

Article

A Multi-Agent Consensus Framework for Explainable Automated Assessment of Language Ability in Children with Autism Spectrum Disorder

Saige Qin ¹, Tongxin Yin ¹, Min Liu ¹ and Qiaoyun Liu ^{1,2,*}

¹ Shanghai Institute of Artificial Intelligence for Education, East China Normal University, Shanghai 200062, China; saige_qin@stu.ecnu.edu.cn (S.Q.); txyincuzn@stu.ecnu.edu.cn (T.Y.); liumin@mail.ecnu.edu.cn (M.L.)

² Innovation Lab of Education and Rehabilitation, Faculty of Education, East China Normal University, Shanghai 200062, China

* Correspondence: qyliu@spe.ecnu.edu.cn

Abstract

Language ability reflects the developmental level and social communication in children with autism spectrum disorder (ASD), making scalable and explainable assessment essential. Existing large language model (LLM)-based methods mainly rely on single-agent reasoning or aggregate independent outputs, offering limited insight into disagreement identification, evidence exchange, and consensus formation. To address this limitation, this paper proposes a Multi-Agent Consensus framework that formulates language assessment as a structured consensus process among LLM-based agents. Using parent–child free-play transcripts, three Expert Agents independently assess language comprehension, vocabulary competence, logical expression, and communication skills under shared criteria. A Judge Agent identifies dimension-level disagreements, marks dimensions with consistent judgments as resolved, coordinates evidence-based negotiation, and adjudicates remaining conflicts. On a real-world dataset of 79 children, the framework achieves a Pearson correlation of 0.843 with an external standardized language development measure, outperforming Single Agent (0.802) and Initial Expert Aggregation (0.814), while maintaining strong cross-run stability. Ablation and process analyses show that consensus negotiation integrates initial judgments, resolves disagreements, and improves assessment stability. Case analysis and human expert evaluations support the plausibility, evidential grounding, and transparency of the generated reports. These findings suggest that multi-agent consensus can enhance automated language assessment for children with ASD while providing structured, traceable explanations. The correlations support criterion validity, but clinical validity and effectiveness in practice require further evaluation given the limited sample size and data sources.



Academic Editor: Flavio Canavero

Received: 3 September 2026

Revised: 23 September 2026

Accepted: 5 October 2026

Published: 9 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: autism spectrum disorder; child language assessment; large language models; multi-agent consensus; explainable assessment

1. Introduction

Child language ability is an important indicator of neurodevelopmental level and social communication competence. It also plays a critical role in the early identification of autism spectrum disorder (ASD), as well as in functional assessment and intervention planning for children with ASD. Previous studies have shown that language ability is closely

associated with social interaction, communicative behavior, and overall developmental level in children with ASD [1–4]. Compared with structured testing environments, natural parent–child interaction can better capture children’s language performance in vocabulary use, sentence organization, semantic understanding, and communicative turn-taking. Therefore, automated, stable, and explainable language ability assessment based on natural parent–child dialogue has become an important research direction in child developmental assessment and clinical support for children with ASD [5,6].

Although standardized scales and clinical judgments provide professionally grounded assessment, they require trained assessors and are difficult to scale for continuous monitoring [7–10]. Existing automated methods improve scalability, but many rely on manually designed features or task-specific formats and still have limited ability to model contextual relations, pragmatic functions, and turn-taking behavior in natural parent–child interaction [11–14]. These limitations make it difficult to support comprehensive and explainable language ability judgments from naturalistic dialogue.

In recent years, large language models (LLMs) have provided a new approach to child language assessment, owing to their strong natural language understanding, contextual integration, and reasoning capabilities. Recent studies have used LLMs to analyze child–adult interactions, predict stages of language development, and assess children’s developmental abilities in free-play settings [6,15]. Qin et al. further demonstrated that LLMs can perform multidimensional language assessment of children with ASD using parent–child dialogue, achieving strong correlations with standardized developmental measures [5]. These studies indicate that LLMs can model natural parent–child dialogue and support multidimensional language ability assessment.

However, existing LLM-based methods for child language ability assessment mainly focus on single-pass assessment results. Some methods improve stability by aggregating multiple independent assessment outputs, but they pay limited attention to key elements of expert judgment, such as disagreement identification, evidence exchange, and opinion revision. A key remaining limitation concerns not only whether LLMs can output accurate scores but also whether the assessment process can make disagreements explicit, support evidence-based revision, and make the final decision path reviewable. Formal expert consensus methods provide procedural references for this purpose, including independent judgments, structured feedback, and opportunities for judgment revision [16–19]. Therefore, there is a need for methods that support evidence-based negotiation among LLM agents while providing transparent, traceable decision rationales.

Based on these observations, this paper proposes a Multi-Agent Consensus framework for child language ability assessment. The framework uses three Expert Agents to generate independent multidimensional assessments from the same parent–child dialogue evidence. A Judge Agent then identifies dimension-level disagreements, coordinates evidence-based negotiation, and adjudicates dimensions that remain unresolved after negotiation. Unlike traditional result aggregation methods, the proposed framework outputs not only final assessment levels but also expert opinion changes, negotiation trajectories, and decision rationales. This makes the assessment process traceable and reviewable.

The main contributions of this study are summarized as follows:

1. We propose a Multi-Agent Consensus framework for language ability assessment in children with ASD using parent–child dialogue as evidence, extending the task from single-pass scoring to traceable consensus-based assessment.
2. We design a dimension-level consensus negotiation mechanism for child language ability assessment, which records the decision path from initial judgments to final conclusions through disagreement detection, evidence-driven negotiation, dimension resolution, and final adjudication.

3. We evaluate Multi-Agent Consensus on a real-world parent–child dialogue dataset collected from 79 children through controlled comparisons, ablation experiments, process analysis, and human expert evaluation, demonstrating its effectiveness and cross-run stability and providing human evaluation evidence of the explanatory quality of its reports.

2. Related Work

2.1. Child Language Ability Assessment

Child language ability assessment has long relied on standardized scales, parent reports, and professional clinical judgments. For example, tools such as the Mullen Scales of Early Learning (MSEL), the Clinical Evaluation of Language Fundamentals (CELF), the Peabody Picture Vocabulary Test (PPVT), and the Chinese Communicative Development Inventory have been widely used to assess children’s language and developmental abilities [7–10]. These methods are professionally grounded in the design of assessment dimensions, standardized procedures, and norm-based interpretation. However, they usually require professionals to conduct face-to-face testing. This leads to relatively high administration costs and limits their scalability for large-scale screening and long-term continuous monitoring.

With advances in speech processing, natural language processing, and machine learning, researchers have begun to explore automated assessment methods based on children’s language samples [20]. Some studies use acoustic features in children’s speech to identify ASD-related language characteristics or predict language development indicators [11–13]. Others extract lexical, syntactic, and semantic features from children’s language samples and combine them with machine learning models to characterize language ability [14,21]. In addition, some studies directly predict language development levels from children’s natural language samples [22–24]. These methods have advanced automated child language assessment. However, they often depend on manually designed features or specific task formats and do not adequately model contextual relations, semantic continuity, turn-taking organization, or communicative functions in natural parent–child interaction.

2.2. LLM-Based Language Ability Assessment

In recent years, LLMs have been increasingly applied to automated language evaluation and educational assessment because of their strong semantic understanding and contextual modeling capabilities. More broadly, studies on LLM-as-a-judge have shown that LLMs can evaluate text quality according to evaluation criteria, structured templates, or reasoning prompts. In some tasks, these models can achieve high agreement with human judgments [25]. In automated essay scoring and second-language writing assessment, recent studies have used LLMs for scoring, feedback generation, and multidimensional ability assessment [26–28].

In the context of child development and interaction assessment, Feng et al. explored the use of general-purpose LLMs to analyze clinical interactions between children and adults. Their results show that LLMs can help identify ASD-related behavioral characteristics in children [6]. Yang et al. further validated the feasibility of using LLMs to assess children’s multidimensional developmental abilities in free-play settings [15]. Qin et al. developed a child language ability assessment method based on parent–child dialogue. Their method evaluates language comprehension, vocabulary competence, logical expression, and communication skills and shows strong correlations between LLM-based assessment results and standardized developmental measures [5].

However, existing LLM-based studies on child language ability assessment mainly focus on the agreement between assessment results and external standardized measures.

That is, they primarily examine whether models can accurately output scores or levels. Although some LLM-as-a-judge and educational assessment studies can generate evaluation rationales, feedback, or structured reports [25–30], these explanations are usually provided as post-hoc rationales for single-pass judgments. Such explanations rarely reveal disagreements or opinion changes among different evaluators. When multiple models, prompts, or repeated assessments are used to improve stability, existing methods usually adopt result aggregation strategies. These include answer selection, score averaging, or consistency checking [31,32]. They do not sufficiently model evidence exchange or judgment revision. Therefore, moving from independent scoring to a traceable consensus formation process remains an underexplored problem in LLM-assisted child language ability assessment.

2.3. Multi-Agent Consensus for Explainable Assessment

Multi-agent collaborative evaluation offers a way to make differing judgments and their revision processes explicit. ChatEval organizes evaluators with different role descriptions to discuss text quality and examines how communication strategies affect assessment [33]. CollabEval combines independent assessment, multi-round discussion, and final judgment, allowing agents to exchange evaluation results and justifications and using consensus checks for early termination [34]. These frameworks extend evaluation beyond isolated judgments by incorporating exchanges among evaluators into the assessment process.

In collaborative evaluation, integrating different judgments, organizing exchanges, and reaching final conclusions are important aspects of process design. Related studies offer different methodological references: Mixture-of-Agents (MoA) synthesizes candidate responses through a layered architecture [35], whereas multiagent debate enables model instances to revise their answers after considering others' responses and reasoning [36]. For decision-making following discussion, Kaesberg et al. compare voting and consensus protocols and find differences in their relative effectiveness across the knowledge and reasoning tasks examined [37]. In addition, the RAND/UCLA Appropriateness Method combines independent ratings with feedback, disagreement discussion, and re-rating [18], providing a reference from human expert procedures for organizing the exchange and revision of assessment judgments.

