# Copiloting Diagnosis of Autism in Real Clinical Scenarios via LLMs

Yi Jiang, Qingyang Shen, Shuzhong Lai, Shunyu Qi Qian Zheng\*, Lin Yao\*, Yueming Wang, Gang Pan

Zhejiang University

#### **Abstract**

Autism spectrum disorder(ASD) is a pervasive developmental disorder that significantly impacts the daily functioning and social participation of individuals. Despite the abundance of research focused on supporting the clinical diagnosis of ASD, there is still a lack of systematic and comprehensive exploration in the field of methods based on Large Language Models (LLMs), particularly regarding the real-world clinical diagnostic scenarios based on Autism Diagnostic Observation Schedule, Second Edition (ADOS-2). Therefore, we have proposed a framework called ADOS-Copilot, which strikes a balance between scoring and explanation and explored the factors that influence the performance of LLMs in this task. The experimental results indicate that our proposed framework is competitive with the diagnostic results of clinicians, with a minimum MAE of 0.4643, binary classification F1-score of 81.79%, and ternary classification F1-score of 78.37%. Furthermore, we have systematically elucidated the strengths and limitations of current LLMs in this task from the perspectives of ADOS-2, LLMs' capabilities, language, and model scale aiming to inspire and guide the future application of LLMs in a broader fields of mental health disorders. We hope for more research to be transferred into real clinical practice, opening a window of kindness to the world for eccentric children.

# 1 Introduction

Autism spectrum disorder (ASD) is a pervasive neurodevelopmental disorder characterized by restricted social communication, repetitive behaviors, and specific interests. It significantly impacts the quality of life and well-being of individuals affected [27]. According to a 2021 report from the Centers for Disease Control and Prevention (CDC) in the United States, the prevalence of ASD among children under the age of eight is approximately 1 in 44 [44]. However, research indicates that the neural impairments in individuals can be early detected and timely intervention have been shown to improve language abilities and behavioral habits in children with autism, enabling them to reintegrate into society successfully. Currently, clinical diagnosis primarily relies on the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) [46], and heavily depends on the professional expertise of doctors. However, due to limitations in the number of qualified professionals and the inherent individual differences among children, diagnostic errors and missed cases are inevitable [26], causing children to miss the optimal treatment window.

With the advancement of computer technology, numerous computer-assisted methods for ASD diagnosis have emerged. One category involves simplifying clinical paradigms and collecting data from various modalities such as electroencephalography (EEG) [25], eye-tracking (ET) [61],

<sup>\*</sup> Correspondence to: {qianzheng,lin.yao}@zju.edu.cn

video [41] or multi-modal [12, 38] for diagnostic purposes. However, due to the simplification of real clinical paradigms, their clinical utility is significantly compromised. On the other hand, another category involves utilizing authentic paradigms and conducting automatic assessments by directly collecting data, such as speech [54, 57], from real diagnostic. Their limitations lie in the fact that the method operates as a black box, lacking sufficient interpretability, and therefore cannot effectively assist doctors in decision-making.

Recently, the remarkable capabilities of LLMs have sparked a research frenzy. Their unique ability to comprehend and reason with long contexts, coupled with the powerful knowledge base acquired through pre-training, allows them to better adapt to downstream tasks even without fine-tuning. This has led to the increasing utilization of LLMs in various diagnostic scenarios [48, 41, 28], with significant improvements in their effectiveness. Therefore, we aim to input interactive dialogue texts between child and doctor from the real clinic diagnosis into LLMs to enable the generation of score and explanation for corresponding items for diagnosis. But base on our preliminary experiments, we must acknowledge that the current large-language models are not yet capable of directly handling this task. Therefore, our paper focuses on three main aspects:

- We proposed an evaluation framework called ADOS-Copilot, which leverages the techniques
  of In-context Enhancement, Interpretability Augmentation and Adaptive Fusion to
  address the limitations of LLMs in ASD diagnosis task based on real ADOS-2 clinic
  scenario.
- We conducted a comprehensive analysis of the reasons why existing mainstream largelanguage models are unable to perform well in ADOS-2 clinical diagnosis tasks.
- Extensive experimental results and analysis demonstrated that our proposed framework
  is competitive with doctors' diagnosis result and provides detail evidence to support the
  assessment, thereby assisting doctors in making more accurate and objective diagnoses and
  treatments.

# 2 Related work

# 2.1 Computer-aided ASD diagnosis

To develop automated diagnosis of ASD using computer technology, researchers have extensively investigated atypical behavioral features of ASD through methods like EEG [25, 9], fMRI [55], and ET [21, 61, 62, 20]. [25] proposed a multimodal diagnostic framework for identifying ASD by combining EEG and ET data. [16] designed a simulated interaction task that uses a standard 7-minute simulated dialog via video to assess multiple biomarkers of social interaction deficits, including gaze behavior, facial expressions, and voice characteristics. This approach specifically targets joint attention (JA) impairments, emotional disorders, and language impairments in ASD. Acoustic and text-based features have also shown relevance in assessing children's language and communicative behaviors for ASD diagnosis [49, 6]. [3] explored the semantic and pragmatic language features in children to understand their significance in diagnosing ASD. To overcome the requirement for specialized professionals and extensive resources, some studies have adopted a natural language processing approach using electronic health records [53] and even online text [34, 11].

# 2.2 ADOS-2 with machine learning

Although these studies above have made considerable progress in ASD diagnosis, translating research findings from lab settings back into clinical practice remains challenging. To address this, researchers extended their studies to the clinical environment of ADOS-2, for more details about ADOS-2 see Appendix C, despite the difficulties in data acquisition. Various paradigms based on ADOS-2 have been proposed for feature extraction and diagnosis. For instance, [41] assessed ASD language, cognition, and attention by transforming ADOS-2 evaluation tasks into 9 social skill task scenarios, utilizing image-language pre-training models for score prediction. Based on the ADOS-2 process, [13] designed and proposed a standardized platform for stimulating, gathering, analyzing, modeling, and interpreting human behavioral data for ASD diagnosis. Additionally, [13, 43, 47, 54] directly used speech data from the ADOS-2 clinical assessment process for ASD prediction. Among these, [43] used speech data from the entire process, while [54] predicted ADOS-2 Calibrated Severity Scores (CSS).

Table 1: Comparison of our work with other ADOS-2-based approaches. Our work leverages the full spectrum of speech data from the ADOS-2 M3 clinical evaluation, allowing for automated scoring of all language-related items, with enhanced accuracy achieved through a combination of large language models and rule-based methods.

Works	Participant	Paradigm	Clinical Data	Used ADOS-2 Tasks	Data Type	Multimodal	LLM	ML	Explainability	ASD Diagnosis	ADOS-2 item scoring	CSS scoring
Liu [41]	Children	Based on ADOS-2	×	NA	Video	1	×	✓	×	×	NA	NA
Cheng [13]	Children	Bsaed on ADOS-2	×	NA	Video	✓	×	✓	✓	✓	NA	NA
MacFarlane [43]	Children	ADOS-2	✓	1	Audio	×	×	1	✓	✓	×	×
Ahn [4]	Children	ADOS-2	✓	All	Video	×	×	1	×	✓	×	Regression
Sadiq [54]	Children	ADOS-2	✓	All	Audio	×	×	1	×	1	×	Regression
Hu [28]	Adults	ADOS-2	✓	11/15	Audio	×	✓	×	×	×	Binary Classification	×
Our Work	Children	ADOS-2	✓	All	Audio	×	1	1	1	1	Ternary Classification	Ternary Classification

# 2.3 LLMs in healthcare

Although there are still some doubts regarding the application of large language models (LLMs) in the medical field, research has been extensively conducted in this area [15, 24, 29, 70, 45]. LLMs have demonstrated promising contextual understanding and zero-shot and few-shot capabilities for many medical scales and subjective descriptions. Studies [48, 19, 1, 40] used case vignettes to evaluate the differences between LLM diagnoses and expert diagnoses in areas such as depression [48, 19] and diabetes [1]. Studies [64, 35] compared the performance of various LLMs on health tasks. [56] introduced a benchmark that combines 6 existing medical question-answering datasets spanning professional medicine, research, consumer queries, and a new dataset of medical questions searched online. In the field of ASD, LLMs have effectively aided in diagnosis [28], treatment [14], and daily life management [30]. In [28], researchers used LLMs to perform binary classification on the scores of ASD adults executing the ADOS-2 A4 task, achieving an accuracy rate of 81.82%. They also used LLMs to analyze 10 social language impairment features of ASD.

Our study utilized the complete range of ADOS-2 clinical voice data and advanced the individual scoring of ADOS-2 through In-context Enhancement, Interpretability Augmentation, and Adaptive Fusion Method to propose the framework **ADOS-Copilot**. Our method can reduce healthcare costs, including patient testing fees and physician training expenses, and minimize the potential for subjectivity in human diagnoses. The explainability of LLMs also contributes to the objectivity of results, aiding patients in understanding their condition and facilitating targeted treatment. However, it is important to acknowledge that the use of LLMs raises privacy and ethical concerns. Ensuring adherence to relevant ethical codes by LLM providers and further testing voice-based de-identification methods is essential. Furthermore, increased reliance on technology by doctors may lead to a higher risk of misdiagnosis.

# 3 Methodology

Our proposed framework is shown in Figure 1. In concisely, we obtained scores for the 8 items of the ADOS-2 Module 3(M3) language section using the **In-context Enhancement** and **Adaptive Fusion** method. Then, we utilized **Interpretability Augmentation** to generate more detailed explanations. It is worth mentioning that ADOS-2 M3 has a total of 14 clinical diagnostic items, with the excluded items being purely non-verbal assessments.

# 3.1 In-context enhancement prompt

We introduced the technique of **in-context enhancement** to address the challenge of current LLMs being unable to utilize generic prompt templates for solving task of ADOS-2 scoring and explanation in real clinical scenarios.

In-context Enhancement Prompt details are shown in Appendix D.2 with the Figure 4. We aim to design such a prompt that enables the model to focus on scoring while providing corresponding justifications simultaneously called **Scoring&Explanation**(zero-shot). We design such prompt based on the following assumptions: reasoning is aimed at better scoring, and scoring is aimed at better reasoning. This assumption aligns with intuition: if the task focuses solely on scoring without providing justifications, the model may resort to tricks, become lazy in thinking, and consequently yield poorer scoring results. Conversely, if the emphasis is solely on reasoning without scoring, there won't be a reasonable standard for the reasoning process, resulting in lower interpretability of the justifications provided [65]. It is worth mentioning that in Figure 4, the -(optional) prefix text can

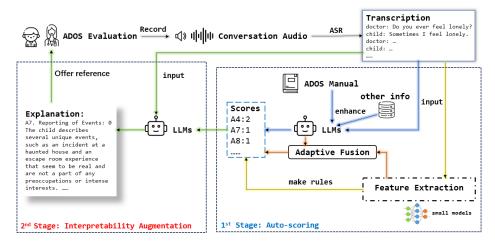


Figure 1: Overall pipeline of our framework. Firstly, we preprocess and transcribe the complete long dialogue text based on the original recordings from clinical ASD diagnoses. Then, we design prompts for In-context Enhancement, using the long text dialogue as input for LLMs to generate scores and justifications for the eight items used in the ADOS clinical diagnoses: A4, A7, A8, B4, B7, B9, B10, and B11. Furthermore, we also feed the long dialogue text into a rule-based method to extract features and then integrated with the results from LLMs, combining the strengths of both approaches to achieve the best scoring performance. Finally, in order to delve deeper into the meaning behind the scores and assist in clinical decision-making, we further input the output fusion results and the original long dialogue text into LLMs to obtain more detailed textual support and additional explanatory evaluations via Interpretability Augmentation in second stage.

be replaced with more settings called Only-Scoring and Scoring&Explanation(few-shot). More zero or few shot setting detail and experiment will be discussed in Appendix I.

