy that the “Communication Skills” dimension exhibits a counterproductive effect in the classification results based on the GPT-4o model evaluation. In the ASD vs. DD vs. TD classification task, removing this dimension leads to an improvement in accuracy from 75 % to 76 %, and in Macro-F1 from 73 % to 75 %. This phenomenon may be attributed to: the model’s tendency to overfit, whereby XGBoost may excessively rely on noisy signals within the communication skills features; and task-specific feature suitability, as the differences in communication skills between children with ASD and DD are less distinct than those in other dimensions, thus interfering with the effectiveness of core discriminative features.

Overall, in high-difficulty classification tasks (ASD vs. DD, ASD vs. DD vs. TD), removing any individual language ability dimension causes varying degrees of performance degradation. This effect is particularly pronounced in the ASD vs. DD task, where the removal of any dimension results in an average decrease of 4.3 % in Accuracy and 5.1 % in Macro-F1. These findings indicate that distinguishing complex disorders requires a comprehensive evaluation of language abilities. In contrast, in lower-difficulty tasks (ASD vs. TD, DD vs. TD), the impact of removing each dimension is relatively balanced (ranging from 1.2 % to 1.5 %), yet vocabulary competence remains critically important, demonstrating its stable discriminative value in differentiating between disorder and typical development.

Appendix D. Performance comparison of multiple classification models

Tables D.1–D.4 present a comparison of classification results using different classifiers (SVM, XGBoost, RF). The classification tasks include three binary classification tasks (ASD vs. DD, ASD vs. TD, DD vs. TD) and one three-class classification task (ASD vs. DD vs. TD).

The experimental results demonstrate the following overall trends: In the “ASD vs. TD” and “DD vs. TD” binary classification tasks, both overall accuracy and Macro-F1 scores are generally high, with models exhibiting stable performance. However, performance is relatively weaker in the “ASD vs. DD” binary classification task and the three-class classification task. Particularly in the three-class task, the DD category shows the poorest recognition performance, with recall rates ranging from a minimum of 0 to a maximum of only 0.67, indicating significant identification difficulties. This finding suggests that the linguistic features of children with DD may share certain similarities with those of children with ASD, thereby increasing the difficulty for models to distinguish between these groups. All three classifiers demonstrate consistency in exhibiting these trends.

Further analysis reveals notable differences in performance across different classifiers for various tasks. Tree-based classifiers (such as Random Forest and XGBoost) generally outperform Support Vector Machine (SVM) in most tasks, particularly achieving near-perfect results in binary classification tasks. For instance, in the “ASD vs. TD” task based on Doubao-Pro-32k evaluation results, the RF classifier achieves perfect accuracy (1.0). In contrast, SVM demonstrates relatively weaker overall performance, with particularly inadequate recognition capability in tasks involving DD. Notably, in both the “ASD vs. DD” and “ASD vs. DD vs. TD” tasks, SVM achieves zero recall for the DD category, and since the model makes no predictions for this category, precision cannot be calculated, indicating complete classification failure for this category. Several factors may contribute to this phenomenon: First, the high similarity in linguistic features between children with ASD and children with DD (as demonstrated in Section 3.1) may result in ambiguous feature boundaries, reducing the model’s discriminative capacity. Second, class imbalance exists in the dataset, with ASD samples (29 cases) outnumbering DD samples (21 cases). During training, SVM tends to bias toward predicting the more prevalent ASD category, leading to systematic neglect of the DD category. These results highlight that when confronted with highly similar inter-class features combined with imbalanced sample distribution, SVM models struggle to establish effective decision boundaries, resulting in significant limitations in recognition performance.

Table D.1
Three-class classification results for children with ASD, DD, and TD using multiple classification models. Bold and underlined values denote the optimal performance among the three classification models for each assessment condition. "-" indicates that the metric is undefined due to zero denominator (the classifier made no predictions for this class).