These studies provide a methodological foundation for integrating different judgments and documenting the assessment process. In child language ability assessment based on natural parent–child dialogue, judgments need to be grounded in the conversational context and linguistic evidence. Beyond presenting final ratings and their supporting evidence, documenting disagreements among evaluators and subsequent judgment revisions can also support the interpretation and review of assessment conclusions. To address this need, this study combines multi-agent negotiation with linguistic evidence analysis to generate reports that make the formation of assessment conclusions traceable.

3. Method

As shown in Figure 1, the framework consists of four stages: input preparation, independent expert assessment, consensus negotiation, and final adjudication with explainable report generation. The system first constructs a unified prompt containing child information, dialogue evidence, and assessment criteria. Three Expert Agents then independently generate dimension-level assessments. The Judge Agent subsequently checks agreement, organizes negotiation for unresolved dimensions, and adjudicates any remaining conflicts. The final output is an explainable report that records final levels, supporting evidence, decision sources, and the consensus trajectory.

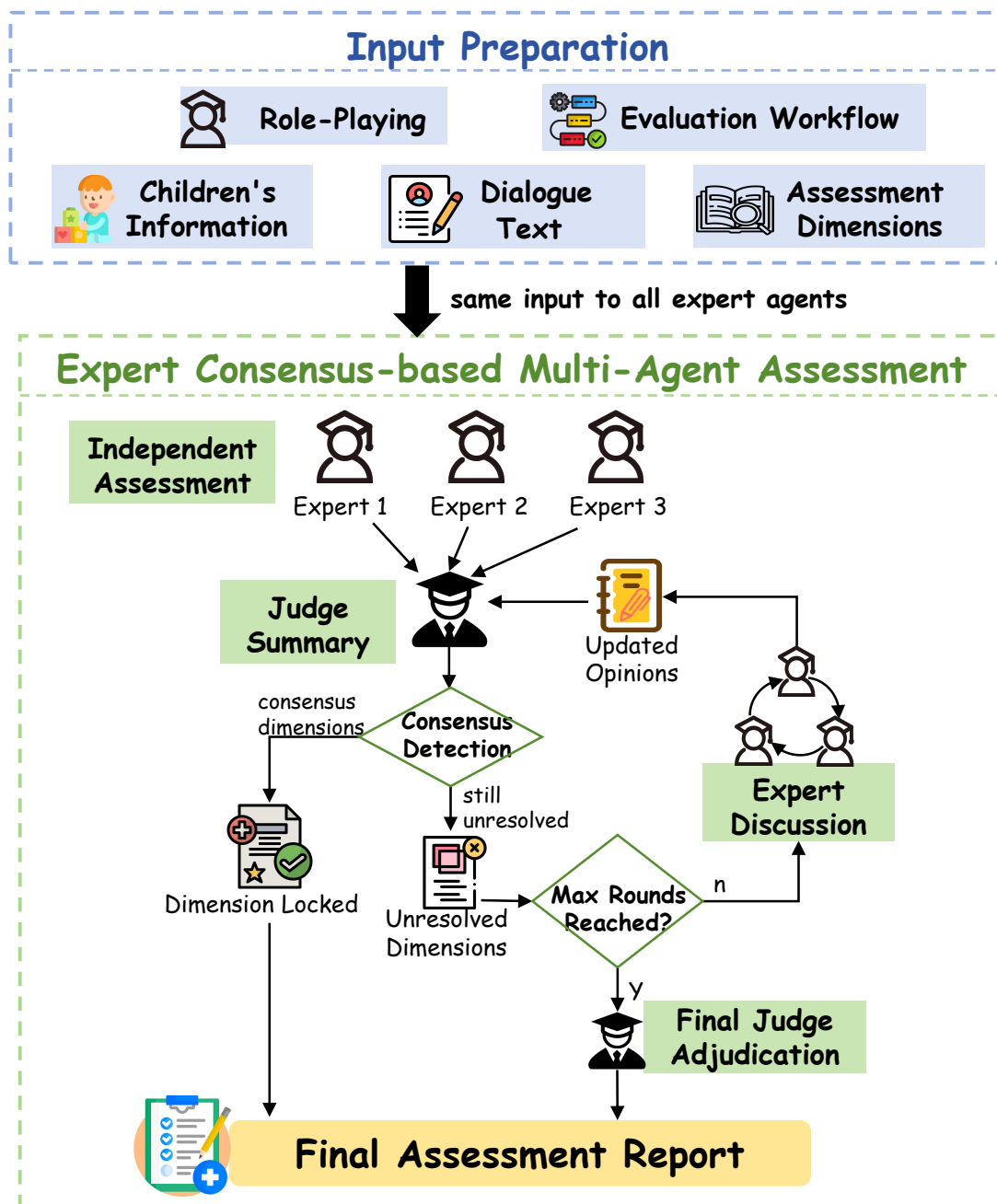


Figure 1. Overall framework of the proposed consensus-based multi-agent assessment method.

3.1. Input Preparation

The input preparation stage organizes the information required for child language ability assessment into a unified input for the Expert Agents. The goal is to provide all Expert Agents with the same task context, evidence scope, and rating scale. This ensures that their independent judgments are based on the same evidence base. Specifically, the input includes role specification, assessment dimensions, child information, parent–child free-play dialogue transcripts, and the assessment workflow.

For role specification, the system assigns each agent the role of an expert in child language assessment. Each agent is instructed to make careful judgments based on the child's developmental background and natural interaction evidence. The child information includes age and sex, which provide developmentally appropriate reference information for expert judgment. MSEL scores and diagnostic labels were not included in the inputs to either the Expert Agents or the Judge Agent. The raw evidence consists of dialogue

audio collected during parent–child free-play sessions, while the evidence provided to the Expert Agents is the manually corrected, anonymized, and speaker-labeled transcript. Specifically, we first used Paraformer-large (https://www.modelscope.cn/models/iic/speech_paraformer-large-vad-punc_asr_nat-zh-cn-16k-common-vocab8404-pytorch; accessed on 4 October 2026) for automatic speech transcription and applied the ERes2Net speaker identification model (https://www.modelscope.cn/models/iic/speech_eres2net_sv_zh-cn_16k-common; accessed on 4 October 2026) to distinguish child and parent speech segments [38,39]. The automatic transcripts were then manually corrected, anonymized, and formatted. This process produced parent–child dialogue texts with speaker labels and timestamps, ensuring that the evidence used for subsequent language ability assessment is reliable and traceable.

Following existing studies on child language ability assessment [5], this paper divides language ability into four dimensions: language comprehension, vocabulary competence, logical expression, and communication skills. Language comprehension refers to the understanding of literal and contextual meanings. Vocabulary competence reflects vocabulary richness and contextual appropriateness. Logical expression captures language organization and expressive coherence. Communication skills refer to initiation, response, and turn maintenance. Each dimension is evaluated using a five-level rating scale. The level set is defined as $\mathcal{L} = \{\text{very poor, poor, fair, good, excellent}\}$.

In the assessment workflow, each expert first reviews the four assessment dimensions and the scope of each dimension. The expert then uses the child’s age and sex as developmental reference information, analyzes the child’s performance in each dimension based on language behavior evidence from the parent–child dialogue, and finally outputs ratings for the four dimensions. The resulting structured prompt is provided identically to all three Expert Agents, ensuring that their initial judgments are made under the same task context, assessment criteria, and dialogue evidence. These independent judgments serve as the starting point for subsequent disagreement detection and consensus negotiation, as illustrated in the input preparation module of Figure 1. Complete prompt templates, input fields, and output specifications for the Expert and Judge Agents are provided in Supplementary File S1.

3.2. Independent Expert Assessment

The independent assessment stage is conducted in parallel by three Expert Agents. For the i -th child sample, the input preparation stage generates a unified initial assessment prompt p_i , which contains the child’s basic information, parent–child dialogue text, four assessment dimensions, and the assessment workflow. The system sends the same prompt p_i to the three Expert Agents. All three Expert Agents use the same base large language model. Under identical assessment inputs, evidence scopes, and rating criteria, they form initial judgments through mutually independent generation calls. At this stage, each Expert Agent cannot access the outputs of the other Agents. Here, Expert Agent denotes an assigned assessment role rather than distinct professional expertise; the initial assessments are separate stochastic generations from the shared model.

Formally, let the expert set be $K = \{1, 2, 3\}$, and let the language ability dimension set be $D = \{d_1, d_2, d_3, d_4\}$, corresponding to language comprehension, vocabulary competence, logical expression, and communication skills, respectively. The initial free-text assessment generated by the k -th Expert Agent for sample i is denoted as $z_{i,k} = M_\theta(p_i; \epsilon_k)$, where M_θ denotes the shared-parameter language model, and ϵ_k denotes the independent generation state associated with the k -th Expert Agent. The three Agents receive exactly the same assessment input p_i , but their generation processes are independent. Therefore, their outputs may differ and serve as the initial state for subsequent disagreement detection and

consensus negotiation. The experimental model and generation parameters are described in Section 4.1.2.

To support subsequent disagreement detection and negotiation, the system extracts the rating level and rationale for each dimension from the free-text assessments of the three experts. These results are organized into structured dimension-level positions. For each expert k and dimension $d \in D$, the structured initial position is represented as

$$a_{i,k,d}^0 = (l_{i,k,d}^0, r_{i,k,d}^0), \quad l_{i,k,d}^0 \in \mathcal{L},$$

where $l_{i,k,d}^0$ is the initial level assigned by expert k for dimension d , and $r_{i,k,d}^0$ is the evidence-based rationale for that level. The structured initial assessment set of the three experts for sample i is defined as $A_i^0 = \{a_{i,k,d}^0 \mid k \in K, d \in D\}$. This structured representation enables the system to compare the initial judgments of the three experts at the dimension level. It also provides the input for subsequent dimension-level agreement checking, disagreement detection, and consensus negotiation.

The purpose of the independent assessment stage is not to directly merge the three expert ratings. Instead, it provides multiple independent initial viewpoints for subsequent consensus negotiation. Because the three experts generate judgments under the same evidence and rating criteria, the system can preserve initial disagreement signals caused by insufficient evidence, ambiguous rating boundaries, or different interpretations of evidence. These signals provide the basis for subsequent dimension-level agreement checking, disagreement detection, and expert negotiation.