By leveraging the contextual learning ability [10], LLMs are capable of completing the various task without updating parameters. Therefore, we introduce the criteria and procedures of the ADOS-2 as well as the prior statistical infomation as in-context enhancement prompts. For criteria, some researches have shown that LLMs tend to make more errors when confronted with more complex conversational contexts [8, 39, 64]. Therefore, we introduce more detailed criteria to guide LLMs in properly activating existing clinical knowledge to better understand this complex task. Inspired by [50], we incorporate the scoring criteria of the eight items used in clinical ASD diagnosis into the system prompt. For procedures, we consider that the evaluation of ADOS-2 consists of a series of consecutive scenario-based interactions. However, it is challenging to segment the individual scenes during the process of speech recognition. In order to facilitate the model's understanding of the evaluation scenarios in ADOS-2, we include excerpts from the assessment process manual used as a reference in hospitals as part of the prompt. For prior information, to better align with the assessment results of doctors and mitigate potential biases inherent in LLMs, we introduce prior statistical information into the prompt. The statistics are collected by hospital, including the mean of each item and the proportion of ASD and TD children.

# 3.2 Adaptive integrating LLMs with rule-based model for enhanced ASD diagnosis

This section introduces an advanced, hybrid methodology that merges the data-driven rule-based model, aligning with ADOS-2 criteria, with the adaptability of LLMs. Our aim is to establish a more comprehensive and reliable diagnostic system for Autism Spectrum Disorder (ASD), addressing the lack of objective and quantitative tools in current ASD diagnosis.

We presents a data-driven rule-based model, addressing the need for objective, quantitative diagnosis tools. It extracts tailored features from child's communication and social interactions linked to ADOS-2 and clinical guidelines for quantifying [6, 47]. Subsequently, rule-based scoring is applied, translating these features into quantitative assessments for individual ADOS-2 sub-items. Readers are encouraged to refer to Appendix E, which provides supplementary information on the design frameworks underlying our rule-based scoring mechanism.

To augment the robustness and validity of scoring framework, we integrate LLMs with rule-based model through an adaptive fusion strategy. This pioneering fusion recognizes the complementary strengths of LLMs and rule-based systems, optimizing performance of our scoring system. Central to this integration is a adaptive weighting schema, which assigns variable importance to each model's predictions based on Mean Absolute Error (MAE) assessments. Consequently, the model that demonstrates superior performance in real-world scenarios is given greater influence, enhancing overall accuracy and adaptability to diverse data intricacies and contexts.

In conclusion, we presents an approach directly addressing the challenge of enhancing prediction reliability across diverse scenarios where traditional single model fall short, through a data-driven balancing act between LLMs and rule-based systems. The fused prediction score  $\hat{y}_i$  is expressed as:

$$\hat{y}_i = \alpha_{LLMi} \cdot f_{LLMi}(x) + \alpha_{Rulei} \cdot f_{Rulei}(x) . \tag{1}$$

Here,  $\alpha_{LLM}$  and  $\alpha_{Rule}$  denote the adaptive weights allocated to the LLM and rule-based predictions respectively for the different ADOS-2 items,  $f_{LLM}{}_i(x)$  and  $f_{Rule}{}_i(x)$  denote the predictions of the LLM and rule-based. Specifically, the calculation of these coefficients is expressed as:

$$\alpha_{LLMi} = \frac{\exp(-MAE_{LLMi})}{\exp(-MAE_{LLMi}) + \exp(-MAE_{Rulei})}, \quad \alpha_{Rulei} = 1 - \alpha_{LLMi}.$$
 (2)

# 3.3 Interpretability augmentation via the second stage

Based on our experimental results, we find that under the prompt setting of Scoring&Explanation, all LLMs perform better in terms of scoring results, including MAE and classification evaluation metrics, in the zero-shot setting compared to the few-shot or Only-Scoring setting. This suggests that the few-shot setting may primarily provide a format guideline and allow LLMs to extract more task-related information from the format guidelines [63], thereby appearing a more professional illusion. However, we aim to provide more reasonable justifications and evidence while ensuring accurate scoring. Therefore, we need a method that can achieve a balance between the two objectives. So partially inspired by [42], we introduce a second stage to enhance the interpretability: in the first stage, we utilize the zero-shot setting to perform the initial scoring and evidence search, aiming to obtain the best scores. In the second stage, we use the scores obtained from the first stage and the original dialogue text as the user prompt. Based on this prompt, we employ a reasoning chain to guide the LLMs to directly extract the most relevant excerpts from the original dialogue text that correspond to the scores and generate the output accordingly. This two-stage approach allows us to achieve both accurate scoring and provide more relevant justifications based on the original dialogue text. The prompt's detail using for second stage is shown in Appendix D.3 with the Figure 6.

# 4 Experiments and results

#### 4.1 Dataset

The dataset utilized in this study was recorded in a clinical context, consisting of 28 audio samples of the whole ADOS-2 assessment processes. We then transcribe the audio into text. Written informed consent was obtained from the parents or caregivers of all participants prior to the assessment. See more details in Appendix A.

#### 4.2 Experimental setup

**LLMs** We explore a wide span of LLMs in our experiment, including GPT-4-turbo [2], Gemini1.5-Pro [52], Claude3, Llama3-8b [5], Mixtral [31, 32], Qwen1.5 [7], Glm [18], Yi-34b [66], Kimi. It is worth noting that our data is derived from a Chinese dataset. Therefore, we have opted to use more LLMs that have been trained on a larger proportion of Chinese pretraining data. Additionally, some models were not trained using Chinese word embeddings, consequently, during experimentation, we will translate both the data and prompts into English when using them, which may lead to inevitable accuracy loss. All LLMs use the default temperature set by their creators.

**Metrics** We use MAE to measure the performance of each 8 items (A4,A7,A8,B4,B7,B9,B10,B11) like many previous works [13], since the magnitude of these scores reflect the degree that the item

Table 2: In-context enhancement ablation result in Qwen1.5-32b, where *Concise* refers to a simplified version of the criteria, retaining only the names of the scoring items and the scoring ranges while excluding specific scoring details (Appendix D.1 for detailed comparison), *C* refers to the standard criteria, *M* refer to the ADOS-2-M3 clinical diagnostic procedures and *S* refers to the prior statistical information.

prompt	A4	A7	A8	B4	B7	В9	B10	B11	avg	2-acc	2-precision	2-f1	3-acc	3-precision	3-f1
Random	0.8934	0.7739	0.8566	0.9879	0.8217	0.8095	0.8568	0.8813	0.8601	0.5927	0.6451	0.5167	0.4493	0.5534	0.4037
Concise	0.9643	0.4286	0.5714	0.8214	0.6071	0.4643	0.6429	0.7500	0.6266	0.5714	0.5714	0.5429	0.4286	0.5533	0.4667
C	0.6071	0.4643	0.5714	0.5357	0.4643	0.4286	0.5714	0.6429	0.5227	0.6071	0.7734	0.5158	0.5000	0.7384	0.4556
C+M	0.7143	0.6429	0.6071	0.7500	0.5000	0.4286	0.6071	0.6429	0.5747	0.6786	0.7991	0.6306	0.5000	0.7173	0.5107
C+S	0.7857	0.3214	0.5714	0.2857	0.4286	0.6071	0.5000	0.6786	0.4870	0.7500	0.7500	0.7490	0.6071	0.6625	0.6336
C+M+S	1.0000	0.4643	0.4643	0.2857	0.5000	0.4643	0.4643	0.6429	0.4805	0.7500	0.7813	0.7381	0.6786	0.7753	0.6969

aims to evaluate. We also calculate the mean MAE of the 8 items to reflect the overall performances. Besides, accounting that the ultimate goal of ADOS-2 is diagnosis, some classification metrics are included. The classification results are calculated using the items scored by model combined with 6 remianing items directly copied from doctor's scoring. The classification task can be either binary (Autism Spectrum Disorder and Non-Spectrum Disorder) or ternary (Autism, Autism Spectrum Disorder, and Non-Spectrum Disorder). The scoring criteria align with clinical judgment, where the overall judgment is based on the sum of scores from the 14 items for classification.

#### 4.3 Results

Validation for in-context enhancement prompt strategy Based on the excellent performance of the Qwen1.5 series, we conducted detailed prompt ablation experiments on the 32b model shown in Table 2. The results showed that the prompt consisting of standard ADOS-2 scoring criteria, the diagnostic procedures of ADOS-2, and prior information(C+M+S) which is described in Figure 4 yielded the best results. This prompt design effectively maximized the potential of LLMs and enabled them to successfully tackle this complex task. Compared to the concise setting, the results of the in-context enhancement prompt show a decrease of 14.61% in average MAE across the eight items. Additionally, there is an increase of 17.86% in binary classification accuracy, 19.52% in F1 score, 25.00% in ternary classification accuracy, and 23.02% in F1 score which clearly demonstrates that utilizing in-context enhancement prompts can effectively tap into the potential of LLMs in this task.

Additionally, we observed that different prompt settings can indirectly and significantly impact the accuracy of certain item, thereby influencing the MAE and classification. For standard criteria, it is crucial for LLMs to include it to understand more objectives and justifications of each scoring criterion, allowing for more objective and professional scoring. We can observe significant improvements in all performance metrics when switching from concise to standard criteria. For the knowledge of ADOS-2 procedure, it could help models to understand the scenario of the conversation with incorporating domain-specific knowledge into the models. For the prior statistical information, it is clearly helpful to lowering B4's MAE. The statistics of ADOS-2 scoring could guide LLMs to give more suitable scores. For example, if mean of an item is pretty low, the model will be more hesitant to give a score of 2. In this sense, the prior information could help to align models' pattern to humans'.

**Validation for rule-based model and fusion strategy** By scrutinizing the entries within the 'rule' row of Table 3, it becomes evident that the rule-based model excels in the majority of the evaluated criteria, thereby attesting to its proficient performance. Of particular note, items A4 and B4 yield MAE scores that are less than LLM's. This notable accomplishment can be attributed to the meticulous quantification of ADOS-2 alongside the adept utilization of sentiment analysis techniques, underscoring the efficacy of these methodologies in enhancing predictive accuracy and precision.

Table 3 showcases the efficacy of combining LLMs with rule-based systems. The Fusion model, by merging the two, achieves superior outcomes: compared to the LLMs, fusion model show a decrease of 2.38% in average MAE across the eight items. In addition, there is an increase of 5.00% in ternary classification accuracy, This evidently illustrates that employing a fusion strategy can efficiently harness the capabilities of both LLMs and rule-based models for this particular task. Thus, Fusion models improve overall performance and robustness, affirming the merit in hybrid solutions. For more detailed results of rule-based model and fusion strategy, please refer to Appendix G.