Assessment model	Classification model	Accuracy	Macro-F1	Precision			Recall		
				ASD	DD	TD	ASD	DD	TD
Doubao-Pro-32k	RF	**0.78**	**0.74**	**0.75**	0.8	**0.82**	**0.93**	**0.38**	**0.93**
	SVM	0.61	0.47	0.56	–	0.66	0.79	0	0.86
	XGBoost	0.72	0.67	0.73	**0.86**	0.69	0.83	0.29	**0.93**
GPT4o	RF	0.72	0.7	0.69	0.69	0.76	**0.62**	0.52	**0.97**
	SVM	0.7	0.67	0.64	**0.75**	0.72	**0.62**	0.43	**0.97**
	XGBoost	**0.75**	**0.73**	**0.73**	0.7	**0.78**	0.59	**0.67**	**0.97**
Qwen2-72B-Instruct	RF	0.7	0.63	0.61	0.83	0.78	0.86	0.24	0.86
	SVM	0.66	0.62	0.59	0.6	0.75	0.76	0.29	0.83
	XGBoost	**0.82**	**0.8**	**0.76**	**0.86**	**0.87**	**0.9**	**0.57**	**0.93**

Table D.2
Binary classification of children with ASD and TD using multiple classification models. Bold and underlined values denote the optimal performance among the three classification models for each assessment condition.

Assessment model	Classification model	Accuracy	Macro-F1	Precision		Recall	
				ASD	TD	ASD	TD
Doubao-Pro-32k	RF	**1**	**1**	**1**	**1**	**1**	**1**
	SVM	0.91	0.92	0.96	0.88	0.86	0.97
	XGBoost	0.98	0.98	**1**	0.97	0.97	1
GPT4o	RF	**0.91**	**0.92**	**0.96**	**0.88**	**0.86**	**0.97**
	SVM	0.83	0.83	0.95	0.76	0.69	0.97
	XGBoost	**0.91**	**0.92**	**0.96**	**0.88**	**0.86**	**0.97**
Qwen2-72B-Instruct	RF	0.88	0.88	0.82	0.96	0.97	0.79
	SVM	0.84	0.85	0.88	0.81	0.79	0.9
	XGBoost	**0.98**	**0.98**	**0.97**	**1**	**1**	**0.97**

Table D.3
Binary classification of children with DD and TD using multiple classification models. Bold and underlined values denote the optimal performance among the three classification models for each assessment condition.

Assessment model	Classification model	Accuracy	Macro-F1	Precision		Recall	
				DD	TD	DD	TD
Doubao-Pro-32k	RF	**1**	**1**	**1**	**1**	**1**	**1**
	SVM	0.74	0.72	0.75	0.74	0.57	0.86
	XGBoost	0.92	0.92	0.95	0.9	0.86	0.97
GPT4o	RF	**0.96**	**0.96**	**1**	**0.94**	**0.9**	**1**
	SVM	0.92	0.92	0.9	0.93	0.9	0.93
	XGBoost	0.94	0.94	0.95	0.93	0.9	0.97
Qwen2-72B-Instruct	RF	0.94	0.94	0.95	0.93	0.9	**0.97**
	SVM	0.84	0.84	0.84	0.84	0.76	0.9
	XGBoost	**0.96**	**0.96**	**0.95**	**0.97**	**0.95**	**0.97**

Table D.4
Binary classification of children with ASD and DD using multiple classification models. Bold and underlined values denote the optimal performance among the three classification models for each assessment condition. "-" indicates that the metric is undefined due to zero denominator (the classifier made no predictions for this class).

Assessment model	Classification model	Accuracy	Macro-F1	Precision		Recall	
				ASD	DD	ASD	DD
Doubao-Pro-32k	RF	**0.84**	**0.84**	**0.84**	0.84	0.9	**0.76**
	SVM	0.58	0.37	0.58	–	1	0
	XGBoost	0.82	0.79	0.76	**1**	**1**	0.57
GPT4o	RF	0.72	0.7	0.71	0.73	0.86	0.52
	SVM	0.74	0.72	0.74	0.75	**0.86**	0.57
	XGBoost	**0.78**	**0.77**	**0.8**	**0.75**	0.83	**0.71**
Qwen2-72B-Instruct	RF	**0.86**	**0.86**	**0.84**	**0.89**	**0.93**	**0.76**
	SVM	0.7	0.67	0.69	0.71	0.86	0.48
	XGBoost	0.84	0.83	0.82	0.88	0.93	0.71

References

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M. et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, *183*(6), 589–596.