3.3. Consensus Negotiation

After the independent initial assessments, the Judge Agent compares the three structured outputs at the dimension level. Dimensions with identical ratings are marked as resolved, with their assessment level, supporting evidence, and resolution status recorded. Dimensions with divergent ratings are marked as unresolved. For these unresolved dimensions, the Judge Agent summarizes the main sources of disagreement and the issues to be discussed in subsequent negotiation. Dimensions that have already reached agreement are excluded from further discussion, which avoids unnecessary interaction and preserves the dimension-by-dimension consensus formation process.

Formally, let $t = 0$ denote the initial assessment summary stage, and let $t \geq 1$ denote the state after the t -th round of expert negotiation. For dimension d of sample i , the current levels assigned by the three experts after round t are denoted as $l_{i,1,d}^t$, $l_{i,2,d}^t$, and $l_{i,3,d}^t$. The dimension-level consensus indicator is defined as

$$C_{i,d}^t = \begin{cases} 1, & l_{i,1,d}^t = l_{i,2,d}^t = l_{i,3,d}^t, \\ 0, & \text{otherwise.} \end{cases}$$

Let L_i^t denote the set of resolved dimensions after round t , and let U_i^t denote the set of unresolved dimensions that still require negotiation. After the initial assessment summary, the system initializes L_i^0 and U_i^0 according to $C_{i,d}^0$. After negotiation round t , the dimension status is updated as follows:

$$L_i^t = L_i^{t-1} \cup \{d \in U_i^{t-1} \mid C_{i,d}^t = 1\},$$

$$U_i^t = D \setminus L_i^t.$$

In negotiation round t , the Expert Agents discuss only the dimensions in U_i^{t-1} . Each expert can access the initial opinions of all three experts, the resolved dimensions, the

current unresolved dimensions, the previous Judge Agent summary, the negotiation history, and the original dialogue evidence.

For each unresolved dimension, the expert outputs an updated level, supporting evidence, responses to other Agents' opinions, and an indication of whether the initial judgment is maintained or revised. The negotiation process does not determine the final level by majority voting. Instead, the final state reflects consensus formed through evidence exchange and judgment revision. Algorithm 1 presents the pseudocode of the dimension-level consensus negotiation process.

Algorithm 1 Dimension-level consensus negotiation

Require: Structured initial assessments A_i^0 , dimension set D , maximum rounds T
Ensure: Resolved dimensions L_i , unresolved dimensions U_i , negotiation history H_i

```

1:  $L_i \leftarrow \emptyset, U_i \leftarrow D, H_i \leftarrow \emptyset$ 
2: Judge Agent summarizes  $A_i^0$  and obtains  $\{l_{i,k,d}^0 \mid k \in K, d \in D\}$ 
3: for all  $d \in D$  do
4:   if  $l_{i,1,d}^0 = l_{i,2,d}^0 = l_{i,3,d}^0$  then
5:      $L_i \leftarrow L_i \cup \{d\}$ 
6:   end if
7: end for
8:  $U_i \leftarrow D \setminus L_i$ 
9: for  $t = 1$  to  $T$  do
10:  if  $U_i = \emptyset$  then
11:    break
12:  end if
13:  for all Expert Agent  $k \in K$  do
14:    Generate a negotiation response only for dimensions in  $U_i$ 
15:  end for
16:  Update current expert positions  $\{l_{i,k,d}^t \mid k \in K, d \in U_i\}$  based on negotiation responses
17:  Judge Agent summarizes current expert positions and appends the round summary to  $H_i$ 
18:   $N_i \leftarrow \emptyset$ 
19:  for all  $d \in U_i$  do
20:    if  $l_{i,1,d}^t = l_{i,2,d}^t = l_{i,3,d}^t$  then
21:       $N_i \leftarrow N_i \cup \{d\}$ 
22:    end if
23:  end for
24:   $L_i \leftarrow L_i \cup N_i$ 
25:   $U_i \leftarrow D \setminus L_i$ 
26: end for
27: return  $L_i, U_i, H_i$ 

```

After each negotiation round, the Judge Agent produces a structured summary of the latest opinions of the three Expert Agents. It updates the current level, supporting rationale, and negotiation status for each dimension. The system then detects newly resolved dimensions according to the strict level-agreement rule and removes these dimensions from subsequent negotiation. In the algorithm, H_i denotes the negotiation history. It records the Judge Agent summary for each round, changes in expert positions, and dimension status updates. Dimensions that remain unresolved are kept in U_i^t and enter the next negotiation round. In this paper, the maximum number of negotiation rounds is set to $T = 5$. When $U_i^t = \emptyset$, the system ends negotiation and directly proceeds to final result determination and report generation. When $U_i^T \neq \emptyset$ after the maximum number of negotiation rounds, unresolved dimensions are passed to the final adjudication process described below. Through this dimension-level status update mechanism, the system obtains final assessment levels and records the complete process through which different dimensions move from initial disagreement to opinion revision and finally to consensus or adjudication. This provides a traceable consensus formation trajectory for the assessment results.

3.4. Final Adjudication and Explainable Report

When all dimensions have been resolved, or when unresolved dimensions remain after the maximum number of negotiation rounds, the system enters the final result determination and report generation stage. Each final dimension-level result has one of two decision sources. For dimensions that reach level agreement during negotiation, the system retains the agreed level assigned by the three Expert Agents, the supporting rationale, and the corresponding resolution round. For dimensions that still do not reach agreement after the maximum number of negotiation rounds, the Judge Agent performs final adjudication based on the original dialogue evidence, initial assessment results, latest negotiation status, and complete negotiation history.

Formally, let the final level and final rationale for dimension d of sample i be $\hat{l}_{i,d}$ and $\hat{r}_{i,d}$, respectively. If dimension d is resolved in round $\tau_{i,d}$ through the level-agreement check among the three Expert Agents, its final level satisfies

$$\hat{l}_{i,d} = l_{i,1,d}^{\tau_{i,d}} = l_{i,2,d}^{\tau_{i,d}} = l_{i,3,d}^{\tau_{i,d}}.$$

If dimension d still fails to reach agreement after the maximum number of negotiation rounds, the Judge Agent generates the final adjudication based on the three experts' initial opinions, latest positions, negotiation history, and original dialogue evidence:

$$(\hat{l}_{i,d}, \hat{r}_{i,d}) = J(d, A_i^0, H_i, L_i^T, p_i), \quad d \in U_i^T,$$

where $J(\cdot)$ denotes the final adjudication process of the Judge Agent. Its inputs include the dimension d , the structured initial assessments A_i^0 of the three Expert Agents, the negotiation history H_i containing the latest expert positions and Judge Agent summaries from previous rounds, the set of resolved dimensions L_i^T , and the original assessment input p_i . By integrating this information, the Judge Agent adjudicates unresolved dimensions and generates final levels and explanatory rationales.

After obtaining the final levels for the four dimensions, the system converts the levels into dimension scores according to the scoring design adopted in prior child language ability assessment research [5]. This scoring design assigns equal weight to the four language ability dimensions. Each dimension has a maximum score of 37.5, and the total score has a maximum of 150. The five levels are mapped to dimension scores at equal intervals of 7.5 points. Let $g(\cdot)$ be the mapping function from level to score. The mapping is defined as

$$\begin{aligned} g(\text{very poor}) &= 7.5, & g(\text{poor}) &= 15.0, & g(\text{fair}) &= 22.5, \\ g(\text{good}) &= 30.0, & g(\text{excellent}) &= 37.5. \end{aligned}$$

This linear mapping is used only to convert discrete levels into continuous scores for correlation analysis with external standardized measures. It is not involved in model reasoning or negotiation. Therefore, the score for sample i on dimension d and the total score are defined as

$$s_{i,d} = g(\hat{l}_{i,d}), \quad S_i = \sum_{d \in D} s_{i,d}.$$

The structure of the final explainable language ability assessment report is shown in Figure 2. The report includes anonymized child information, a description of the assessment input, final levels, dimension scores, and the overall score. For each dimension, it records the final rationale, decision source (Expert Consensus or Final Judge Adjudication), and resolution stage. The process trace further summarizes initially resolved dimensions, opinion revisions during negotiation, and final adjudication results. Compared with traditional assessment outputs that mainly present final results, this report provides an

explanation chain from assessment result to supporting evidence, opinion revision, and final decision, making the decision process traceable and analyzable.

This report presents the final explainable language assessment generated from an anonymized parent-child free-play dialogue. It summarizes the child's language ability across four dimensions and records the final scores, expert rationales, decision sources, and consensus-negotiation trace produced by the multi-agent framework.	
## Child Profile	
- Case ID: <anonymous_case_id>	- Age / Gender: <age> / <gender>
- Evidence: anonymized parent-child free-play dialogue with speaker labels and utterance timestamps	
## Overall Language Assessment	
- Total score: <total_score> / 150.0	
- Dimension scores: LC <score>; VC <score>; LE <score>; CS <score>	
## Dimension-Level Final Assessment	
### Language Comprehension	
- Level: <level>	- Score: <score>
- Decision source: <expert consensus final judge adjudication>	
- Resolution point: <round_id final_judge>	
- Final rationale: <final_rationale>	
### Vocabulary Competence ...	
### Logical Expression ...	
### Communication Skills ...	
## Decision Explanation	
- Representative dimension: <dimension_name>	
- Expert positions: <expert_1 level+rationale> / <expert_2 level+rationale> / <expert_3 level+rationale>	
- Final rationale: <final_rationale>	
- Decision source: <decision_source>	
## Process Trace	
- Round 0: <initial_locked_dimensions none> locked; <initial_unresolved_dimensions none> unresolved	
- Round 1-T: <negotiation_updates none>	
- Final judge: <judge_resolved_dimensions none>	

Figure 2. Template of the final explainable language assessment report. Placeholders in angle brackets are filled with case-specific assessment outputs. LC, VC, LE, and CS denote Language Comprehension, Vocabulary Competence, Logical Expression, and Communication Skills, respectively. Blue, green, and yellow shading distinguishes the overall assessment, decision explanation, and process trace sections, respectively. In the decision-source field, yellow and pink highlighting indicates expert consensus and final judge adjudication, respectively.