Table 3: Comparative result of LLM, rule-based system, and their fusion. Among these, LLM refers to the LLM with the best performance, qwen1.5-72b; Rule denotes the rule-based model; and Fusion represents the model that integrates both.

model	A4	A7	A8	B4	В7	В9	B10	B11	avg	2-acc	2-precision	2-f1	3-acc	3-precision	3-f1
LLM	0.8214	0.4643	0.5000	0.1250	0.3214	0.6071	0.6071	0.6786	0.4756	0.8214	0.8345	0.8179	0.7143	0.8627	0.7451
Rule	0.6938	0.5137	0.5190	0.1178	0.3245	0.6809	0.7196	0.8252	0.4982	0.7903	0.8271	0.7887	0.6854	0.7426	0.6833
Fusion	0.6885	0.5165	0.4823	0.1877	0.3036	0.5383	0.6240	0.6951	0.4643	0.8214	0.8661	0.8129	0.7500	0.8626	0.7837
Random	0.8934	0.7739	0.8566	0.9879	0.8217	0.8095	0.8568	0.8813	0.8601	0.5927	0.6451	0.5167	0.4493	0.5534	0.4037

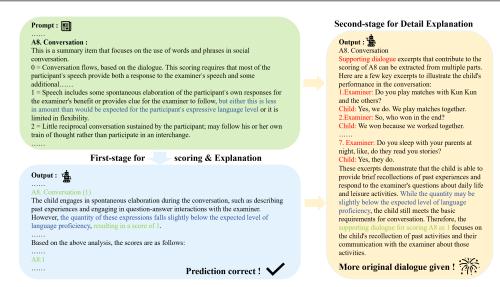


Figure 2: Case study for our framework generated by Qwen1.5-32b. The left part refers to the Scoring&Explanation Stage, where the explanations are relatively general. The right part refers to the Interpretability Augmentation Stage, which include truncated segments from the original dialogue that support the scoring decisions made by our framework. Upon analyzing the output text, the blue part refers to the references to the ADOS-2-M3 scoring criteria, the green part signifies scoring consistency, and the red part indicates several original dialogue segments that support the scoring decisions made by our framework.

Validation for interpretability augmentation via case study From the result of our experiment, we can observe that in the first stage, due to the constraint of tokens and the dual task of scoring and explanation, the explanations provided are often general summaries rather than specific dialogue excerpts. For instance, in Figure 2, when scoring the A8 dialogue item, LLMs provide a general summary stating, "Describes past experiences and engages with the doctor's questions, with a lower-than-expected level of expressive language." This leads to a score of 1. However, using Interpretability Augmentation in the second stage, several brief dialogue excerpts are extracted from the original dialogue to support this assertion, accompanied by more detailed explanations. For example, in Figure 2, seven dialogue excerpts related to daily life and leisure activities are provided from the original dialogue to support the before assertion. Furthermore, after providing the original excerpts, ADOS-Copilot responded to the detailed scoring criteria from the first stage and reaffirmed the scores from the initial phase. Therefore, our framework can achieve consistency and accuracy in reasoning while providing accurate scoring and detailed explanations. More case studies on our framework can be found in the Appendix F.

### 5 Discussion

# 5.1 Further experiments conducted on the LLMs baseline

To validate the effectiveness and generalizability of proposed framework, We conducted extensive comparative experiments on a wide range of LLMs. Due to limitations in handling long contexts, not all famous LLMs such as llama-2 [59] were selected. From Table 4, it can be observed that the Qwen1.5 series of models perform SOTA. The 72b model achieved an average MAE of 0.4805,

Table 4: LLMs baseline Result. All LLMs using the prompt mentioned by Figure 4. The table below presents the single MAE and average MAE (avg) for eight items, as well as the comprehensive metrics for binary and ternary classifications based on ADOS-2 M3 clinical diagnostic criteria.

model	A4	A7	A8	B4	B7	В9	B10	B11	avg	2-acc	2-precision	2-f1	3-acc	3-precision	3-f1
Yi-34b	0.6071	0.7143	0.4643	0.4286	0.5893	0.6964	0.5179	0.7500	0.5860	0.5714	0.5681	0.5577	0.4286	0.4748	0.4420
glm4	0.9643	0.6429	0.3929	0.1071	0.6071	0.6429	0.6429	0.7857	0.5649	0.6071	0.6571	0.5890	0.5357	0.6482	0.5190
glm-3-turbo	0.8214	0.5714	0.5714	0.4643	0.6786	0.7143	0.7500	0.7857	0.6071	0.6429	0.6494	0.6429	0.5714	0.6468	0.5815
kimi	0.5357	0.3571	0.6071	0.9649	0.4286	0.3929	0.6071	0.6071	0.5487	0.5357	0.2870	0.3738	0.4286	0.2571	0.3187
qwen-max	0.7500	0.2142	0.8571	0.8214	0.6429	0.7500	0.9286	0.5000	0.6396	0.6071	0.6607	0.5485	0.4643	0.5665	0.4348
qwen1.5-110b	0.7143	0.4286	0.5715	0.6429	0.4643	0.3214	0.5179	0.6071	0.5179	0.6071	0.7734	0.5158	0.5357	0.7709	0.5262
qwen1.5-72b	0.8571	0.4643	0.7143	0.2857	0.3929	0.5714	0.5000	0.6071	0.4968	0.7857	0.8470	0.7714	0.7143	0.8224	0.7204
qwen1.5-32b	1.0000	0.4643	0.4643	0.2857	0.5000	0.4643	0.4643	0.6429	0.4805	0.7500	0.7813	0.7381	0.6786	0.7753	0.6969
qwen1.5-14b	0.9286	0.7321	0.6786	0.1786	0.4643	0.5400	0.6354	0.6429	0.5533	0.7500	0.7526	0.7503	0.5714	0.7530	0.5943
gpt4	0.8214	0.7500	0.8929	1.1429	0.7143	0.4643	0.7143	0.6429	0.7013	0.6071	0.7733	0.5158	0.4643	0.6827	0.3957
gemini-1.5-pro	0.6429	0.2857	0.5714	1.2857	0.4643	0.3571	0.5714	0.5357	0.5649	0.5714	0.7619	0.4490	0.4286	0.6607	0.3282
claude-3-opus	0.5714	0.3929	0.8929	0.6071	0.6786	0.5714	0.8214	0.8214	0.6558	0.5000	0.4333	0.4063	0.3929	0.3615	0.3199
mixtral-8x22b	0.6071	0.6429	0.7143	0.5714	0.7143	0.8571	0.9643	1.0000	0.6786	0.6786	0.6778	0.6773	0.5715	0.5744	0.5720
mixtral-8x7b	0.6429	0.5357	0.6786	0.4286	0.6071	0.4286	0.6786	0.7143	0.5649	0.6429	0.6633	0.6190	0.5000	0.6094	0.5274
mistral-7b	0.8214	0.7857	0.6071	0.7500	0.5714	0.6071	0.5714	0.7143	0.6526	0.5357	0.5298	0.5265	0.4286	0.4821	0.4519
llama-3-8b	0.7500	0.6429	0.6786	0.6429	0.6429	0.7857	0.7500	0.6429	0.6494	0.5714	0.7619	0.4490	0.4643	0.6893	0.3600
random	0.8934	0.7739	0.8566	0.9879	0.8217	0.8095	0.8568	0.8813	0.8601	0.5927	0.6451	0.5167	0.4493	0.5534	0.4037

Table 5: Average MAE of all LLMs for single scoring criteria. The blue font refers to the item LLMs prefer and the red font refers to the item LLMs struggle.

LLMs	A4	A7	A8	B4	В7	В9	B10	B11
avg-MAE	0.7584	0.5557	0.6492	0.6175	0.5872	0.5748	0.6697	0.6912

while the 72b model excelled in classification metrics. It achieved an accuracy of 78.57%, precision of 84.70%, and an F1 score of 77.14% for binary classification. For the ternary classification, it achieved an accuracy of 71.43%, precision of 82.24%, and an F1 score of 72.04%. Indeed, it is surprising that GPT-4 did not achieve better results in our task compared to the renowned Qwen1.5 series. Additionally, some models even performed worse than random in certain aspects, which indicates that the complexity of our task surpasses the models'ability to comprehend the context. Despite using methods like in-context enhancement, the inherent limitations in their foundational capabilities cannot be compensated. So what effects the evaluation made by LLMs? We will delve into this question in the subsequent section.

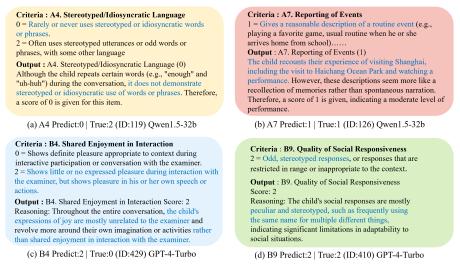


Figure 3: Case study for LLMs'preference Analysis. We provided the scoring items of input and corresponding outputs generated in first stage of our framework. The blue font in the figure indicates the basis for scoring.

#### 5.2 What effects the evaluation made by LLMs?

In this section, we will systematically explain the factors that contribute to the biases of LLMs towards various scoring items. We have calculated the average MAE results for the eight items from all experimental models of its SOTA settings shown in Table 5. We can observe that LLMs have higher average MAE in the A4, A8, B10, and B11 items, indicating a larger deviation from the clinical diagnoses made by doctors. On the other hand, LLMs show smaller average MAE deviations in the A7, B4, B7, and B9 items, aligning more closely with the results provided by doctors. Based on the ADOS-2 criteria details for the eight items, we can conduct that LLMs demonstrate effectiveness in providing scores for certain items primarily due to two characteristics: Firstly, the ADOS scoring criteria for these items have clear descriptions, and there are evident differences between the scores. Such task descriptions serve as effective prompts to guide LLMs' performance. Secondly, these items are mainly based on objective factors that are less influenced by non-verbal elements.

In Figure 3, we can observe how LLMs perform scoring: in Figure 3(a), A4 involves stereotyped and idiosyncratic use of words or phrases, which typically extends beyond just the text and includes aspects of speech and intonation. LLMs may not capture the stereotyped usage solely from the text, leading to an explanation of why they give a score of 0, whereas doctors would consider other factors to cause the bias. In Figure 3(b), A7 pertains to the reporting of events, and the scoring criteria provide detailed explanations and examples. LLMs, therefore, pay particular attention to the events mentioned in the conversation and provide a relatively reliable score for the overall conversation quality. In Figure 3(c), B4 involves shared enjoyment in interaction, which is a highly subjective experience that relies on the doctor's expertise and real-time experience. It also encompasses nonverbal aspects. LLMs can only assess the child's expression of joy based on the text and may perceive it as one-sided rather than shared enjoyment with the doctor, leading to a score of 2. However, the doctor's on-site experience could be completely different and might not align with the LLM's assessment. In Figure 3(d), B9 pertains to the quality of social responsiveness, and its description is concise, impactful, and exhibits significant differences between scores. So LLMs can summarize from the conversation that "the child's social responses are mostly peculiar and stereotyped" and provide an example of "frequently using the same name for multiple different things." Finally, LLMs give a score of 2 based on the evidence of ADOS criteria that "the child demonstrates obvious limitations in adapting to social situations". The explanation is logical, well-supported, and persuasive. Based on the aforementioned analysis, we suggest that the scoring criteria for each assessment item in the ADOS-2 manual should be described in a more detailed and clear manner. It would be highly beneficial to include subjective examples that can help enhance understanding. It is worth mentioning that there is significant variation in the bias of the B4 item across different LLMs. Models such as Qwen1.5 and glm4 scoring more accurately, while models like gpt-4 and Gemini1.5 perform poorly. More detailed discussion of eight item can be found in Appendix H.

# 6 Conclusion

In conclusion, this paper explores the potential of LLMs in ASD diagnosis, addressing the limitations of current methods and computer-assisted tools. We propose an evaluation framework called ADOS-Copilot, to enhance LLMs' performance in real-world diagnostic scenarios. Experimental results show that ADOS-Copilot is competitive with doctors' diagnoses, providing evidence-based assessments. Also our findings have significant implications for ASD diagnosis and the broader field of mental health disorders.