Benoit, J. R. A. (2023). ChatGPT for clinical vignette generation, revision, and evaluation. *MedRxiv*, (pp. 2023–02).

Boucher, J. (2012). Research review: Structural language in autistic spectrum disorder–characteristics and causes. *Journal of Child Psychology and Psychiatry*, *53*(3), 219–233.

Caruccio, L., Cirillo, S., Polese, G., Solimando, G., Sundaramurthy, S., & Tortora, G. (2024). Can chatGPT provide intelligent diagnoses? A comparative study between predictive models and chatGPT to define a new medical diagnostic bot. *Expert Systems with Applications*, *235*, 121186.

Chen, J., Wang, X., Xu, R., Yuan, S., Zhang, Y., Shi, W., Xie, J., Li, S., Yang, R., Zhu, T. et al. (2024). From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Chen, Y., Zheng, S., Wang, H., Cheng, L., Chen, Q., & Qi, J. (2023a). An enhanced res2net with local and global feature fusion for speaker verification. *arXiv preprint arXiv:2305.12838*.

Chen, Z., Balan, M. M., & Brown, K. (2023b). Boosting transformers and language models for clinical prediction in immunotherapy. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 5: Industry track)* (pp. 332–340).

Coplan, J. (2005). Early language milestone scale. *Edisi ke-2*.

DA, K. (1978). Autism screening instrument for educational planning: Background and development. *Autism: Diagnosis, Instruction, Management and Research*.

DeepSeek-AI (2024). Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Dunn, D. M., & Dunn, D M. (2019). Peabody picture vocabulary test [measurement instrument]. *NCS Pearson*.

Eigsti, I. M., Bennetto, L., & Dadlani, M. B. (2007). Beyond pragmatics: Morphosyntactic development in autism. *Journal of Autism and Developmental Disorders*, *37*, 1007–1023.

Feng, T., Xu, A., Lahiri, R., Tager-Flusberg, H., Kim, S. H., Bishop, S., Lord, C., & Narayanan, S. (2024). Can generic LLMs help analyze child-adult interactions involving children with autism in clinical observation? https://arxiv.org/abs/2411.10761.

Gale, R., Dolata, J., Prud'hommeaux, E., Van Santen, J., & Asgari, M. (2020). Automatic assessment of language ability in children with and without typical development. In *2020 42nd Annual international conference of the IEEE engineering in medicine & biology society (EMBC)* (pp. 6111–6114). IEEE.

Gao, Z., Zhang, S., McLoughlin, I., & Yan, Z. (2022). Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *arXiv preprint arXiv:2206.08317*.

Gernsbacher, M. A., Morson, E. M., & Grace, E. J. (2016). Language and speech in autism. *Annual Review of Linguistics*, *2*(1), 413–425.

Geurts, H. M., & Embrechts, M. (2008). Language profiles in ASD, SLI, and ADHD. *Journal of Autism and Developmental Disorders*, *38*, 1931–1943.

Gilkerson, J., & Richards, J. A. (2008). The LENA natural language study. *Boulder, CO: LENA Foundation. Retrieved March*, *3*(2009), 15–17.

GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Rojas, D., Feng, G., Zhao, H., Lai, H., Yu, H., Wang, H., Sun, J., Zhang, J., Cheng, J., Gui, J., Tang, J., Zhang, J., Li, J.,…, Wang, Z. (2024). ChatGLM: A family of large language models from GLM-130b to GLM-4 all tools.

Gretter, R., Matassoni, M., Allgaier, K., Tchistiakova, S., & Falavigna, D. (2019). Automatic assessment of spoken language proficiency of non-native children. In *ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing (icassp)* (pp. 7435–7439). https://doi.org/10.1109/ICASSP.2019.8683268

Happé, F., & Frith, U. (2006). The weak coherence account: Detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *36*, 5–25.