4. Experiments

4.1. Experimental Setup and Implementation Details

4.1.1. Experimental Data

The dataset comprised parent-child free-play dialogue data from 79 children whose primary language of communication was Mandarin Chinese. The children were 2–6 years old and included 29 typically developing (TD) children, 21 children with developmental delay (DD), and 29 children with autism spectrum disorder (ASD). The ASD and DD diagnoses were made by developmental-behavioral pediatricians. All children were assessed using the MSEL [7]. We used the MSEL Verbal Developmental Quotient (VDQ) as the external criterion for examining the criterion validity of the automated language ability assessment scores. Table 1 summarizes the demographic characteristics and MSEL VDQ scores. The study was approved by the institutional ethics committee, and written informed consent was obtained from the guardians of all children before data collection.

Table 1. Demographic characteristics and MSEL VDQ scores of the participants.

Group	Age (Months)	Sex (Male/Female)	MSEL VDQ
TD (<i>n</i> = 29)	42.12 ± 15.18	15/14	108.12 ± 12.80
DD (<i>n</i> = 21)	60.43 ± 11.70	15/6	50.96 ± 23.14
ASD (<i>n</i> = 29)	62.28 ± 9.20	23/6	35.31 ± 21.34

Note: TD = typically developing; DD = developmental delay; ASD = autism spectrum disorder; MSEL VDQ = Mullen Scales of Early Learning Verbal Developmental Quotient; SD = standard deviation; *n* = number of participants. Age and MSEL VDQ are reported as mean ± SD; sex is reported as male/female counts.

To obtain naturalistic child-language interaction data, we recorded parent–child free-play interactions in a laboratory setting. Each interaction lasted approximately 15 min, yielding about 1185 min of raw audio in total. The recordings were automatically transcribed, manually corrected, anonymized, and converted into speaker-labeled dialogue transcripts for subsequent analysis. To reduce the influence of initial adaptation and end-of-session fatigue, the analyses used the interaction content between minutes 3 and 13 of each recording. Dialogue segments crossing these boundaries were adjusted when necessary to preserve complete conversational turns.

4.1.2. Overall Experimental Protocol and Model Configuration

This study evaluates Multi-Agent Consensus from three perspectives: assessment validity, consensus-configuration comparison, and framework explainability. Assessment validity is measured by the correlation between automated language ability assessment scores and the external MSEL VDQ criterion. The consensus-configuration comparison examines how the full framework performs relative to initial expert aggregation and ablated configurations in which specific consensus mechanisms are removed or altered. Framework explainability is evaluated using a representative case study and human expert evaluation, with emphasis on the reliability of generated explanations for both final assessment results and the consensus-formation process.

All language ability assessment experiments were conducted with a locally deployed Qwen3.6-35B-A3B model (developed by Alibaba, Hangzhou, China) [40]. The model was served by vLLM (version 0.18.0) through an OpenAI-compatible application programming interface (API) on a server equipped with four NVIDIA L20 graphics processing units (GPUs) (NVIDIA Corporation, Santa Clara, CA, USA). The inference service used tensor parallelism with a tensor-parallel size of 4, a maximum context length of 65,536 tokens, and prefix caching. The assessment workflow used instruct (non-thinking) mode; thinking was disabled by setting `enable_thinking=false` in the chat-template configuration. Following the recommended Qwen3.6 sampling settings for instruct mode, temperature, top-p, top-k, min-p, presence penalty, and repetition penalty were set to 0.7, 0.8, 20, 0.0, 1.5, and 1.0, respectively. The maximum output length was set to 8,192 tokens to support the structured outputs required for assessment, negotiation summaries, and final adjudication.

4.2. Comparative Results on Assessment Validity

4.2.1. Validity Evaluation Protocol

This section evaluates the assessment validity of the proposed Multi-Agent Consensus framework through two comparisons. First, Multi-Agent Consensus is compared with Single Agent under the same dataset, base model, and generation parameters to isolate the effect of the consensus mechanism. Second, the proposed method is compared with previously reported LLM-based results from Qin et al. [5], including GPT-4o, Doubao-Pro-32k, and Qwen2-72B-Instruct. These reference results provide a benchmark for evaluating the proposed method relative to existing LLM-based automated assessment methods.

The Pearson correlation coefficient r was used to measure the linear correlation between the automated language ability total score and the MSEL Vdq external criterion in the pooled sample of 79 children. To reduce the influence of randomness from a single generation, both Single Agent and Multi-Agent Consensus were independently run five times. The mean Pearson correlation coefficient across the five runs is reported. The comparative results are shown in Table 2.

To examine criterion validity within the ASD group, we additionally analyzed all 29 children with ASD using the existing outputs from Single Agent, Initial Expert Aggregation, and Multi-Agent Consensus. Initial Expert Aggregation is the score-averaging baseline described in Section 4.3. For each method, Pearson and Spearman correlations with MSEL Vdq were calculated separately for each of the five evaluations and summarized as means and sample standard deviations.

Table 2. Pearson correlations for large language model (LLM)-based language assessment, including controlled results from this work and reference results from prior work.

Source	Method	Pearson r
Qin et al. [5]	GPT-4o	0.80
	Doubao-Pro-32k	0.73
	Qwen2-72B-Instruct	0.74
This work	Single Agent (Qwen3.6-35B-A3B)	0.80
	Multi-Agent Consensus (Qwen3.6-35B-A3B)	0.84

Note: The results for Qin et al. are reference values reported in the original study. For this work, Pearson r denotes the mean over five independent runs using the same dataset and model configuration. Bold indicates the highest Pearson r among the values shown.

4.2.2. Comparative Results

In the controlled comparison, Single Agent achieved an average Pearson r of 0.80, whereas Multi-Agent Consensus achieved 0.84, corresponding to an absolute improvement of 0.04. Because both methods used the same dataset, base model, and generation parameters, their main difference was the assessment workflow. Single Agent directly completed the four-dimensional judgments using a single assessment agent, whereas Multi-Agent Consensus introduced disagreement identification, multi-round negotiation, dimension locking, and Judge Agent adjudication after independent initial assessments by three Expert Agents. This controlled comparison indicates that the consensus process improves consistency between automated assessment results and the standardized language development measure.

In the literature benchmark comparison, the best Pearson r values previously reported by Qin et al. for GPT-4o, Qwen2-72B-Instruct, and Doubao-Pro-32k are 0.80, 0.74, and 0.73, respectively, with GPT-4o achieving the highest result. The Pearson r of Multi-Agent Consensus in this work is 0.84, which is approximately 0.04 higher than the best previously reported result. Overall, the controlled comparison validates the contribution of the consensus negotiation mechanism to automated language assessment validity, and the literature benchmark comparison further shows that the proposed method is competitive with existing related studies.

As shown in Table 3, Multi-Agent Consensus achieved higher mean Pearson and Spearman correlations with MSEL Vdq within the ASD group than Single Agent and Initial Expert Aggregation. Its mean Pearson correlation was 0.674, compared with 0.843 in the pooled sample. These results provide additional evidence of criterion validity among children with ASD, while highlighting that the pooled correlation should not be interpreted as ASD-specific performance.

Table 3. Within-ASD correlations between automated assessment scores and MSEL VDQ ($n = 29$).

Method	Pearson r	Spearman ρ
Single Agent	0.595 $\pm$ 0.123	0.560 $\pm$ 0.107
Initial Expert Aggregation	0.629 $\pm$ 0.065	0.609 $\pm$ 0.039
Multi-Agent Consensus	0.674 $\pm$ 0.022	0.621 $\pm$ 0.039

Note: Values are means $\pm$ sample standard deviations across five evaluations. Bold values indicate the highest mean correlation in each column.

4.3. Performance Comparison Under Different Consensus Configurations

4.3.1. Comparison Settings and Ablation Design

This section analyzes the effects of different consensus configurations on the performance of automated language ability assessment and further examines the role of each component in Multi-Agent Consensus. The experiments include six assessment configurations. Single Agent directly generates final results using one assessment agent. Initial Expert Aggregation averages the total scores from the three Expert Agents' independent initial assessments and does not use the Judge Agent. This setting serves as an aggregation baseline that introduces only independent judgments from multiple experts. Multi-Agent Consensus denotes the full multi-agent consensus framework proposed in this paper. Based on the full framework, three mechanism-level ablations are further designed: Without Negotiation, Single-Round Negotiation, and Without Dimension Locking. The Without Negotiation configuration retains Judge Agent disagreement detection and final adjudication but removes inter-expert negotiation rounds. Single-Round Negotiation retains only one negotiation round. The Without Dimension Locking configuration removes the mechanism that locks dimensions after expert agreement. These settings are used to analyze how different components affect assessment performance.

Pearson and Spearman correlation coefficients are used to evaluate the correlation between the automated language ability assessment total score and the MSEL VDQ external criterion. For both coefficients, the mean and sample standard deviation over five independent runs are reported. Cross-run stability is measured using the intraclass correlation coefficient (ICC), specifically ICC(A,1) [41], which is calculated from the child-aligned 79×5 prediction matrix.

4.3.2. Overall Performance Results

Table 4 summarizes the performance of different assessment configurations over five independent runs. Compared with Single Agent, Initial Expert Aggregation improved Pearson r from 0.802 to 0.814 and Spearman ρ from 0.792 to 0.807, indicating that multiple independent expert judgments already provide a performance gain. Building on this aggregation baseline, the full Multi-Agent Consensus further increased Pearson r and Spearman ρ to 0.843 and 0.835, respectively, corresponding to improvements of 0.029 and 0.028 over Initial Expert Aggregation.

These gains suggest that the advantage of the full framework comes not only from aggregating multiple expert outputs but also from revising and integrating initial judgments through procedural mechanisms such as disagreement identification, expert negotiation, dimension locking, and final adjudication. The standard deviations of Pearson r and Spearman ρ for the full framework were 0.014 and 0.013, respectively, which are lower than those of Single Agent (0.030 and 0.025). ICC(A,1) also increased from 0.839 for Single Agent to 0.920 for the full framework, indicating that Multi-Agent Consensus improves assessment validity while maintaining strong cross-run stability.