**Limitations.** To facilitate future work, we point out several limitations of our framework. First, it can only be improved by modifying the prompts, unlike doctors who can learn and adapt from their experiences. The framework's ability to comprehend and reason with long contexts may not translate to a deep understanding of children's developmental levels, crucial for accurate ASD assessment. The quality of ASR impacts the later processing of diagnostic data, and ASR transcriptions can lose information such as tone and emotion, important for ASD diagnosis. Lastly and most importantly, the framework only uses textual data from ASR transcriptions, but ASD diagnosis is a multi-modal task, involving the assessment of a child's behavior, social interaction, and communication.

# References

- [1] Mahyar Abbasian, Zhongqi Yang, Elahe Khatibi, Pengfei Zhang, Nitish Nagesh, Iman Azimi, Ramesh Jain, and Amir M Rahmani. Knowledge-infused llm-powered conversational health agent: A case study for diabetes patients. *arXiv preprint arXiv:2402.10153*, 2024.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [3] J Adilakshmi, G Vinoda Reddy, Krishan Dev Nidumolu, Renzon Daniel Cosme Pecho, and M Jahir Pasha. A medical diagnosis system based on explainable artificial intelligence: autism spectrum disorder diagnosis. *International Journal of Intelligent Systems and Applications in Engineering*, 11(6s):385–402, 2023.
- [4] Yeojin A Ahn, Jacquelyn M Moffitt, Yudong Tao, Stephanie Custode, Meaghan Parlade, Amy Beaumont, Sandra Cardona, Melissa Hale, Jennifer Durocher, Michael Alessandri, et al. Objective measurement of social gaze and smile behaviors in children with suspected autism spectrum disorder during administration of the autism diagnostic observation schedule. *Journal of autism and developmental disorders*, pages 1–14, 2023.
- [5] AI@Meta. Llama 3 model card. 2024.
- [6] B Ashwini, Vrinda Narayan, and Jainendra Shukla. Spasht: Semantic and pragmatic speech features for automatic assessment of autism. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [8] Sourangshu Bhattacharya, Avishek Anand, et al. In-context ability transfer for question decomposition in complex qa. *arXiv preprint arXiv:2310.18371*, 2023.
- [9] William J Bosl, Helen Tager-Flusberg, and Charles A Nelson. Eeg analytics for early detection of autism spectrum disorder: a data-driven approach. *Scientific reports*, 8(1):6828, 2018.
- [10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.
- [11] Junya Chen, Matthew Engelhard, Ricardo Henao, Samuel Berchuck, Brian Eichner, Eliana M Perrin, Guillermo Sapiro, and Geraldine Dawson. Enhancing early autism prediction based on electronic records using clinical narratives. *Journal of Biomedical Informatics*, 144:104390, 2023.
- [12] Shi Chen and Qi Zhao. Attention-based autism spectrum disorder screening with privileged modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1181–1190, 2019.
- [13] Ming Cheng, Yingying Zhang, Yixiang Xie, Yueran Pan, Xiao Li, Wenxing Liu, Chengyan Yu, Dong Zhang, Yu Xing, Xiaoqian Huang, et al. Computer-aided autism spectrum disorder diagnosis with behavior signal processing. *IEEE Transactions on Affective Computing*, 2023.
- [14] Yujin Cho, Mingeon Kim, Seojin Kim, Oyun Kwon, Ryan Donghan Kwon, Yoonha Lee, and Dohyun Lim. Evaluating the efficacy of interactive language therapy based on llm for highfunctioning autistic adolescent psychological counseling. arXiv preprint arXiv:2311.09243, 2023.
- [15] Ismail Dergaa, Feten Fekih-Romdhane, Souheil Hallit, Alexandre Andrade Loch, Jordan M Glenn, Mohamed Saifeddin Fessi, Mohamed Ben Aissa, Nizar Souissi, Noomen Guelmami, Sarya Swed, et al. Chatgpt is not ready yet for use in providing mental health assessment and interventions. *Frontiers in Psychiatry*, 14:1277756, 2024.
- [16] Hanna Drimalla, Tobias Scheffer, Niels Landwehr, Irina Baskow, Stefan Roepke, Behnoush Behnia, and Isabel Dziobek. Towards the automatic detection of social biomarkers in autism spectrum disorder: Introducing the simulated interaction task (sit). *NPJ digital medicine*, 3(1):25, 2020.
- [17] Jiayu Du, Jinpeng Li, Guoguo Chen, and Wei-Qiang Zhang. Speechcolab leaderboard: An open-source platform for automatic speech recognition evaluation, 2024.

- [18] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.
- [19] Zohar Elyoseph, Inbar Levkovich, and Shiri Shinan-Altman. Assessing prognosis in depression: comparing perspectives of ai models, mental health professionals and the general public. *Family Medicine and Community Health*, 12(Suppl 1), 2024.
- [20] Yi Fang, Huiyu Duan, Fangyu Shi, Xiongkuo Min, and Guangtao Zhai. Identifying children with autism spectrum disorder based on gaze-following. In 2020 IEEE International Conference on Image Processing (ICIP), pages 423–427. IEEE, 2020.
- [21] Thomas W Frazier, Eric W Klingemier, Sumit Parikh, Leslie Speer, Mark S Strauss, Charis Eng, Antonio Y Hardan, and Eric A Youngstrom. Development and validation of objective and quantitative eye tracking- based measures of autism risk and symptom levels. *Journal of the American Academy of Child & Adolescent Psychiatry*, 57(11):858–866, 2018.
- [22] Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *INTERSPEECH*, 2022.
- [23] Jacob D Gronemeyer, Mikayla Haas, Heba Elsayed, Michael Dacanay, and Helen Genova. 20 using automated sentiment analysis to examine self-evaluation in youth with autism spectrum disorder. *Journal of the International Neuropsychological Society*, 29(s1):629–629, 2023.
- [24] Zhijun Guo, Alvina Lai, Johan Hilge Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. Large language model for mental health: A systematic review. arXiv preprint arXiv:2403.15401, 2024.
- [25] Junxia Han, Guoqian Jiang, Gaoxiang Ouyang, and Xiaoli Li. A multimodal approach for identifying autism spectrum disorders in children. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2003–2011, 2022.
- [26] Jennie Hayes, Tamsin Ford, Rose McCabe, and Ginny Russell. Autism diagnosis as a social process. Autism, 26(2):488–498, 2022.
- [27] Tomoya Hirota and Bryan H King. Autism spectrum disorder: A review. Jama, 329(2):157–168, 2023.
- [28] Chuanbo Hu, Wenqi Li, Mindi Ruan, Xiangxu Yu, Lynn K Paul, Shuo Wang, and Xin Li. Exploiting chatgpt for diagnosing autism-associated language disorders and identifying distinct features. *arXiv preprint arXiv:2405.01799*, 2024.
- [29] Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, and Andrew Beam. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*, 2024.
- [30] JiWoong Jang, Sanika Moharana, Patrick Carrington, and Andrew Begel. "it's the only thing i can trust": Envisioning large language model use by autistic workers for communication assistance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024.
- [31] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- [32] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
- [33] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.
- [34] Michelle YoungJin Kim, Junghwan Kim, Bryan Woosung Kim, Kristen Marie Johnson, and Jee-In Kim. Asdclaims: Twitter dataset of claims on autism spectrum disorder. In 2022 IEEE International Conference on Big Data (Big Data), pages 4385–4389. IEEE, 2022.

- [35] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Healthllm: Large language models for health prediction via wearable sensor data. *arXiv* preprint *arXiv*:2401.06866, 2024.
- [36] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213, 2022.
- [37] Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18471–18480, 2024.
- [38] Jicheng Li, Anjana Bhat, and Roghayeh Barmaki. A two-stage multi-modal affect analysis framework for children with autism spectrum disorder. arXiv preprint arXiv:2106.09199, 2021.
- [39] Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353, 2023.
- [40] Sue Lim and Ralf Schmälzle. Artificial intelligence for health message generation: an empirical study using a large language model (llm) and prompt engineering. *Frontiers in Communication*, 8:1129082, 2023.
- [41] Wenxing Liu, Ming Cheng, Yueran Pan, Lynn Yuan, Suxiu Hu, Ming Li, and Songtian Zeng. Assessing the social skills of children with autism spectrum disorder via language-image pretraining models. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 260–271. Springer, 2023.
- [42] Yucong Luo, Mingyue Cheng, Hao Zhang, Junyu Lu, Qi Liu, and Enhong Chen. Unlocking the potential of large language models for explainable recommendations, 2024.
- [43] Heather MacFarlane, Alexandra C Salem, Liu Chen, Meysam Asgari, and Eric Fombonne. Combining voice and language features improves automated autism detection. *Autism Research*, 15(7):1288–1300, 2022.
- [44] Matthew J Maenner. Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2018. MMWR. Surveillance Summaries, 70, 2021.
- [45] Nikita Mehandru, Brenda Y Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J Butte, and Ahmed Alaa. Evaluating large language models as agents in the clinic. *NPJ digital medicine*, 7(1):84, 2024.
- [46] Cynthia A Molloy, Donna S Murray, Rachel Akers, Terry Mitchell, and Patricia Manning-Courtney. Use of the autism diagnostic observation schedule (ados) in a clinical setting. *Autism*, 15(2):143–162, 2011.
- [47] Keiko Ochi, Nobutaka Ono, Keiho Owada, Masaki Kojima, Miho Kuroda, Shigeki Sagayama, and Hidenori Yamasue. Quantification of speech and synchrony in the conversation of adults with autism spectrum disorder. *PloS one*, 14(12):e0225377, 2019.
- [48] Roy H Perlis, Joseph F Goldberg, Michael J Ostacher, and Christopher D Schneck. Clinical decision support for bipolar depression using large language models. *Neuropsychopharmacology*, pages 1–5, 2024.
- [49] Nadine Probol and Margot Mieskes. Autism detection in speech-a survey. *arXiv preprint arXiv:2402.12880*, 2024.
- [50] Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media. arXiv preprint arXiv:2305.05138, 2023.
- [51] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.
- [52] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv* preprint arXiv:2403.05530, 2024.

- [53] Sergio Rubio-Martín, María Teresa García-Ordás, Martín Bayón-Gutiérrez, Natalia Prieto-Fernández, and José Alberto Benítez-Andrades. Enhancing asd detection accuracy: a combined approach of machine learning and deep learning models with natural language processing. Health Information Science and Systems, 12(1):20, 2024.
- [54] Saad Sadiq, Micheal Castellanos, Jacquelyn Moffitt, Mei-Ling Shyu, Lynn Perry, and Daniel Messinger. Deep learning based multimedia data mining for autism spectrum disorder (asd) diagnosis. In 2019 international conference on data mining workshops (ICDMW), pages 847–854. IEEE, 2019.
- [55] Caio Pinheiro Santana, Emerson Assis de Carvalho, Igor Duarte Rodrigues, Guilherme Sousa Bastos, Adler Diniz de Souza, and Lucelmo Lacerda de Brito. rs-fmri and machine learning for asd diagnosis: A systematic review and meta-analysis. *Scientific reports*, 12(1):6030, 2022.
- [56] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [57] Chunyi Song, Shigang Wang, Meimei Chen, Honghua Li, Feiyong Jia, and Yunxiu Zhao. A multimodal discrimination method for the response to name behavior of autistic children based on human pose tracking and head pose estimation. *Displays*, 76:102360, 2023.
- [58] Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. Benchmarking chinese commonsense reasoning of llms: From chinese-specifics to reasoning-memorization correlations, 2024.
- [59] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.
- [60] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022.
- [61] Weijie Wei, Zhi Liu, Lijin Huang, Alexis Nebout, and Olivier Le Meur. Saliency prediction via multi-level features and deep supervision for children with autism spectrum disorder. In 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pages 621–624. IEEE, 2019.
- [62] Chen Xia, Hexu Chen, Xinran Guo, Kuan Li, and Shuai Ren. Identification of asd via graph convolutional network with visual semantic encoding of saccade. In 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 1553–1560. IEEE, 2023.
- [63] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv* preprint arXiv:2111.02080, 2021.
- [64] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32, 2024.
- [65] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. Towards interpretable mental health analysis with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- [66] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv* preprint arXiv:2403.04652, 2024.
- [67] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.
- [68] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*, 2023.
- [69] Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, et al. Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. arXiv preprint arXiv:2311.11797, 2023.

