Contextualizing Parental Behaviors in Bilingual Datasets from In-Person and Telehealth Language Assessment

Anonymous ACL submission

Abstract

The increasing adoption of telehealth technologies presents both opportunities and challenges, offering greater convenience for patients while increasing clinicians workload, particularly in managing remotely collected data. Bilingual speech-language pathologists (SLPs) spent substantial effort in evaluating parent behaviors when conducting family-centered language assessments. In this study we collaborate with SLPs to examine how large language models (LLMs) can support clinical workflows and address real-world challenges in telehealth. We collected a detailed bilingual dataset of 59 Mandarin-English child language assessment sessions (16 in-person and 43 via telehealth) and benchmarked three open-source LLMs and one closed-source LLM on this task. All four LLMs are still inferior to human experts despite notable accuracy, and additional error analysis revealed that domain complexity, cultural context, and multimodal cues pose significant challenges for both LLMs and human annotators. This work highlights the need for domain-specific NLP advancement and evaluation methods that extend beyond standard benchmarks to include clinical utility, workflow integration, and cultural appropriateness in bilingual telehealth assessment.

1 Introduction

Language assessments play a critical role in the early detection and intervention of children's communication disorders, particularly in bilingual contexts where assessments must account for linguistic proficiency in multiple languages (Wang et al., 2020; Gorman et al., 2015; Wang et al., 2024). With the rapid expansion of telehealth technologies recently, especially after COVID-19, more bilingual families can access essential medical resources remotely, such as interacting with web-based child language assessment tools at home (Pratt et al., 2022; Dam and Pham, 2023). However, bilingual speech-language pathologists (SLPs) are already in severe shortage in the United States because of the challenging but necessary expertise to evaluate children's language abilities across different linguistic and cultural contexts (Du et al., 2020; Pratt et al., 2022; Dam and Pham, 2023). The adoption of telehealth technologies further poses a significant workload for SLPs because they need to spend a considerable amount of time reviewing large-scale data collected remotely.

During remote assessments, parents are required to supervise and facilitate children's interaction with the telehealth tool. For instance, parents could help provide technical assistance with digital platforms and offer behavioral support (Pozniak et al., 2024; Edwards-Gaither et al., 2023). However, due to the lack of specialized training, parents can unintentionally exhibit interference behaviors, such as repeating assessment instructions and "leaking" the correct answer, which can compromise the validity of assessment results (Du et al., 2020; Tomlinson et al., 2018). Identifying these behaviors typically requires SLPs to conduct manual transcription and meticulous behavioral coding from video-recorded sessions, which is an extremely time-intensive and laborious process (Sun et al., 2024; Cao et al., 2019; Lønfeldt et al., 2023).

Recent advances in Natural Language Processing, particularly large language models (LLMs), have shown promise in automating complex behavioral coding tasks in domain-specific contexts like motivational interviewing and mental health counseling (Cao et al., 2019; Tavabi et al., 2020; Mayer et al., 2024; Pellemans et al., 2024). Nevertheless, these applications are mainly limited to monolingual, adult, and face-to-face interactions, and the application scenarios, unlike medical assessments, often do not require highly professional expertise and strict step-by-step requirements. Little research has explored the potential of

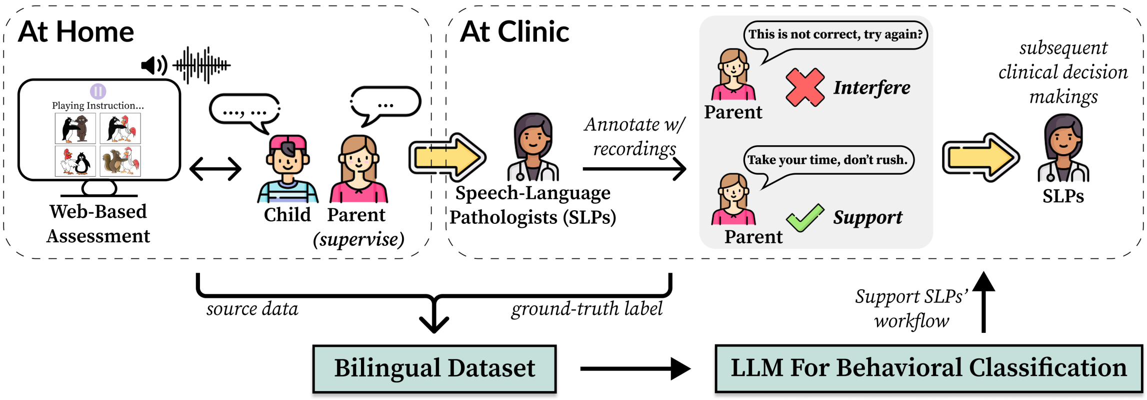


Figure 1: The clinical workflow of speech-language pathologists (SLPs) for remote patient assessment.