The mechanism-level comparisons showed that the full framework achieved the highest mean Pearson and Spearman correlations, with modest differences relative to the three

ablation configurations. Without Negotiation yielded mean Pearson and Spearman correlations of 0.838 and 0.830, respectively. Single-Round Negotiation was closest to the full framework, with corresponding values of 0.840 and 0.832, whereas Without Dimension Locking yielded 0.837 and 0.828. These comparisons characterize the final-score performance of the complete consensus workflow, while the following process analysis examines how negotiation and dimension locking shape consensus formation and the allocation of discussion across dimensions.

Table 4. Validity and cross-run consistency across different consensus configurations.

Method	Pearson r	Spearman ρ	ICC(A,1)
Single Agent	0.802 ± 0.030	0.792 ± 0.025	0.839
Initial Expert Aggregation	0.814 ± 0.010	0.807 ± 0.008	0.918
Without Negotiation	0.838 ± 0.015	0.830 ± 0.015	0.923
Single-Round Negotiation	0.840 ± 0.016	0.832 ± 0.017	0.927
Without Dimension Locking	0.837 ± 0.015	0.828 ± 0.013	0.926
Multi-Agent Consensus	0.843 ± 0.014	0.835 ± 0.013	0.920

Note: Pearson r and Spearman ρ are reported as mean $\pm$ sample SD over five independent runs. ICC(A,1) = intraclass correlation coefficient, absolute-agreement single-measure form; it measures cross-run consistency across the five independent runs. Bold values indicate the best Pearson r and Spearman ρ .

4.4. Consensus Formation Process Analysis

To complement the comparisons of criterion validity and cross-run stability, this section examines the consensus formation trajectory, dimension-level variation, and interaction efficiency. It first investigates how multi-round negotiation gradually reduces disagreements at the child level. It then compares the extent to which language ability dimensions differ in negotiation-based resolution and reliance on judge adjudication. Finally, it analyzes how dimension locking reduces repeated discussion and maintains the stability of established consensus.

4.4.1. Dynamic Dimension Agreement Distribution

To characterize the child-level consensus formation process in Multi-Agent Consensus, this study counts, for each child at each stage, the number of the four language ability dimensions that have reached expert agreement. Figure 3 summarizes the results for 79 children in each run, with percentages averaged across five independent runs. Each bar represents the proportion of children with 0, 1, 2, 3, or 4 agreed dimensions at stages 0–5 on the horizontal axis. Stage 0 denotes the agreement status obtained from the independent initial assessments of the three Expert Agents before negotiation. Stages 1–5 denote the cumulative agreement status after negotiation rounds 1 through 5. In the stacked categories, “4 dimensions agreed” means that all four language ability dimensions for a child have reached expert agreement, whereas “0–3 dimensions agreed” indicates that varying degrees of dimension-level disagreement remain in the child’s assessment.

As shown in Figure 3, the number of agreed dimensions at the child level generally increases as negotiation proceeds, indicating that multi-agent negotiation effectively moves expert disagreements toward consensus. In the independent initial assessment stage, substantial multidimensional disagreements are observed. Only 6.6% of children show agreement across all four dimensions, whereas children with zero or one agreed dimension account for 59.2% in total. After the first negotiation round, the distribution changes markedly. The proportion of children with agreement across all four dimensions increases to 17.0%. The proportion with three or four agreed dimensions increases from 19.0% to 45.3%, whereas the proportion with zero or one agreed dimension decreases to 25.6%. This

shows that the first negotiation round resolves a substantial portion of initial disagreements and marks the stage with the most pronounced change in the consensus formation process.

In subsequent rounds, the proportion of agreed dimensions continues to increase, but the rate of growth gradually decreases. After the fifth round, the proportion of children with agreement across all four dimensions further increases to 29.9%. The proportion with three or four agreed dimensions reaches 65.6%, whereas the proportion with zero or one agreed dimension decreases to 11.4%. These results show that multi-round negotiation can continuously increase expert agreement. However, as easier disagreements are gradually resolved, the rate at which new consensus is formed decreases. At the same time, some children still do not reach agreement across all four dimensions after the fifth round. This suggests that certain disagreements are difficult to resolve through expert negotiation alone and still require final adjudication by the Judge Agent.

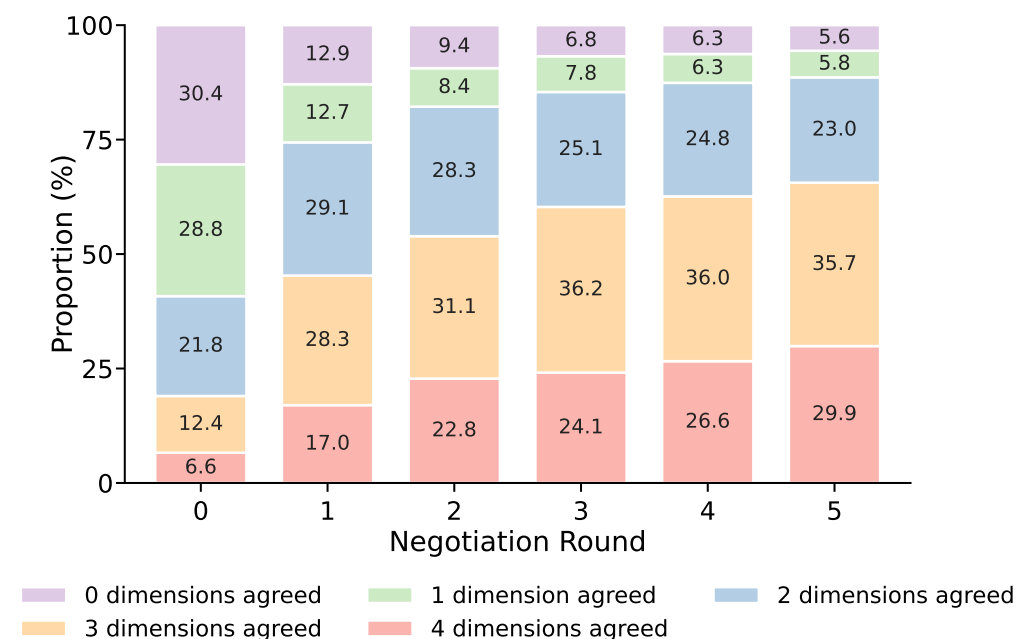


Figure 3. Distribution of the number of agreed dimensions at the child level across negotiation rounds in the Multi-Agent Consensus framework.

To further examine whether these changes in consensus are consistent across language ability dimensions, this study analyzes the distribution of decision sources across negotiation rounds. It also examines whether the remaining disagreements requiring judge intervention are concentrated in specific dimensions. Specifically, this study calculates the proportions of three decision sources for each of the four dimensions: Initial Agreement, Cumulative New Agreement, and Judge Required. The results are shown in Figure 4. For each dimension, the denominator is the 79 child-level assessments in that dimension in each run. Overall, all four dimensions show an increase in the Cumulative New Agreement category and a decrease in the Judge Required category as negotiation rounds proceed. This indicates that the negotiation mechanism can transform initial disagreements into expert agreement across different language ability dimensions.

The difficulty of resolving disagreements varies across dimensions. Although the four dimensions show similar Initial Agreement proportions in the independent initial assessment stage, clear differences emerge after five rounds of negotiation. For Language Comprehension, the Cumulative New Agreement proportion reaches only 20.3%, and 45.3% of cases still require Judge Agent adjudication, representing the highest reliance on Judge Agent adjudication among the four dimensions. In contrast, the Cumulative New

Agreement proportions reach 43.1% for Logical Expression and 43.6% for Communication Skills, while the final Judge Required proportion is 23.0% for both dimensions. Vocabulary Competence falls between these cases, with a final Judge Required proportion of 30.1%. These results indicate that unresolved disagreements are concentrated more strongly in Language Comprehension than in the other dimensions.

These results further show that expert negotiation in Multi-Agent Consensus can resolve a substantial proportion of initial disagreements. However, the negotiability of different language ability dimensions varies. For dimensions such as Logical Expression and Communication Skills, experts can more readily compare evidence related to language organization, response patterns, and turn-taking behavior. As a result, they are more likely to reach agreement through multi-round negotiation. In contrast, Language Comprehension involves more complex judgments, including contextual semantics, implicit intentions, and developmental boundary distinctions. Therefore, a higher proportion of cases in this dimension still require the Judge Agent to make the final adjudication by integrating information from the negotiation process with the original evidence.

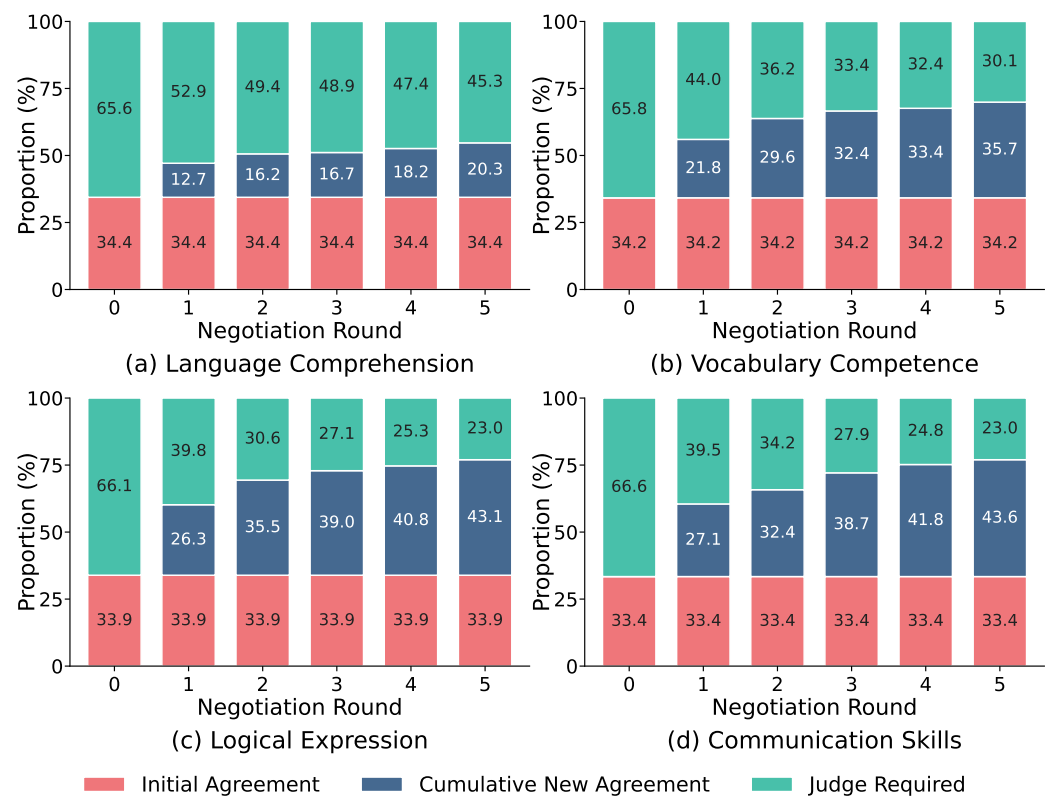


Figure 4. Round-level decision sources for consensus formation across four language ability dimensions.