[70] Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, et al. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*, 2023.

# A Dataset

#### A.1 Data collection

The datasets utilized in this study consist of ADOS-2 (Autism Diagnostic Observation Schedule, Second Edition) assessment processes recorded in a clinical context. After obtaining informed consent from parents or guardians, the complete audio recordings of the ADOS-2 assessment procedures were captured, starting from when the child entered the dedicated assessment room until the child left the room at the end of the evaluation. Any irrelevant data unrelated to the actual assessment, such as the initial and final moments, were removed in subsequent data processing stages. We are committed to protecting the privacy of the children with removing any content that involves children's privacy in data pre-processing phrase.

# A.2 Automatic speech recognition(ASR) and speaker diarization

Audio transcription is our first step in the whole pipeline. Our scenario turns out to be a very challenging one for current ASR systems, largely due to children's still developing language abilities and non-standard pronunciation. We test many commercial services(like iflytek) and open-source ASR models(including Whisper [51] and Paraformer [22]), but none of these give very satisfactory transcriptions, among which large speech models' results are more acceptable, just as [17] indicated. Finally, we choose OpenAI's whisper-large-v3 model as our ASR model. Speaker diarization, the process of segmenting audio recordings by speaker labels and aims to answer the question "who spoke when?", is also needed since the conversations in ADOS involves multiple speakers. Readers are encouraged to refer to Appendix B for supplementary details on the utilized ASR models.

# **B** ASR samples

In this section, we provide an example of a speech transcription text that has undergone preprocessing. Table 6 presents a dialogue excerpt between a doctor and a child in the "Friendship, Relationship, and Marriage" task of the ADOS-2-M3 clinical diagnosis. It is observed that there are illogical segments in the dialogue, such as "". However, it is difficult to determine whether these illogical segments stem from the child's diminished social communication abilities associated with Autism Spectrum Disorder (ASD) or if they are a result of inadequate transcription capabilities. Nonetheless, manual transcription methods are extremely expensive, so further exploration of more advanced methods is still necessary.

Table 7 presents a segment of dialogue manually transcribed with speaker diarization. Table 8 showcases the transcription generated by Paraformer, which notably omits several utterances from the child and contains a considerable number of inaccuracies. Table 9 demonstrates the output transcribed by Whisper, a system that, while capable of transcribing the dialogue, fails to identify speakers and also introduces some errors during the transcription process. Lastly, Table 10 highlights a transcription that underwent an initial phase of Whisper followed by post-processing with GPT for error correction and speaker labeling. This approach evidences a superior performance compared to Paraformer, effectively illustrating the enhancement in both accuracy and speaker distinction achievable through such a sequential refinement process.

# C Introduction of ADOS-2

#### C.1 Procedure

The ADOS-2 consists of a series of structured and semi-structured tasks, typically taking 40-60 minutes to administer. During this time, the examiner provides opportunities for the subject to demonstrate social and communication behaviors relevant to an autism diagnosis. Activities are selected from the module that matches the subject's developmental and language level. Module 3, which includes 14 activities (see 12), is suitable for verbally fluent children or young adolescents.

Table 6: A snip of the transcription

Doctor: I see. So, do you have a boyfriend?

Child: Of course I don't have a boyfriend, but Tongtong is my boyfriend

because he is Yuanyuan, and he is Yuanyuan's cousin.

Doctor: So he is your boyfriend.

Child: He is the boyfriend who came down from the mountain.

Doctor: Oh, I see. How do you know he is your boyfriend?

Child: Because Yuanyuan introduced me to him.

Doctor: Oh, I see. Where do you want to live when you grow up?

Child: You always say not to grow old, so I want to live in my hometown.

Doctor: I mean when you grow up.

Child: When I grow up, oh, I have to go to a university with dorm buildings, so I will live in the dormitory. When I'm old, I will live in a nursing home.

Doctor: Who do you want to live with when you grow up?

Child: I definitely want to live in the dormitory with my classmates.

Doctor: Hmm, have you ever thought about getting married when you grow up?

Child: If I don't want to get married, because the feeling of constantly

having babies after marriage is just too bad.

Doctor: Oh, okay. Do you ever feel lonely?

Child: Sometimes I feel lonely, but then I think about some things I want to do, or things I can't let go of, and remember some things I've done before, and then I don't feel lonely anymore.

Doctor: Oh, do you think your peers feel lonely?

Child: They don't, they all have friends.

Doctor: So, do you feel lonely? Child: I don't feel lonely.

Table 7: A snip of the transcription by manual annotation

Doctor: Have you finished putting it together then?

Child: Yes.

Doctor: I think you've done a fantastic job!

Doctor: Look, there's an airplane.

Child: Hmm, why is there a poison symbol here?

Doctor: Oh, this is the poison sign, do you recognize it?

Doctor: Ah, because this is an eco-friendly bag.

Doctor: And actually, this isn't poison.

Doctor: This is a recycling symbol.

Doctor: Where did you see this poison symbol?

Child: I saw it on a barrel.

Doctor: On a barrel?

Child: That barrel was a car, actually.

Child: It was on the cement mixer truck where I saw the poison symbol.

Doctor: On the cement mixer truck?

Doctor: There might have been hazardous materials, right?

Child: Missiles.

Doctor: Would there be missiles on a mixer truck? Child: No. but the mixed cement could be toxic.

Doctor: The mixed cement is toxic?

Doctor: How do you know that?

Doctor: I don't seem to know.

Child: Because it's all marked there.

Table 8: A snip of the transcription by Paraformer[22]

Doctor: Did you listen carefully?

Doctor: I think your assembly is wonderful.

Doctor: Look, there's an airplane.

Doctor: This is a toxic symbol, do you recognize it?

Doctor: It's an eco-bag. This is not about toxicity but a recycling symbol.

Doctor: Where did you see the toxic symbol?

Doctor: On a barrel.

Doctor: On a cement mixer, perhaps carrying dangerous goods.

Child: Missiles.

Doctor: Missiles.

Doctor: Missiles on a mixer truck?

Doctor: No, the produced concrete is poisonous.

Doctor: How do you know?

Doctor: I don't know where exactly, Doctor: But it's all indicated there.

Table 9: A snip of the transcription by Whisper[51]

Are you done assembling it then?

I think you've done a great job.

I see an airplane.

Hmm, why is there a toxic substance here?

Oh, this is the toxic sign, do you know it?

Ah, this is an environmentally friendly bag.

And this actually isn't toxic.

This is a recycling symbol.

Where did you spot this toxic symbol?

Saw it on a bucket.

On a bucket.

That bucket happened to be a car,

Specifically, saw that toxic symbol on the cement mixer.

On the cement mixer.

Could be dangerous goods,

Right?

Missiles.

Would there be missiles on a mixer?

No, but the mixed concrete would be poisonous.

Mixed concrete is poisonous?

How do you know?

I don't seem to know.

Because it's all labeled there.

Table 10: A snip of the transcription by Whisper+GPT refinement [51]

Doctor: Have you completed the assembly? I must say, your work is impressive.

I spotted an airplane there.

Child: Why is there a toxic symbol here, though?

Doctor: Ah, you recognize the toxic sign? That's interesting.

Child: Oh, because this is an eco-bag, and actually, this isn't toxic—it's a

recycling symbol.

Doctor: Where did you first see the toxic symbol?

Child: On a barrel, which was a car—a cement mixer with that toxic symbol. Doctor: On a cement mixer, possibly carrying hazardous materials, yes?

Child: Missiles.

Doctor: Missiles on a mixer?

Child: Not missiles, but the mixed concrete might be toxic. Doctor: Toxic concrete? How did you learn that?

Child: I'm not sure, but everything is clearly labeled.

Table 11: Statistics of our dataset (28 samples in total)

Item	statistics
gender age(months) ADOS score 2-cls 3-cls	boy:20(71.4%) girl:8(28.6%) 93.6±28.2 (min=57,max=173) 7.25±4.56 (min=1,max=17) TD: 12(42.9%) ASD: 16(57.1%) TD: 12(42.9%) ASD: 4(14.3%) Austim: 12(42.9%)

#### C.2 Scoring

The scoring of ADOS does not strictly follow a one-to-one task scoring method. Instead, it evaluates the subject's overall performance throughout the entire assessment process. This is why we have introduced LLMs to handle the complete ADOS-2 assessment process, rather than using only a portion of its content. Here, we will only discuss the specific scoring criteria actually used in this experiment.

#### C.2.1 A4

This item's assessment includes delayed echolalia or other highly repetitive speech patterns with consistent intonation. These words or phrases can be meaningful and can be somewhat applicable to conversation. The focus of this project is the stereotyped or characteristic nature of the phrasing.

0 =Rarely or never uses stereotyped or idiosyncratic words or phrases.

1 = Uses vocabulary or phrases that are more repetitive or formulaic than most people at the same level of expressive language ability, but not noticeably strange, or occasionally uses strange uses of stereotyped discourse or phrases, with substantial spontaneous and flexible language.

Table 12: ADOS-2 Module 3 Activities

Activities	Summary
Construction Task	Provide a warm-up activity to observe the interactive behavior of the subjects.
Make Believe Play	Observe the extent to which the subjects use toys and dolls creatively in an unstructured task.
Joint Interactive Play	Assess the extent and quality of the subjects' coordination of behavior and emotional expression with the assessor.
Demonstration Task	Assess the ability of the subject to use a set of actions accompanied by gestures or language, and report a familiar event.
Description of Picture	Obtain an example of the subject's spontaneous language and communication, and understand what can attract his interest.
Telling a Story from a Book	Assess the ability to describe a sequential story from a picture and comment on social relationships and emotional expression.
Cartoons	Observe how the subject narrates a story, uses gestures to present events,
Conversation and Reporting	Assess the subject's ability to engage in back-and-forth conversation and describe an event or situation without visual cues.
Emotions	The subject details two emotions, their triggers, and related personal experiences.
Social Difficulties and Annoyance	Assess the subject's insight into their own social difficulties and their sense of responsibility for their own behavior.
Break	Provide the subject with a break from the structured social demands of the assessment and observe their behavior in a less structured setting.
Friends and Marriage	The subject details one or more types of relationships and shares their views on friendship, family or a long-term partnership.
Loneliness	Assess the subject's insight into their own social situations and their ability to describe their emotional responses.
Creating a Story	The subject uses objects creatively to tell a new story.

- 2 = Frequently uses stereotyped discourse or strange words or phrases, along with some other language.
- 3 = Almost exclusively uses strange or stereotyped language. And rarely uses non-stereotyped natural language.

#### C.2.2 A7

This item focuses on the participant's ability to spontaneously select an event or respond to a general question from the assessor and describe it in a clear and understandable way, without requiring specific prompting. This involves a sequential description of events that occurred outside of the immediate environment.