I.,A., Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Yu, T., Xie, W., …, Dai, Z. (2024). Yi: Open foundation models by 01.AI. https://arxiv.org/abs/2403.04652.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Zhou, J., & Sun, H. (2024). Self-prompt tuning: Enable autonomous role-playing in llms. *arXiv preprint arXiv:2407.08995*.

Kraljevic, Z., Bean, D., Shek, A., Bendayan, R., Hemingway, H., Yeung, J. A., Deng, A., Baston, A., Ross, J., Idowu, E. et al. (2022). Foresight–generative pretrained transformer (GPT) for modelling of patient timelines using ehrs. *arXiv preprint arXiv:2212.08072*.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J. et al. (2023). Performance of chatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health*, *2*(2), e0000198.

Li, S., Guo, C., Lian, J., Cho, C. J., Zhao, W., Zhou, X., Zhou, D., Wang, S., Wang, G., Yang, J., Xu, J., Bao, R., Brenner, E., In, B., Pei, F., Gorno-Tempini, M. L., & Anumanchipalli, G. (2025). K-function: Joint pronunciation transcription and feedback for evaluating kids language function. https://arxiv.org/abs/2507.03043.

Lin, C. S., Chang, S. H., Liou, W. Y., & Tsai, Y. S. (2013). The development of a multimedia online language assessment tool for young children with autism. *Research in Developmental Disabilities*, *34*(10), 3553–3565. https://doi.org/10.1016/j.ridd.2013.06.042

Liu, X. (2019). Children language disorders and language assessment. *Chinese Scientific Journal of Hearing and Speech Rehabilitation*, *17*(3), 161–165.

Liu, X. (2019). An overview of child language disorders and assessments. *Chinese Scientific Journal of Hearing and Speech Rehabilitation*, *17*(03), 161–165.

MA, D.m., Zhi-mei, J., Wei, P., Hong, L. I., Lan-min, G., Xin, L. I., Mei-xuan, L. V., & Tian-yu, N. (2019). Exploration and analysis of language functions for autism spectrum disorders. *Chinese Scientific Journal of Hearing and Speech Rehabilitation*, *17*(03), 175–178.

Mullen, E. M. (1995). Mullen scales of early learning. AGS Circle Pines, MN.

Nadwodny, N., Yoder, P. J., Ingersoll, B. R., Wainer, A. L., Stone, W. L., Eisenhower, A., Carter, A. S., Network, R. R., Stone, W. L., Ingersoll, B. R. et al. (2025). The language ENvironment analysis (LENA) system in toddlers with early indicators of autism: Test–retest reliability and convergent validity with clinical language assessments. *Autism Research*. *18*(8), 1568–1579.

Nnamoko, N., Karaminis, T., Procter, J., Barrowclough, J., & Korkontzelos, I. (2024). Automatic language ability assessment method based on natural language processing. *Natural Language Processing Journal*, *8*, 100094. https://doi.org/10.1016/J.NLP.2024.100094

OpenAI (2024a). GPT-4o. https://openai.com/index/hello-gpt-4o/.

OpenAI (2024b). OpenAI o1. https://openai.com/index/introducing-openai-o1-preview/.

OpenAI (2025). OpenAI API documentation. https://platform.openai.com/docs/guides/text?api-mode=responses.

Ouyang, Y., Gao, W., Wang, H., Chen, L., Wang, J., & Zeng, Y. (2024). Medsq: Towards personalized medical education via multi-form interaction guidance. *Expert Systems with Applications*, *267*, 126138.

Rescorla, L. (1989). The language development survey: A screening tool for delayed language in toddlers. *Journal of Speech and Hearing Disorders*, *54*(4), 587–599.

Richards, J. A., Xu, D., Gilkerson, J., Yapanel, U., Gray, S., & Paul, T. (2017). Automated assessment of child vocalization development using LENA. *Journal of Speech, Language, and Hearing Research*, *60*(7), 2047–2063.