4.4.2. Effect of Dimension Locking on Interaction Efficiency

In the full Multi-Agent Consensus framework, once a language ability dimension reaches expert agreement, its current assessment result is locked, and the dimension is excluded from subsequent negotiation rounds. To examine whether this mechanism reduces repeated interaction and maintains the stability of established consensus, this study compares process metrics between the full framework and the configuration Without Dimension Locking. The results are shown in Table 5.

In Table 5, Workload denotes the average number of dimension-round units discussed for each child and is used to measure the interaction cost of the negotiation process. Consensus Drift indicates the proportion of cases in which a dimension that has reached agreement later returns to a non-consensus state when discussion continues. Rating Change

indicates the proportion of cases in which the rating assigned to an agreed dimension changes during subsequent discussion. Because the full framework locks a dimension immediately after agreement is reached, the latter two metrics apply only to the Without Dimension Locking setting and are marked as N/A for the full framework.

Table 5. Effect of dimension locking on interaction efficiency and consensus stability.

Configuration	Workload	Consensus Drift	Rating Change
Multi-Agent Consensus	8.58 ± 0.62	N/A	N/A
Without Dimension Locking	16.65 ± 0.54	0.53% ± 0.33%	0.23% ± 0.17%

Note: Values are reported as mean ± sample SD over five independent runs. Workload denotes the average number of discussed dimension-round units per child. Consensus Drift and Rating Change are calculated only for dimensions that remain in negotiation after consensus has been reached. In Multi-Agent Consensus, such dimensions are locked and excluded from later negotiation; therefore, these metrics are not applicable (N/A).

As shown in Table 5, removing dimension locking increases the average Workload per child from 8.58 to 16.65, which is nearly twice that of the full framework. This indicates that allowing agreed dimensions to continue participating in later negotiation substantially increases repeated dimension-level discussion. Meanwhile, in the Without Dimension Locking setting, Consensus Drift is only 0.53%, and Rating Change is only 0.23%. This shows that most dimensions that have already reached agreement remain stable during repeated discussion, and continued discussion rarely leads to substantive changes in the assessment results. Taken together, these results indicate that dimension locking reduces unnecessary repeated discussion while keeping subsequent negotiation focused on unresolved disagreements. Supplementary File S2 reports a six-child measurement of model requests, token consumption, assessment time, and negotiation workload across assessment configurations.

4.5. Explainability and Human Expert Evaluation

This section evaluates the explanatory credibility of the proposed method from two perspectives: a representative case study and a human expert evaluation. The case study illustrates how Multi-Agent Consensus transforms initial expert disagreements into a traceable process of consensus formation and decision justification. The human expert evaluation further examines whether the final assessment levels, evidence-based rationales, and negotiation trajectories are acceptable from the perspective of domain experts.

4.5.1. Case Study of an Explainable Assessment Report

To illustrate how the negotiation process supports explainable assessment, this study analyzes a representative case involving initial disagreements across all four language ability dimensions. As shown in Figure 5, Vocabulary Competence and Logical Expression reached agreement and were locked after the third negotiation round, while Communication Skills reached agreement after the fifth round. Language Comprehension did not reach unanimous agreement after negotiation and therefore required final adjudication by the Judge Agent, mainly because the experts differed in how they interpreted the child's understanding of abstract concepts.

The disagreement over Language Comprehension centered on the trade-off between context-supported concrete understanding and abstract or decontextualized concept comprehension. Some experts regarded the child's ability to return objects to their proper locations and identify common objects as evidence of basic language comprehension. Other experts placed greater emphasis on the child's difficulties with decontextualized concepts, such as geometric shape names, and argued that comprehension remained unstable when contextual support was removed. After multiple rounds of negotiation, the three experts

still failed to reach a unanimous judgment on this dimension; therefore, the Judge Agent performed final adjudication. Considering both the child's concrete-context understanding and limitations in abstract concept comprehension, the Judge Agent finally rated this dimension as "fair".

The discussion of Vocabulary Competence focused mainly on vocabulary repertoire and lexical retrieval fluency. The child was able to use some concrete words, such as object names, indicating a certain vocabulary foundation. However, repeated expressions such as "guo guo guo", disfluent lexical retrieval, and occasional inaccurate word use also suggested limitations in vocabulary use. Through negotiation, the experts gradually converged on a shared judgment. Expert 2 revised the original higher rating to "fair" and acknowledged limiting factors such as lexical retrieval difficulties. This dimension reached unanimous agreement among the three experts in the third round and was locked as "fair".

A: Case Overview

Case ID: Child-16		
Final Explainable Assessment:		
Dimension	Level	Source
LC	Fair	Judge Adjudication
VC	Fair	Expert Consensus
LE	Fair	Expert Consensus
CS	Fair	Expert Consensus

B: Consensus Negotiation Process

Dimension	Initial assessment	Key negotiation	Negotiation outcome	Final source
LC	E1: Fair E2: Excellent E3: Good	Abstract language understanding vs contextual reasoning	Unresolved → Judge Adjudication	Judge Adjudication
VC	E1: Poor E2: Excellent E3: Fair	Vocabulary richness vs retrieval difficulty	Round 3 → Fair (Consensus & Locked)	Expert Consensus
LE	E1: Poor E2: Good E3: Fair	Linguistic complexity vs contextual coherence	Round 3 → Fair (Consensus & Locked)	Expert Consensus
CS	E1: Fair E2: Good E3: Poor	Passive response vs active social engagement	Round 5 → Fair (Consensus & Locked)	Expert Consensus

Figure 5. A representative case illustrating explainable language assessment and consensus formation in the Multi-Agent Consensus framework. LC, VC, LE, and CS denote Language Comprehension, Vocabulary Competence, Logical Expression, and Communication Skills, respectively. E1–E3 denote the three independent expert agents. The green and orange elements indicate expert consensus and Judge Agent adjudication, respectively. Blue shading is used for descriptive information and table headers.

The disagreement over Logical Expression mainly arose from different emphases on expressive complexity and context-based expression. Some experts emphasized that the child's utterances were mostly short sentences and lacked complex sentence structures, independent narration, and causal elaboration. Other experts argued that the child could produce basic expressions in the current interactional context and that the sentence structures were generally complete. As negotiation progressed, the expert judgments gradually

converged: Expert 1 revised the rating from “poor” to “fair”, and Expert 2 revised the rating from “good” to “fair”. The three experts eventually reached agreement in the third round and locked this dimension as “fair”.

The core disagreement over Communication Skills concerned how to evaluate the child’s overall performance across spontaneous initiation, response quality, and interaction maintenance. On the one hand, the experts noted limited spontaneous initiation, limited topic extension, and shallow interactional depth. On the other hand, they also considered that the child still demonstrated joint attention and basic social responsiveness. After further negotiation, the experts gradually reached agreement on the balance between limited initiative and basic social participation. Expert 3 revised the rating from “poor” to “fair”, and the three experts ultimately reached agreement in the fifth round, locking this dimension as “fair”.

This case demonstrates the complete consensus formation process in Multi-Agent Consensus, from independent assessment and disagreement identification to expert negotiation and final adjudication. Unlike assessment methods that output only final scores, the proposed framework preserves the decision source for each dimension: some dimensions reach consensus through expert negotiation, whereas others are finally adjudicated by the Judge Agent when disagreement persists. In addition, the negotiation process records changes in expert opinions, key points of disagreement, and the rationale for the final decisions, thereby providing a traceable formation path for the assessment results.

4.5.2. Human Evaluation of Explainability

To assess the explanatory quality of the assessment reports generated by Multi-Agent Consensus, this study conducted a human expert evaluation. The assessment reports for all 79 children served as the evaluation materials and were independently rated by two doctoral students in intelligent education whose research focuses on child language development and one postdoctoral researcher in speech and hearing rehabilitation science. The anonymized evaluation materials included each child’s age and sex, dialogue transcript, final assessment results, supporting evidence, decision sources, and consensus negotiation trace. The evaluators were unaware of the children’s MSEL results during rating. The evaluation included five metrics. Assessment Reasonableness assesses whether the final assessment levels are consistent with the child’s language performance and the relevant evidence. Evidence Sufficiency examines whether the key assessment evidence is adequate to support the final assessment results. Consensus Trace Reliability examines whether the identification of expert disagreements, revision of opinions, and formation of consensus are consistent with reasonable expert decision-making logic. Decision Transparency examines whether the rationale and formation process of the final decisions are clear and understandable. Overall Explainability provides a comprehensive evaluation of the explanatory quality of the assessment report as a whole.

The human experts rated each report using a 5-point Likert scale, where 1 indicated complete disagreement and 5 indicated complete agreement [42]. For each evaluation metric, we first calculated each expert’s average score across the 79 assessment reports and then reported the mean and sample standard deviation across the three expert-level average scores. The study also used the two-way random-effects absolute-agreement average-measures intraclass correlation coefficient (ICC), specifically $ICC(A,k)$, with $k = 3$, to evaluate inter-rater reliability based on the report-level ratings from the three experts. The human evaluation results are shown in Table 6. The reported mean scores for all five evaluation metrics exceed 4.2, indicating that the experts generally recognized the quality of the explanations generated by the framework. Decision Transparency achieved the highest score (4.46), suggesting that the reports clearly presented the rationale for the final

decisions and the process through which they were formed. Assessment Reasonableness also received a high score (4.41), indicating that the final levels were largely consistent with the children's language performance and the relevant evidence.