- 0 = Reports a specific, real, and unconventional event in a non-stereotyped manner (e.g., a vacation, a trip, a shopping spree). The participant provides this description without being directly asked, but may need an initial general question to get started.
- 1 = Provides a well-structured description of a routine event (e.g., what they did on the way home from school, playing their favorite game). This is not related to their profession or hobbies and is likely a truthful account. The participant offers this description without prompting, but may initially need to be asked to elaborate on the event. Descriptions from the "Demonstration Task" are also included here.
- 2 = Provides a description of a regular or unconventional event, but relies on specific prompting from the assessor, or describes an event that seems unlikely to be true.
- 3 = Response to specific prompting is inconsistent or insufficient.

#### C.2.3 A8

This summary item evaluates the participant's ability to engage in reciprocal conversation. The rating should encompass all conversation opportunities, not just the most successful ones.

- 0 = Fluent conversation builds upon the assessor's contributions. The participant primarily delivers responses that address the assessor's prompts and adds their own thoughts or questions based on the ongoing dialogue. This allows for a back-and-forth exchange with at least four elements: assessor initiates, participant comments, assessor responds, and participant responds.
- 1 = The conversation includes some elaboration by the participant on their responses or provides cues to guide the assessor's direction. However, this elaboration may be limited in quantity or flexibility compared to what's expected at the participant's expressive language level.
- 2 = The participant initiates minimal reciprocal conversation and may prioritize their own train of thought over actively engaging. While some spontaneous information or comments might be present, there's a weak sense of true back-and-forth exchange.
- 3 = The participant demonstrates very little spontaneous communicative language (though there may be a significant amount of repetitive or non-communicative language). This rating applies to participants with limitations who struggle to respond meaningfully to the assessor's conversation openers.

#### C.2.4 B4

This item assesses the participant's ability to express pleasure to the assessor, beyond simply interacting or responding.

- 0 = Exhibits pleasure in at least one task or conversation topic during engagement or interaction with the assessor.
- 1 = Exhibits context-appropriate pleasure during interactions with the assessor or shows clear pleasure in one interaction.
- 2 = Shows little or no expressed pleasure in interactions with the assessor, but may show pleasure in their own speech or behavior or in non-interactive components of the ADOS-2 materials or activities.
- 3 = Expresses very little or no pleasure throughout the ADOS-2 assessment.

# C.2.5 B7

This summary item focuses on the quality of the participant's attempts to initiate social interaction with the assessor, rather than the frequency of such attempts.

- 0 = Effectively uses nonverbal and verbal means to communicate clear social overtures to the assessor. These overtures must be appropriate for the current context.
- 1 = Some social overtures are slightly unusual in nature. Initiations may be limited to personal needs or related to the participant's own interests, but in some cases, may draw the assessor in.
- 2 = Inappropriate overtures; many overtures lack the ability to blend into the context and/or have social quality. This includes the focus of the participant's overtures, with few attempts to draw the assessor in.
- 3 =No social overtures of any kind.

#### C.2.6 B9

Another summary item that evaluates the participant's social responses.

- 0 = Exhibits a range of appropriate responses that vary depending on the current social context and pressure.
- 1 = Responds to most social cues but with some limitations, inappropriateness, inconsistency, or consistent negativity.
- 2 = Strange, stereotyped, or limited-range responses, or responses that are not contextually appropriate.
- 3 = Little or no response to the assessor's attempts to engage the participant.

#### C.2.7 B10

This item focuses on the frequency of interaction using any communication mode during the ADOS-2 assessment. Frequency here is defined by the number of occurrences and distribution across a range of contexts. The rating for this summary project should describe all aspects of nonverbal and verbal/vocal behavior (which do not need to be coordinated), but there must be at least a brief interaction with the assessor (not others who may be present in the ADOS-2 assessment room).

- 0 = Extensive use of verbal or nonverbal behavior (regardless of the level of proficiency) for social interaction (i.e., chatting, commenting, making remarks, or nonverbal behavior that appears to be reciprocal).
- 1 = Some reciprocal social interaction (as described in the 0 rating above), but with a reduced frequency or quantity or the number of contexts in which such interactions occur (regardless of the amount of non-social talk).
- 2 = Most communication is either object-oriented (i.e., asking about things), answering questions, echolalic, or has a specific premise, with little or no social interaction/back-and-forth talk.
- 3 =Very little or no interaction with the assessor.

# C.2.8 B11

This is another summary item that reflects the overall judgment of the rapport that the assessor establishes with the participant during the ADOS-2 assessment. Particular consideration is given to the extent to which the assessor has to change their own behavior in order to successfully maintain the interaction.

- 0 = Comfortable interaction between the participant and the assessor, appropriate for the context of the ADOS-2 assessment.
- 1 = Interaction is sometimes comfortable but not sustained (e.g., may feel awkward or constrained at times, or the participant's behavior may seem mechanical or slightly inappropriate).

# System Prompt ## Role (Task\_Desciption<sub>sys</sub>) You are a doctor with extensive experience in the field of Autism Spectrum Disorder (ASD) diagnosis and are very familiar with the ADOS (Autism Diagnostic Observation Schedule)—a standardized diagnostic test used to assess ASD. - If the user provides the text of a conversation between a child and a doctor during an ADOS assessment, you need to score each item based on the scoring criteria provided afterward. - Although some items involve non-verbal communication, try to score based on the text as much as possible - (optional) You are required to provide specific reasons for the scores, including references to the text. ## Considerations - Only discuss topics related to ASD and ADOS. - Since the conversation text is based on voice recognition, it may not be entirely accurate, and you need to discern carefully. Prompt<sub>criteria</sub>: ADOS Scoring Criteria: \${Criteria} Prompt<sub>m3</sub>: ADOS Assessment Procedure: \${ADOS-m3} Prompt<sub>stat</sub>: ADOS Assessment Scoring Statistics \${Stat\_info} **User Prompt** $(Task\_Description_{user})$ The following is the text of a conversation between a child and a doctor during an ADOS assessment: Transcipt: Please score these 8 items (A4,A7,A8,B4,B7,B9,B10,B11) based on the above scoring criteria and the conversation.

Figure 4: In-context Enhancement prompt for ADOS-Copilot. Where  $Prompt_{criteria}$  refers to the prompt of clinical ADOS criteria,  $Prompt_{m3}$  refers to the prompt of clinical ADOS-2 Module 3 procedures,  $Prompt_{stat}$  refers to the prior information of the ASD and TD children,  $Transcript_i$  refers to the pro-processing dialogue texts between doctor and child.

- 2 = One-sided or unusual interactions lead to a conversation that is persistently mildly uncomfortable, or the conversation would be difficult if the assessor did not repeatedly change this outside of the standard activities in the ADOS-2 assessment.
- 3 = The participant shows very little attention to the assessor, or the conversation is noticeably uncomfortable for an extended period of time.

# C.3 Classification

In clinical practice, children's scores are limited to 0, 1, and 2 (with a score of 3 being merged into 2). For Module 3, the final ADOS-2 M3 score consists of social affect(SA), communication (3 items, 0-6 points) and reciprocal social interaction (7 items, 0-14 points), and restricted and repetitive behaviors(RRB) (4 items, 0-8 points). This results in a total of 14 items and a score range of 0-28 points. The ADOS-2 uses a cutoff method for diagnosing ASD in children. Specifically, for children assessed with M3, a total score of 0-6 indicates a non-spectrum disorder, 7-8 indicates a spectrum disorder, and 9 or above indicates Autism. The scoring items, standards, and cutoff algorithms vary between modules according to the child's developmental level.

# D More details for prompts using by ADOS-Copilot

#### D.1 Concise vs standard criteria

Figure 5 shows a comparison of the concise and standard(C) prompts used in the ablation experiment. It is evident that the concise criteria only include the item name and an overall scoring description, whereas the standard criteria provide a detailed item description and corresponding criteria justifications for each score.

# D.2 Details for prompt of in-context enhancement

As shown in Figure 4, we can see more detail of our in-context enhancement prompt.

#### **Concise Criteria:**

A4. Stereotyped/idiosyncratic Use of Words or Phrases

A7. Reporting of Events

A8. Conversation

B4. Shared Enjoyment in Interaction

B7. Quality of Social Overtures

B9. Quality of Social Response

B10. Amount of Reciprocal Social Communication

B11. Overall Quality of Rapport

Each item is scored as a 0, 1, or 2, with lower scores indicating more typical behavior in children, and higher scores indicating a greater tendency towards autism.

#### Standard Criteria:

A7. Reporting of Events

The focus of this item is on the participant's ability to select an event spontaneously or in response to the examiner's general questioning and to describe it in a comprehensible fashion without requiring specific probes. This should involve a sequential description of an event outside the immediate environment. Evaluate the "best" example according to the given constraints described below.

- 0=Reports a specific nonroutine event (e.g., a holiday, a vacation, a shopping trip) that is not part of any preoccupations or intense interests and seems likely to be real. Gives a reasonable account without specific probes, but may need to be asked a general question to get started.
- 1=Gives a reasonable account of a routine event (e.g., playing a favorite game, usual routine when he or she arrives home from school) that is not part of a preoccupation or intense interest and seems likely to be real. Gives the account without specific probes, but initially may need to be asked to describe the event. Include accounts from the "Demonstration Task" here.

2=Provides an account of routine or nonroutine events, but dependent on specific probes, OR only describes an event that seems unlikely to have been real.

.....

Figure 5: Concise vs Standard Criteria of ADOS-2. In the ablation experiment, we compared the effects of the concise criteria and standard criteria. This Figure provides specific examples of both to better understand the differences in their effects.

Table 13: An excerpt from the scoring description of the ADOS-2 procedures

# 2. Make-Believe Play

Purpose: This activity provides the opportunity to observe the participant's creative or imaginative use of miniature play objects in an unstructured task.

Materials: Contents of Bag 3—two male action figures and a female action figure, with one "prop" for each; miniature hairbrush; two small tools; toy dinosaur. Contents of Bag 2—small spoons and plates, several pieces of miniature food, small teapot/pitcher/measuring cup, miniature book, toy car, toy rocket, small ball, hologram disk, and two pieces of "junk" (a small piece of cloth and a small "jewelry" box). If necessary, other materials from Module 2 may be added (i.e., from Bag 1), but because they require somewhat less creativity, the difference in contexts from the standard presentation should be taken into account when coding.

#### Instructions:

- -Lay out the materials from Bag 3. Introduce the action figures with descriptions appropriate to their appearance (e.g., "Here are a princess, a wrestler, and a soldier, and their pet dinosaur").
- -Lay out the materials from Bag 2, saying, "Here are some of their things. Can you play with these for a while?"
- -Observe the participant's behavior. If the participant does nothing or seems uncomfortable, then after a few moments, say, "I'll play with these."

# D.3 Prompt of second-stage for interpretability augmentation

As shown in Figure 6, We replaced the -(optional) content from our prompt and enhanced the interpretability separately for each individual scoring item. We extract the single-item criteria by separating the items in ADOS Scoring Criteria. Then, we incorporated the scores and justifications generated by LLMs called  $Output_{LLMs}$  in the first stage into the user prompt.

# E Feature and paradigm design of rule-based model

#### Feature design

In clinical practice, clinicians need to score children's performance based on ADOS-2 annotations. However, current research has not adequately addressed the objective and quantitative assessment of these issues[13]. To identify the risk of ASD in children by analyzing their language patterns, we

# (Task\_Desciption<sub>sys</sub>) from base prompt - (optional) Please provide dialogue excerpts based on the text and the corresponding justifications for your judgments. ## Limitations - When providing justifications, please remain faithful to the actual dialogue text and refrain from fabricating or concocting dialogue content. - The excerpts you provide for autism clinical diagnosis are of utmost importance. Please pay more attention and carefully think step by step. | Item\_name\_i | Scoring Criteria : \$\{Single-item-Criteria\} \quad \text{Prompt}\_{m3} : ADOS Assessment Procedure : \( \frac{\{ADOS-m3\}}{\} \)

#### **User Prompt**

Below is the dialogue text between the child and the doctor from a certain ADOS assessment:

#### $Transcipt_{i}$

The following is the overall evaluation result regarding the dialogue content:

#### Output

Based on the evaluation result, scoring criteria, and assessment process, please find the excerpts in the dialogue text that can support the assessment of the *ltem\_name*; item.