Rose, V., Trembath, D., Keen, D., & Paynter, J. (2016). The proportion of minimally verbal children with autism spectrum disorder in a community-based early intervention programme. *Journal of Intellectual Disability Research*, *60*(5), 464–477.

Rutter, M., Le Couteur, A., & Lord, C. (2003). Autism diagnostic interview-revised. *Los Angeles, CA: Western Psychological Services*, *29*(2003), 30.

Schopler, E., Reichler, R. J., & Renner, B. R. (2010). The childhood autism rating scale (CARS). Western Psychological Services Los Angeles, CA.

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498.

Shao, Y., Li, L., Dai, J., & Qiu, X. (2023). Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Shield, A., Cooley, F., & Meier, R. P. (2017). Sign language echolalia in deaf children with autism spectrum disorder. *Journal of Speech, Language, and Hearing Research*, *60*(6), 1622–1634.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D. et al. (2023). Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Su, Y. E., & Naigles, L. (2020). Expressive language abilities of mandarin-exposed preschool children with autism spectrum disorder (ASD). *Chinese Journal of Language Policy and Planning*, *5*(02), 25–34. https://doi.org/10.19689/j.cnki.cn10-1361/h.20200202

Tardif, T., & Fletcher, P. (2008). Chinese communicative development inventories: User's guide and manual.

Tseng, Y. M., Huang, Y. C., Hsiao, T. Y., Hsu, Y. C., Foo, J. Y., Huang, C. W., & Chen, Y. N. (2024). Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Tu, Q., Fan, S., Tian, Z., & Yan, R. (2024). Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Wang, B., Su, X., Lv, L., & Zhang, R. (2017). Analysis of language development level in children with autism. *Chinese Journal of Child Health Care*.

Wang, Y., Williams, R., Dilley, L., & Houston, D. M. (2020). A meta-analysis of the predictability of LENA™ automated measures for child language development. *Developmental Review*, *57*, 100921.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

wenqian, L. (2018). Factors associated with age at firstdiagnosis among children with autismspectrum disorders of a tertiary hospital. Ph.D. thesis. Chongqing: Chongqing Medical University.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatGPT. https://arxiv.org/abs/2302.11382.

Wiig, E. H., Secord, W. A., & Semel, E. (2013). Clinical evaluation of language fundamentals: CELF-5. Pearson.

Wu, W., Wu, H., Jiang, L., Liu, X., Hong, J., Zhao, H., & Zhang, M. (2024). From role-play to drama-interaction: An LLM solution. *arXiv preprint arXiv:2405.14231*.

Xue, V. W., Lei, P., & Cho, W. C. (2023). The potential impact of chatGPT in clinical and translational medicine. *Clinical and Translational Medicine*, *13*(3), e1216.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J.,…, Fan, Z. (2024). Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Flores, M. G., Zhang, Y. et al. (2022). Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*.

Yang, Y., Shen, Y., Sun, T., & Xie, Y. (2025). Validating the effectiveness of a large language model-based approach for identifying children's development across various free play settings in kindergarten. https://arxiv.org/abs/2505.03369.

Yeo, Y. H., Samaan, J. S., Ng, W. H., Ting, P. S., Trivedi, H., Vipani, A., Ayoub, W., Yang, J. D., Liran, O., Spiegel, B. et al. (2023). Assessing the performance of chatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clinical and Molecular Hepatology*, *29*(3), 721.

Yi, S. U., Qian-qian, X., & Lin-yan, S. U. (2020). Cross-population comparison of early expressive language profiles in ASD, DD and LD. *Chinese Journal of Clinical Psychology*, *28*(03), 508–512 + 517. https://doi.org/10.16128/j.cnki.1005-3611.2020.03.016

Zhao, J., Tang, L., He, T., Xiong, T., Yan, T., Li, J., Wu, D., & Yan, H. (2021). Characteristics of language development in children with autism spectrum disorders. *Chinese Journal of Child Health Care*, *29*(09), 969–972.

Zhou, M., Niu, J., & Lu, H. (2023). Analysis of picture book reading comprehension characteristics in autistic children with language age of 4-6 years. *Modern Special Education*, (02), 68–72.