In addition, the ICC(A,k) values for the five evaluation metrics range from 0.838 to 0.897, indicating high agreement among the three experts in their ratings of the explanatory quality of the Multi-Agent Consensus assessment reports. Decision Transparency achieved an ICC(A,k) of 0.897, while the other evaluation metrics also maintained high levels of consistency. These results suggest that the experts made consistent judgments regarding the explanatory quality of the assessment reports, supporting the reliability of the human evaluation.

Table 6. Human expert evaluation of the explainability of Multi-Agent Consensus assessment reports for all 79 children.

Evaluation Metric	Mean $\pm$ SD	ICC(A,k)
Assessment Reasonableness	4.41 $\pm$ 0.13	0.874
Evidence Sufficiency	4.25 $\pm$ 0.13	0.838
Consensus Trace Reliability	4.23 $\pm$ 0.18	0.879
Decision Transparency	4.46 $\pm$ 0.17	0.897
Overall Explainability	4.35 $\pm$ 0.12	0.869

Note: Mean $\pm$ SD denotes the mean and sample standard deviation across the three expert-level average scores. For each evaluation metric, each expert-level average score was calculated across the 79 assessment reports. ICC(A,k) = intraclass correlation coefficient, two-way random-effects absolute-agreement average-measures form with $k = 3$, calculated from the report-level ratings of the three experts. Higher scores indicate better perceived explainability.

Although the overall ratings are high, the scores for all evaluation metrics remain below the maximum value, indicating that the generated reports still leave room for improvement. The case analysis suggests that, in complex assessment scenarios, the reports should strengthen the explicit connection between assessment evidence and final judgments to improve Evidence Sufficiency. The links among opinion revision, evidence citation, and final decision-making during negotiation could also be refined to improve Consensus Trace Reliability. In addition, improving the organization and presentation of report information may enhance Overall Explainability.

5. Discussion

The proposed Multi-Agent Consensus framework models automated language ability assessment for children with ASD as a structured multi-agent consensus process. Beyond aggregating independent assessments, the framework makes disagreements explicit and records how judgments are revised and integrated through evidence-based negotiation. Because language comprehension, vocabulary competence, logical expression, and communication skills rely on partly different behavioral evidence, a single assessor may overemphasize particular linguistic cues. Multi-agent negotiation makes these judgment rationales explicit and comparable, thereby reducing uncertainty related to evidence selection, rating boundaries, and explanatory preferences. The value of the proposed method therefore lies not only in improving correlation with the external criterion but also in producing a reviewable and traceable record of consensus formation.

The process analysis further reveals how the multi-agent consensus mechanism operates. The proportion of agreed-upon dimensions at the child level increases gradually across negotiation rounds, indicating that expert negotiation can transform a substantial portion of initial disagreements into shared judgments. However, the gains in later rounds gradually diminish, suggesting that consensus formation does not increase linearly with the number of rounds, but instead shows diminishing returns determined by the difficulty

of the remaining disagreements. The dimension-level results also show that the extent to which disagreements can be resolved through negotiation varies across language-ability dimensions. Logical Expression and Communication Skills more readily support consensus because experts can rely on observable evidence such as sentence organization, response quality, turn maintenance, and interactional initiative. In contrast, Language Comprehension involves more complex judgments concerning contextual semantics, implicit intentions, situational support, and developmental boundary distinctions, and is therefore more likely to involve persistent disagreements that require integrative adjudication by the Judge Agent. This finding suggests that disagreement itself has explanatory value in automated child language ability assessment: it not only reflects instability in model outputs but may also correspond to dimensions in real assessment tasks where evidence interpretation is difficult and ability boundaries are ambiguous.

Compared with existing automated assessment methods that mainly focus on final scores or rating prediction, another core contribution of the proposed framework is its explicit documentation of how assessment conclusions are formed. Multi-Agent Consensus produces not only final levels and total scores but also initial expert positions, disputed dimensions, opinion revisions, locking rounds, and adjudication sources. This shifts the explanation of assessment results from “result explanation” to “process explanation”. Both the case analysis and the human expert evaluation show that such process-level records help experts understand why an assessment conclusion was formed, which dimensions involved prior disagreement, which conclusions were derived from expert consensus, and which conclusions depended on integrative judgment by the Judge Agent. For language ability assessment in children with ASD, a task that relies heavily on natural interaction evidence and professional interpretation, improving predictive correlation alone is insufficient to support practical application. The ability to present a traceable evidence chain and decision path also affects the acceptability of automated assessment results to experts and their utility in expert review.

From an application perspective, the proposed method is better suited as an auxiliary analytical tool for human–artificial intelligence (AI) collaborative assessment. The language performance of children with ASD is highly heterogeneous, and the same segment of language behavior may simultaneously involve vocabulary, comprehension, expressive organization, and social communication. Comprehensive judgment therefore requires consideration of the child’s age, interactional context, and specific linguistic evidence. By systematically presenting differences in expert judgments and the process of consensus formation, the proposed framework can help professionals quickly identify high-consensus conclusions and highly disputed dimensions, and support focused review of cases requiring Judge Agent intervention or involving insufficient evidence. The dimension locking mechanism further shows that reducing repeated discussion of dimensions with stable agreement can lower interaction costs and focus subsequent negotiation on dimensions with remaining interpretive difficulty. Thus, the framework provides a feasible path for future AI-assisted assessment systems for child development, balancing assessment validity, process transparency, and review efficiency.

This study still has several limitations. First, the assessment in this study is based primarily on transcripts of parent–child free-play interactions, and multimodal information from speech, visual signals, and behavioral cues was not fully utilized. Therefore, the framework’s ability to model multimodal language performance requires further improvement. Second, although the three Expert Agents are instantiated as independent assessors, they are all based on the same base language model, and their professional backgrounds and cognitive preferences are not explicitly modeled. Future work could introduce Expert Agents with different areas of professional expertise to examine the influence of role hetero-

geneity on the consensus formation process. In addition, the current sample size and data sources remain limited, and the generalizability of the assessment results across regions, language environments, and clinical populations requires further validation. Although the human expert evaluation indicates that the reports generated by the proposed framework provide credible explanations, whether such reports can support sustained expert review, intervention decisions, and follow-up management in real clinical and educational settings still needs to be examined through longitudinal studies and real-world applications.

6. Conclusions

This paper proposed a Multi-Agent Consensus framework for explainable automated assessment of language ability in children with ASD using parent–child free-play dialogue transcripts. The framework models assessment as a consensus-formation process that combines independent Expert Agent judgments, dimension-level disagreement detection, evidence-based negotiation, dimension locking, and Judge Agent adjudication. Experiments on 79 children showed that the proposed method achieved higher consistency with MSEL VDQ scores than Single Agent and Initial Expert Aggregation baselines, while maintaining strong cross-run stability. Ablation experiments, process analyses, case-study analysis, and human expert evaluation further demonstrated that the framework can resolve assessment disagreements, reduce unnecessary repeated discussion, and generate reports with clear evidence, reliable consensus traces, and transparent decision paths. Overall, these findings suggest that multi-agent consensus provides a feasible approach to improving assessment validity while providing structured, traceable explanations for AI-assisted child language assessment. The observed correlations provide evidence of criterion validity; clinical validity and effectiveness in practice remain to be established beyond the current limited sample and data sources. Future work will incorporate multimodal evidence, introduce heterogeneous Expert Agents, and validate the framework in larger and more realistic clinical or educational settings.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/electronics15204581/s1>, File S1: Prompt templates, model inputs, and output specifications. Table S1: Prompt and input template index. File S2: Resource use and efficiency across assessment configurations. Table S2: Model requests and token consumption per successful assessment. Table S3: Assessment time and negotiation workload per successful assessment.

Author Contributions: Conceptualization, S.Q. and Q.L.; methodology, S.Q. and Q.L.; software, S.Q.; validation, S.Q., T.Y. and M.L.; formal analysis, S.Q.; investigation, S.Q., T.Y. and M.L.; resources, Q.L.; data curation, S.Q. and T.Y.; writing—original draft preparation, S.Q.; writing—review and editing, S.Q., T.Y., M.L. and Q.L.; visualization, S.Q.; supervision, Q.L.; project administration, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2022YFC2705201; the Leading Projects in Key Research Fields of Language funded by the National Language Commission, grant number LH24TD08; the Humanities and Social Science Project of the Ministry of Education of China, grant number 23YJC740035; the Youth Project of the National Social Science Fund of China, grant number 24CYY092; the Major Program of Humanities and Social Science of Shanghai Education Commission, grant number 2023SKZD07; the PI Team Building Project of Changning Maternity and Infant Health Hospital, grant number PI202432; and the State Language Commission’s “14th Five-Year Plan” Scientific Research Project, grant number YB145-21.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the University Committee on Human Research Protection of East China Normal University (protocol code HR374-2024) on 3 June 2026.

Informed Consent Statement: Written informed consent was obtained from the guardians of all children involved in the study before data collection.