Figure 6: Prompt of second-stage for Interpretability Augmentation

propose a set of evaluation features based on ADOS-2 and clinical practices inspired by [46]. These features quantify children's language communication and reciprocal Social interaction. Predefined scoring rules then assign evaluation scores to child participants. The evaluation features are introduced as follows:

**Child immediate echolalia rate** This feature assesses the severity of immediate echolalia in children. We utilize the levenshtein distance [67] to determine whether the child's speech mimics the doctor's utterances.

Conversation alternating rate This feature evaluates the flow of conversation between the child and the doctor. It reflects the smoothness of the conversation based on the frequency of turn-taking between the child and the doctor in the dialogue text.

**Child participant rate** This feature assesses the child's level of participation in the conversation. It measures the proportion of the child's speech in the dialogue text to determine the child's engagement in the conversation.

**Child enjoyment rate** This feature evaluates the child's ability to express enjoyment. By conducting sentiment analysis[23] on the child's speech, we assess the degree of enjoyment during the evaluation based on the proportion of positive utterances by the child.

**Child passive rate** This feature assesses the level of passivity in the child. Through sentiment analysis of the child's speech, we evaluate the degree of child's negative emotions during the assessment based on the proportion of negative utterances by the child.

**Child social suggestion rate** This feature evaluates the frequency with which the child offers social suggestions. We determine the child's tendency to provide social suggestions based on the proportion of suggestions within the total number of their utterances.

**Child social response rate** This feature assesses the frequency of the child's social responses. We measure the child's active social responding by calculating the proportion of the child's responses to the doctor's questions relative to the total number of questions asked.

### Paradigm design

Figure 7 illustrates the scoring rules we devised for various items, informed by the ADOS-2 documentation and clinical practice. To ascertain the threshold parameters for these rules, we employed a stratified two-fold cross-validation approach coupled with grid search. The mean Mean Absolute Error (MAE) on the validation sets was adopted as the primary evaluation metric, ensuring a robust and comprehensive assessment of our model's performance across different subsets of the data. This strategy not only validates the scoring rules' Generality but also optimizes their effectiveness by fine-tuning the parameters against real-world variability captured within the validation folds.

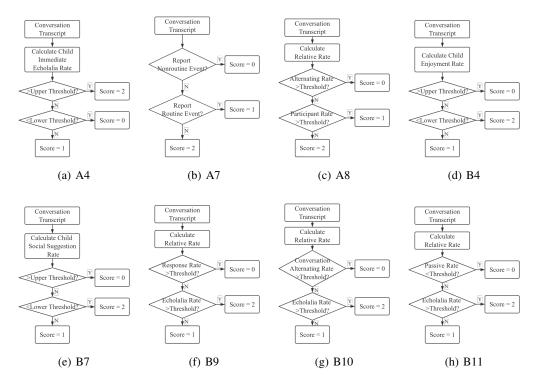


Figure 7: Paradigm design flow chart. Please refer to Appendix C for the scoring criteria of specific items

model A7 A8 B4 В7 B10 B11 2-f1 3-precision 3-f1 A4 avg 2-acc 2-precision 3-acc 0.4643 **0.6071** 0.5398 0.6498 0.5610 **0.7837** Fusion-v1 0.5714 0.5000 0.2500 0.3214 0.6071 0.6786 0.4805 0.6786 0.7278 0.5357 0.6748 0.5152 0.4825 0.3037 0.8129 0.7500 Fusion-v2 0.6859 0.1894 0.6235 0.6947 0.4642 0.8214 0.8661 0.8626 Fusion-v3 0.6755 0.5125 0.4828 0.1912 0.3037 0.5439 0.6220 0.6930 0.4633 0.8214 0.8661 0.8129 0.7500 0.8626 0.7837 0.8661 0.5165 0.4823 0.1877 0.3036 0.5383 0.6240 0.6951 0.8214 0.8129 0.7500 0.8626 0.7837 0.6885 0.4643 Fusion-v4 0.3245 LLM 0.8214 0.4643 0.5000 0.1250 0.3214 0.6071 0.6071 0.6786 0.4756 0.8214 0.8345 0.8179 0.7143 0.8627 0.7451

Table 14: Results of different fusion strategy

# F More details of our proposed framework via case study

Figure 8 illustrates an example of our proposed framework based on Qwen1.5-32b for assessing B9 (Quality of Social Responsiveness). In the Scoring&Explanation stage, ADOS-Copilot determines that the child's social responses still exhibit inconsistency and stereotyped reactions, aligning with the description for the score of 1. In the Interpretability Augmentation stage, two excerpts are extracted from the original dialogue to provide evidence for the reasoning behind the score. These excerpts intuitively demonstrate that the child does respond to the doctor's interaction, but the responses are brief and not entirely appropriate, further validating the assessment in the first stage and reinforcing the objectivity and rationality of assigning a score of 1. This showcases the superiority of our proposed framework in scoring and explanation.

# **G** More analysis of experiment results

# G.1 Detailed results of rule-based and fusion model

The presented experimental data in Table 14 offer insights into the performance of various fusion strategies (Fusion-v1 through Fusion-v4) across multiple items (A4 to B11) and aggregated performance metrics, contrasting these with standalone rule-based and LLM models. The analysis underscores the impact of different weighting schemes on fusion efficacy.

#### ID: 311 Output generated by Interpretability Enhancement: 1. Doctor: There are so many delicious things. Do you want to share with me? Criteria: **B9.** Quality of Social Response Child: Thank you. Child: Then I'll keep it. This is a summary item that focuses Child: It would be nice if you also wanted some. on the participant's social responses - In this excerpt, the child responds to the doctor's sharing request, showing a certain level of social throughout the ADOS -2 evaluation. responsiveness, even though the response may be polite in nature. 0 =Shows a range of appropriate 2.Doctor: You're doing a great job with the puzzle. responses that are varied according to Child: Okay. immediate social situations and Child: Hmm. presses. Child: Blue. 1 = Shows responsiveness to most Doctor: Yes, Auntie has blue ones here too. social contexts, but somewhat limited, Child: I have two more. socially awkward, inappropriate, Doctor: You have two more, and there are two more blue ones here. inconsistent, or consistently negative. Child: I'll give you two blue ones. 2 = Odd, stereotyped responses, or Doctor: I'll give you the blue ones too. responses that are restricted in range Child: Mine is a bit crooked. or inappropriate to the context. Doctor: Yes, because this one might be made of foam. Output generated in Child: Is it like this when it's made of foam? Scoring&Explanation Stage Doctor: Yes. B9. Quality of Social Responsiveness (1 - This excerpt further demonstrates the child's interaction during the puzzle activity, indicating their The child's social responses are ability to follow the conversation topic and provide some level of response, albeit limited in content occasionally appropriate, but at times - These snippets reflect the child's basic social responsiveness in their interaction with the doctor. they also exhibit inconsistent or Despite potential limitations and inconsistencies, they do exhibit fundamental social responses such stereotyped reactions. Therefore, a as answering questions, sharing thoughts, and engaging in simple discussions. These behaviors score of 1 is assigned. align with the score 1 described in B9, which includes "responding to most social situations but with some limitations, inappropriateness, inconsistency, or consistently negative responses."

Figure 8: Case Study for B9 Scoring and Explanation. The blue part refers to the references to the ADOS-2 scoring criteria, the green part signifies scoring consistency, and the red part indicates several original dialogue segments that support the scoring decisions made by our framework.

The rule-based model excels in particular tasks (e.g., B1, B2), as reflected in high accuracy, highlighting its domain-specific precision. However, its modest overall average accuracy points to limitations in generalizability beyond these niches. The LLM demonstrates exceptional performance in certain categories (like A4), yet lags in overall average accuracy and other comprehensive metrics when compared to fused models, indicating a trade-off between specialized accuracy in isolated instances and broader consistency.

Fusion-v1 employs a rudimentary approach, assigning a coefficient of 1 to the model with lower MAE and 0 to the other, effectively silencing one model in the decision-making process. While this straightforward method shows promise in specific categories (such as A4 and A7), its overall average accuracy (avg) and composite metrics lag behind subsequent fusion techniques, suggesting that linear weighting may underutilize the complementary strengths of the two models.

Fusion-v2, v3, and v4 refine the weighting strategy by applying the inverse, square of the inverse, and softmax transformations of MAE values, respectively. These advanced schemes yield improvements across most indicators, with Fusion-v2 and v3 notably excelling in average accuracy, 2-acc, and 2-precision, illustrating the potency of nuanced weight allocation in harnessing the advantages of both models. Fusion-v4, despite adopting a more sophisticated softmax weighting, delivers comparable results, implying that simpler inversions (as in v2 and v3) might suffice for optimal performance in this context.

# **G.2** Effect of language

Apparently in Figure 9, LLMs that specialzes in Chinese perform better than multi-lingual ones in general. This phenomenon is expected, as many papers suggest [58]. Even GPT-4 undergo much disadvantage in our task. We believe that the noise in the transcription adds much difficulty to models with lack of specialization in the Chinese, while LLMs that specialzes in Chinese could still detect the true text in the noisy transcription.

#### **G.3** Effect of scale

We find an interesting phenomenon: medium-sized models perform best in our task Figure 10(a) 10(b). This forms an anti-scaling examples [33]. Why big-sized models fail? We attribute this to the fact that our task requires detecting abnormal through language. Big-sized models could more easily

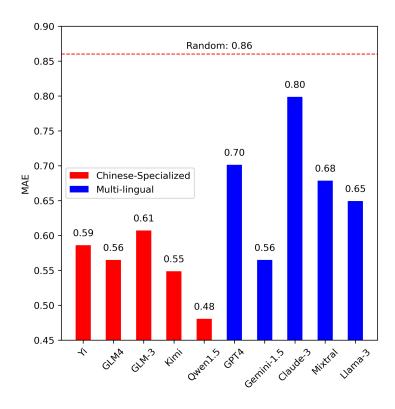


Figure 9: Effect of language. We compare the average MAE of 8 items of 5 Chinese-specialized and 5 multi-lingual famous LLMs.

catch the noise in the transcription thus are misled to give higher scores. As for small-sized models, our task is too challenging for them. Many small-sized models even couldn't understand our task.

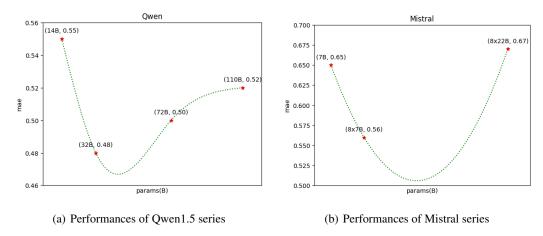


Figure 10: Effect of scale. We compare the average MAE of 8 items of different sizes of two well-known open-source series models, Qwen1.5 and Mistral.

# H More details for what effects LLMs' scoring of the eight ADOS items?

A4 corresponds to the stereotypical/special use of words or phrases, and the experimental results indicate that almost all LLMs perform poorly in this aspect. The reason behind this is that stereotypical/special use heavily relies on subjective judgment and often occurs sporadically in different

#### Criteria: A8. Conversation

0 = This scoring requires that most of the participant's speech provide both a response to the examiner's speech and some additional talking......