Data Availability Statement: The datasets presented in this article are not publicly available due to privacy and ethical restrictions, as they contain sensitive information from children with developmental delay and children with autism spectrum disorder. Requests to access the datasets should be directed to the corresponding author, Qiaoyun Liu (qyliu@spe.ecnu.edu.cn). Access will be considered only where permitted by the relevant ethics approval, institutional requirements, and the consent conditions under which the data were collected.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (model: GPT-5.6 sol; OpenAI, San Francisco, CA, USA) for language editing and minor formatting assistance. The tool was not used to generate research ideas, design the study, collect or analyze data, interpret results, or draw scientific conclusions. The authors reviewed and edited all outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mundy, P.; Sigman, M.; Ungerer, J.; Sherman, T. Nonverbal communication and play correlates of language development in autistic children. *J. Autism Dev. Disord.* **1987**, *17*, 349–364. [\[CrossRef\]](#)
2. Lord, C.; Pickles, A. Language level and nonverbal social-communicative behaviors in autistic and language-delayed children. *J. Am. Acad. Child Adolesc. Psychiatry* **1996**, *35*, 1542–1550. [\[CrossRef\]](#)
3. Bennett, T.; Szatmari, P.; Georgiades, K.; Hanna, S.; Janus, M.; Georgiades, S.; Duku, E.; Bryson, S.; Fombonne, E.; Smith, I.; et al. Language impairment and early social competence in preschoolers with autism spectrum disorders: A comparison of DSM-5 profiles. *J. Autism Dev. Disord.* **2014**, *44*, 2797–2808. [\[CrossRef\]](#)
4. Taylor, L.J.; Maybery, M.T.; Grayndler, L.; Whitehouse, A.J. Evidence for shared deficits in identifying emotions from faces and from voices in autism spectrum disorders and specific language impairment. *Int. J. Lang. Commun. Disord.* **2015**, *50*, 452–466. [\[CrossRef\]](#)
5. Qin, S.; Liu, M.; Wei, T.; Liu, Q. Language proficiency assessment of autistic children using large language models. *Expert Syst. Appl.* **2026**, *298*, 129712. [\[CrossRef\]](#)
6. Feng, T.; Xu, A.; Lahiri, R.; Tager-Flusberg, H.; Kim, S.H.; Bishop, S.; Lord, C.; Narayanan, S. Can Generic LLMs Help Analyze Child-Adult Interactions Involving Children with Autism in Clinical Observation? *arXiv* **2024**, arXiv:2411.10761. [\[CrossRef\]](#)
7. Mullen, E.M. *Mullen Scales of Early Learning*; American Guidance Service: Circle Pines, MN, USA, 1995.
8. Wiig, E.H.; Semel, E.; Secord, W.A. *Clinical Evaluation of Language Fundamentals: CELF-5*; Pearson: Bloomington, MN, USA, 2013.
9. Dunn, D.M. *Peabody Picture Vocabulary Test*, 5th ed.; [Measurement instrument]; NCS Pearson: Bloomington, MN, USA, 2019.
10. Tardif, T.; Fletcher, P. *Chinese Communicative Development Inventories: User's Guide and Manual*; Peking University Medical Press: Beijing, China, 2008.
11. Mohanta, A.; Mittal, V.K. Classifying speech of ASD affected and normal children using acoustic features. In *Proceedings of the 2020 National Conference on Communications (NCC)*; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6. [\[CrossRef\]](#)
12. Vitthal, C.; Shreeharsha B.S.; Sabu, K.; Rao, P. Predicting lexical skills from oral reading with acoustic measures. *arXiv* **2021**, arXiv:2112.00635. [\[CrossRef\]](#)
13. Smith, M.L.; Munson, B. Validating acoustic predictors of sentence intelligibility using a racially diverse corpus. *J. Acoust. Soc. Am.* **2023**, *153*, A79. [\[CrossRef\]](#)
14. Nnamoko, N.; Karaminis, T.; Procter, J.; Barrowclough, J.; Korkontzelos, I. Automatic language ability assessment method based on natural language processing. *Nat. Lang. Process. J.* **2024**, *8*, 100094. [\[CrossRef\]](#)
15. Yang, Y.; Shen, Y.; Sun, T.; Xie, Y. Validating the Effectiveness of a Large Language Model-based Approach for Identifying Children's Development across Various Free Play Settings in Kindergarten. *arXiv* **2025**, arXiv:2505.03369. [\[CrossRef\]](#)
16. Fink, A.; Kosecoff, J.; Chassin, M.; Brook, R.H. Consensus methods: Characteristics and guidelines for use. *Am. J. Public Health* **1984**, *74*, 979–983. [\[CrossRef\]](#)
17. Black, N.; Murphy, M.; Lamping, D.; McKee, M.; Sanderson, C.; Askham, J.; Marteau, T. Consensus development methods: A review of best practice in creating clinical guidelines. *J. Health Serv. Res. Policy* **1999**, *4*, 236–248. [\[CrossRef\]](#)
18. Fitch, K.; Bernstein, S.J.; Aguilar, M.D.; Burnand, B.; LaCalle, J.R.; Lázaro, P.; van het Loo, M.; McDonnell, J.; Vader, J.; Kahan, J.P. *The RAND/UCLA Appropriateness Method User's Manual*; Technical Report MR-1269-DG-XII/RE; RAND Corporation: Santa Monica, CA, USA, 2001. Available online: https://www.rand.org/pubs/monograph_reports/MR1269.html (accessed on 19 August 2026).

19. Nair, R.; Aggarwal, R.; Khanna, D. Methods of formal consensus in classification/diagnostic criteria and guideline development. *Semin. Arthritis Rheum.* **2011**, *41*, 95–105. [\[CrossRef\]](#)
20. Xu, A.; Hebbbar, R.; Lahiri, R.; Feng, T.; Butler, L.; Shen, L.; Tager-Flusberg, H.; Narayanan, S. Understanding Spoken Language Development of Children with ASD Using Pre-trained Speech Embeddings. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, 20–24 August 2023*; Harte, N., Carson-Berndsen, J., Jones, G., Eds.; ISCA: Grenoble, France, 2023; pp. 4633–4637. [\[CrossRef\]](#)
21. Salem, A.C.; MacFarlane, H.; Adams, J.R.; Lawley, G.O.; Dolata, J.K.; Bedrick, S.; Fombonne, E. Evaluating atypical language in autism using automated language measures. *Sci. Rep.* **2021**, *11*, 10968. [\[CrossRef\]](#)
22. Gale, R.; Dolata, J.; Prud'hommeaux, E.; Van Santen, J.; Asgari, M. Automatic assessment of language ability in children with and without typical development. In *Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*; IEEE: Piscataway, NJ, USA, 2020; pp. 6111–6114. [\[CrossRef\]](#)
23. Gretter, R.; Matassoni, M.; Allgaier, K.; Tchistiakova, S.; Falavigna, D. Automatic assessment of spoken language proficiency of non-native children. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2019; pp. 7435–7439. [\[CrossRef\]](#)
24. Lin, C.S.; Chang, S.H.; Liou, W.Y.; Tsai, Y.S. The development of a multimedia online language assessment tool for young children with autism. *Res. Dev. Disabil.* **2013**, *34*, 3553–3565. [\[CrossRef\]](#)
25. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 2511–2522. [\[CrossRef\]](#)
26. Xia, W.; Mao, S.; Zheng, C. Empirical Study of Large Language Models as Automated Essay Scoring Tools in English Composition: Taking TOEFL Independent Writing Task for Example. *arXiv* **2024**, arXiv:2401.03401. [\[CrossRef\]](#)
27. Lee, S.; Cai, Y.; Meng, D.; Wang, Z.; Wu, Y. Unleashing large language models' proficiency in zero-shot essay scoring. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 181–198. [\[CrossRef\]](#)
28. Bannò, S.; Vydana, H.K.; Knill, K.M.; Gales, M.J.F. Can GPT-4 do L2 analytic assessment? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), Mexico City, Mexico*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 149–164. Available online: <https://aclanthology.org/2024.bea-1.14/> (accessed on 19 August 2026).
29. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623. [\[CrossRef\]](#)
30. Kim, S.; Shin, J.; Cho, Y.; Jang, J.; Longpre, S.; Lee, H.; Yun, S.; Shin, R.S.; Kim, S.; Thorne, J.; et al. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. In *Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024*. Available online: <https://openreview.net/forum?id=8euJaTveKw> (accessed on 19 August 2026).
31. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.V.; Chi, E.H.; Narang, S.; Chowdhery, A.; Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023*. Available online: <https://openreview.net/forum?id=1PL1NIMMrw> (accessed on 19 August 2026).
32. Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, Q.; Liu, T.; et al. Large Language Models are Not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 9440–9450. [\[CrossRef\]](#)
33. Chan, C.M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; Liu, Z. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. In *Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024*. Available online: https://proceedings.iclr.cc/paper_files/paper/2024/hash/25cc3adf8c85f7c70989cb8a97a691a7-Abstract-Conference.html (accessed on 21 September 2026).
34. Qian, Y.; Zhang, S.; Zhou, Y.; Ding, H.; Socolinsky, D.; Zhang, Y. Enhancing LLM-as-a-Judge via Multi-Agent Collaboration. In *Proceedings of the AAAI 2025 Workshop on Advancing LLM-Based Multi-Agent Collaboration, Philadelphia, PA, USA, 4 March 2025*. Available online: <https://www.amazon.science/publications/enhancing-llm-as-a-judge-via-multi-agent-collaboration> (accessed on 21 September 2026).
35. Wang, J.; Wang, J.; Athiwaratkun, B.; Zhang, C.; Zou, J.Y. Mixture-of-Agents Enhances Large Language Model Capabilities. In *Proceedings of the Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025*. Available online: https://proceedings.iclr.cc/paper_files/paper/2025/hash/5434be94e82c54327bb9dcdf7fca52b6-Abstract-Conference.html (accessed on 21 September 2026).
36. Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J.B.; Mordatch, I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In *Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024*; Volume 235, pp. 11733–11763. Available online: <https://proceedings.mlr.press/v235/du24e.html> (accessed on 21 September 2026).

37. Kaesberg, L.B.; Becker, J.; Wahle, J.P.; Ruas, T.; Gipp, B. Voting or Consensus? Decision-Making in Multi-Agent Debate. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 27 July–1 August 2025; pp. 11640–11671. Available online: <https://aclanthology.org/2025.findings-acl.606/> (accessed on 21 September 2026).
38. Gao, Z.; Zhang, S.; McLoughlin, I.; Yan, Z. Paraformer: Fast and Accurate Parallel Transformer for Non-autoregressive End-to-End Speech Recognition. In Proceedings of the Interspeech 2022, ISCA, Incheon, Republic of Korea, 18–22 September 2022; pp. 2063–2067. [CrossRef]
39. Chen, Y.; Zheng, S.; Wang, H.; Cheng, L.; Chen, Q.; Qi, J. An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification. In Proceedings of the Interspeech 2023, ISCA, Dublin, Ireland, 20–24 August 2023; pp. 2228–2232. [CrossRef]
40. Qwen Team. Qwen3.6-35B-A3B. 2026. Available online: <https://huggingface.co/Qwen/Qwen3.6-35B-A3B> (accessed on 18 August 2026).
41. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef]
42. Likert, R. A technique for the measurement of attitudes. *Arch. Psychol.* **1932**, *140*, 1–55.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.