2 = Little reciprocal conversation sustained by the participant; may follow his or her own train of thought.....

Output: A8. Conversation Scoring: 2

In conversations, the child's responses often revolve around their own thoughts and interests rather than actively engaging in the dialogue. Although they make attempts to interact with the clinician, the conversation lacks a genuine sense of reciprocity. It frequently centers around their fantasy storylines.

(a) A8 Predict:2 | True:0 (ID:311) Qwen1.5-72b

#### Criteria: B10. Amount of Reciprocal Social Communication

0 = Extensive use of verbal or nonverbal behaviors for social interchange (i.e., chat, comments, remarks, or nonverbal behaviors that appear to have reciprocal intent).

1 = Some reciprocal social communication but reduced in frequency or amount.

Output: B10: 1. The child provides a lot of information during the conversation, but these contributions are often spontaneous rather than responsive to the clinician's speech. Their communication tends to be one-sided narratives rather than genuine reciprocal exchanges.

(c) B10 Predict:1 | True:0 (ID:124) Qwen1.5-72b

#### Criteria: B7. Quality of Social Initiative

0 = Effectively using non-verbal and verbal means to clearly demonstrate social initiative to the evaluator. The initiative must be appropriate to the immediate context.

#### Output

B7. Quality of Social Initiative

Scoring: 0

The child actively engages in games, shares stories, and asks questions during the conversation, demonstrating good social initiative, albeit sometimes centered around their own interests.

(b) A7 Predict:0 | True:0 (ID:328) Qwen1.5-72b

#### Criteria: B11. Overall Communication Quality

2 = One-sided or unusual interaction resulting in a inconsistently mildly uncomfortable session or a session that would have been difficult if the examiner had not continuously modified the structure of the situation beyond the standard activities in the ADOS-2 evaluation.

Output: B11. Overall Communication Quality: 2

Output: B11. Overall Communication Quality: 2
During the conversation, the child's communication tends to be one-sided, requiring the evaluator to constantly adjust to understand their thoughts. Additionally, there are many unusual topics and vocabulary used in the dialogue, making communication challenging. Therefore, a score of 2 is given.

(d) B11 Predict:2 | True:2 (ID:429) Qwen1.5-72b

Figure 11: More case study for LLMs' preference Analysis. The blue font in the figure indicates the basis for scoring.

dialogue contexts. In clinical practice, doctors can often determine stereotypical usage and assign a higher score based on a single interaction fragment and other non-verbal behavior. However, LLMs find it challenging to detect abnormalities from similar fragments just based on text, making it difficult for them to perceive the exception. As a result, LLMs can't align with the clinical doctor leading to this bias.

A7 refers to event reporting, and LLMs perform not bad in this item. The purpose of this item is to identify communication between doctors and children regarding routine/unusual events. On one hand, LLMs need to detect relevant event fragments from the dialogue, and this type of task is widely present in the fine-tuning process of various LLMs. Therefore, LLMs are competent in this aspect. On the other hand, the ADOS scoring criteria for this item are described clearly, with distinct justifications for the three scores and corresponding examples provided. The combination of a clear prompt and LLMs' inherent capabilities enables the successful completion of this task.

A8 corresponds to conversation and focuses on the fluency of social interactions and the level of word and phrase usage. LLMs struggle in this item because the evaluation of word and phrase usage is subjective and relies on the clinical experience of doctors. Moreover, assessing fluency and flexibility is more accurately done at the linguistic level, where aspects such as hesitations and interruptions are more evident in speech and tone than in written text. As shown in Figure 11(a), LLMs perceive that from textual perspective, the child is immersed in their own world and lacks reciprocity during the conversation. However, the clinician's viewpoint is exactly the opposite. Although it is a conversation item, the nonverbal aspects of the clinical setting, such as eye contact or facial expressions, are certainly taken into consideration by the clinician, leading to significant discrepancies in scoring.

**B4** is a highly controversial item that involves the sharing of joy during interactions, and there is a significant variation in MAE among different LLMs in this item. On one hand, B4 involves the expression of happiness during interactions, which is not solely dependent on language. Factors such as body language, tone of voice, and facial expressions all influence the judgment of clinical doctors. LLMs, on the other hand, tend to exhibit more extreme scoring criteria based on the presented text. If there are no explicit scenes or semantic cues related to share enjoyment in the dialogue, the model tends to assign a higher score. However, once phrases like "hahaha, your story is very funny, i am happy." are detected, the model immediately assigns a lower score without considering the whole enjoyment of dialogues. Doctors, on the other hand, tend to have a more lenient and multidimensional approach to scoring this item, leading to the dramatic variation observed in the scores. On the other

hand, The shared enjoyment in interactions can be regarded as a form of emotional analysis task, but it is more complex than traditional sentiment analysis due to its practical application in multi-turn dialogues where participants' emotional states undergo changes. Existing large models excel in simple sentiment analysis tasks, but they are outperformed by fine-tuned smaller models in complex tasks[68]. Therefore, one possible reason for the underperformance of some large models on B4 is their insufficient capability to handle complex scenarios. We believe that part of the reason why some models achieve good results in this task is the utilization of prior statistical information in the prompt. This injection of prior statistics lowers the probabilities of tokens that were originally more likely in the model, causing the model to carefully consider extreme scoring and align its results with the doctor's thought process.

**B7&B9** primarily assess social initiative and response and are summary items. LLMs can provide reasonably objective scores for the overall conversation situation. On one hand, this is attributed to the clear scoring descriptions for each score in the ADOS scoring criteria, with evident differences between the scores. On the other hand, in real assessment scenarios, doctors tend to carefully consider their scores for summary items, resulting in a more compromise scoring. LLMs, when evidence is not abundant, typically exhibit conservative scoring tendencies that align with doctors' preferences. As shown in Figure 11(b), LLMs keenly observed the child's social initiative in games, storytelling, and questioning, leading to a score of 0.

**B10&B11** assess the quality of reciprocal communication and evaluate the overall quality of communication. Unlike B7 and B9, the ADOS scoring criteria for these two items include descriptions of nonverbal behaviors. This makes LLMs less certain about their own scores when they only have text inputs. Additionally, the corresponding descriptions in the ADOS scoring criteria can be abstract, and unless a doctor is experienced and seasoned, they may struggle to fully understand the implications of their evaluations. This difficulty extends even more to LLMs. As shown in Figure 11(c), LLMs perceive that the child's communication is spontaneous and one-sided rather than responsive to the clinician, resulting in fewer instances of reciprocal social communication, warranting a score of 1. However, the clinician's comprehensive observation from nonverbal and behavioral et al. aspects may subjectively lean towards a score of 0. Moreover, the boundary between scores 0 and 1 is not clearly defined, and relying solely on the phrase "reduced in frequency or amount" can lead to confusion for LLMs. In Figure 11 (d), B11 is a summary of the overall communication quality. LLMs perceive that the child's expressions during the conversation are one-sided, requiring the clinician to make huge efforts to sustain the dialogue. This aligns with the criteria for a score of 2, which is consistent with the clinician's judgment.

# I Zero/few-shot learning settings

Moreover, we have devised three prompt settings, incorporating both zero-shot and few-shot settings, to investigate the hypotheses we have put forward in Method Section.

- Only-Scoring (zero-shot):In the prompt, we require the model to provide scores for the eight items only, without the need for justifications.
- Scoring&Explanation (zero-shot):In the prompt, we introduce CoT (Chain of thought)[36] to encourage the model to provide not only scores but also the reasoning and justifications behind those scores. This approach aims to showcase the model's thought process and achieve a more rigorous assessment of the scores.
- Scoring&Explanation (few-shot):In the prompt, inspired by [60, 69], we provide some few-shot examples that includes score, reasoning, and evidence, with the expectation that the model will generate similar outputs to gain more detail explanation.

Table 15 refers to the comparison of results for several models in the **Only-Scoring** (os) and **Scoring&Explanation** (**few-shot**) (fs) settings compared to the **Scoring&Explanation** (**zero-shot**) setting. It is evident that in the only-scoring setting, all LLMs show a significant decrease in performance. The average MAE decreases by 2.84%, the average F1 score for binary classification decreases by 10.61%, and the average F1 score for ternary classification decreases by 13.66%. This confirms our hypothesis that explanations aimed at better scoring. In the few-shot setting, there is an issue with Kimi's refusal to score certain items, which may be attributed to the sensitive settings. On

#### Prompt : Just Scoring(Zero-shot)

- Only need to provide the scores for each item.

#### Prompt: Scoring & Explanation(Zero-shot)

- Required to provide specific scores, reasoning, and evidence. The evidence should include excerpts from the dialogue text that support the given scores.

#### Prompt: Scoring & Explanation(Few-shot)

- Required to provide specific scores, reasoning, and evidence. The evidence should include excerpts from the dialogue text that support the given scores, following the format of the output example.

#### ## Output Example:

- A8: Score 2
- *Reason:* The child can speak in the dialogue, but in limited quantity. He mainly focus on repeating and imitating the dialogue without engaging in a conversation with the participants. Most of the time, he express his thoughts without reciprocal interaction. This aligns with the description of "2 = Few instances of sustained reciprocal conversation; may follow their own train of thought rather than engaging in a dialogue; may offer some spontaneous information or comments, but with little reciprocity." Hence, I would assign a score of 2.
- Evidence:
- Child: He asks Xiao Ming Curiously, "Why?" Doctor: "Who is Xiao Ming?"
- Child: "Why does the dog suddenly have a stomachache?"- Doctor: "Because the dog didn't wash its hands before eating."
- Child: "He had something dirty in his hands, so his stomach hurts."

Figure 12: Optional Prompt For Zero/Few-shot Setting. This Figure illustrates the differences in three prompt settings. We only display the part that needs to be replaced for the -(optional) shown in Figure 4 while the left parts remain unchanged.

Table 15: Performances Comparison of Different Prompts Setting. This table displays the performance comparison results between the Scoring&Explanation(zero-shot) setting, where *os* refers to the Only-Scoring(zero-shot) setting and *fs* refers to the Scoring&Explanation(few-shot) setting. The blue indicate a decrease in performance compared to the Scoring&Explanation(zero-shot) setting, while the red indicate an improvement in performance.

model	Kimi	qwen1.5-32b	Yi-34b	glm4	Avg
$Avg - MAE_{os}$ $F1 - 2_{os}$ $F1 - 3_{os}$	↓ -0.0259 ↓ -0.0880 ↓ -0.1641	↓ -0.0292 ↓ -0.1681 ↓ -0.1015	↓ -0.0259 ↓ -0.0127 ↓ -0.0727	↓ -0.0324 ↓ -0.1556 ↓ -0.2080	↓ -0.0284 ↓ -0.1061 ↓ -0.1366
	\ \	↓ -0.0178 ↓ -0.1227 ↓ -0.0931	↓ -0.1039 ↓ -0.0853 ↓ -0.0673	↑+0.0293 ↑+0.0914 ↑+0.0704	↓ -0.0326 ↓ -0.0389 ↓ -0.0300

the other hand, GLM4 shows a slight performance improvement while Qwen1.5 and Yi-34b show the performance decline. This once again confirms that the few-shot setting does not necessarily lead to significant performance improvements in novel tasks which is never used in pre-training and fine-tuning for LLMs[37]. Therefore, in the main text, we opted to use the **Scoring&Explanation** (zero-shot) prompt setting as the In-context Enhancement Prompt which is shown in Figure 4.