LLMs to support clinicians' workflow within the unique context of remote language assessments for bilingual children, where the interplay of linguistic, cultural, and technological factors significantly complicates analysis (Zhang et al., 2023a,b; Lin et al., 2022; Karacan et al., 2024).

In this paper, we address this critical gap by systematically examining how state-of-the-art (SOTA) LLMs, including open-sourced and closed-sourced ones, can support bilingual SLPs' clinical workflows by automating the coding of parental behaviors during Mandarin-English language assessments. We introduce a novel and comprehensive dataset consisting of transcripts and meticulously annotated behavioral descriptions from 59 bilingual parent-child dyads, including both in-person (16) and telehealth (43) sessions. Our dataset encompasses 1,304 annotated parental behaviors categorized into eight clinically validated labels that are collaboratively developed with bilingual SLPs.

We benchmark the performance of four SOTA LLMs: GPT-4 (Achiam et al., 2023), Llama 3 (Grattafiori et al., 2024), Qwen2 (Yang et al., 2024a), and DeepSeek-V3 (Liu et al., 2024) on this bilingual dataset using zero-shot and few-shot prompting techniques. Our analysis reveals that while some models (GPT-4 and DeepSeek-V3) achieve commendable accuracy, all models are consistently outperformed by expert human annotators, particularly when faced with Mandarin-English utterances. An in-depth error analysis further elucidates critical challenges posed by cultural nuances, contextual complexities, and intricate clinical procedures, which highlight substantial space for future exploration in multilingual and culturally sensitive NLP methodologies.

To our knowledge, this study introduces the first publicly available bilingual Mandarin-English dataset specifically to encode parental behaviors in bilingual language assessment contexts. Our findings provide a rigorous benchmark and an ecologically valid challenge for NLP research in clinical scenarios, emphasizing the urgent need to enhance the multilingual and domain-specific capabilities of LLMs. By advancing in this direction, NLP technologies hold the potential to significantly alleviate clinician workload, enhance the accuracy and efficiency of clinical assessments, and ultimately improve patient outcomes.

2 Related Work

2.1 Multilingual LLMs for Real-World Tasks

Recent studies have explored LLMs' capabilities in real-world scenarios that require domain expertise, such as education (Chen et al., 2023; Nayeem and Rafiei, 2024; Valentini et al., 2023; Ghanizadeh and Dousti, 2025) and health care (Riina et al., 2024; Labrak et al., 2024). LLMs like GPT-4 (OpenAI, 2023), Llama (Touvron et al., 2023a,b), Qwen2 (Yang et al., 2024b), and DeepSeek (Bi et al., 2024; Peng et al., 2025; Neha and Bhati) have been trained on multilingual data and demonstrated impressive performance, in tasks like question answering and logical reasoning (Wei et al., 2021; Sanh et al., 2021; Chung et al., 2022), although performances may differ due to language features (e.g., morphosyntax) (Hlavnova and Ruder, 2023; Weissweiler et al., 2023) or use of in-context learning (ICL) (Brown et al., 2020; Zhang et al., 2022; Rubin et al., 2022; Li et al., 2023) or multimodal prompting (Yang et al., 2024c). Prior NLP research has focused on various domain-specific tasks, such as assessment (Wang et al., 2020; Gorman et al., 2015; Laverghetta Jr and Licato, 2023), behavioral analysis (Van Aken et al., 2021; Sun et al., 2024; Cao et al., 2019; Yang et al., 2023), and narrative tasks (Prud'hommeaux and Roark, 2015; Chen et al., 2023).

2.2 Annotating Clinical Assessment Data

Computational researchers have attempted to annotate clinical data (Leeson et al., 2019) using a variety of speech processing techniques (Pérez-Rosas et al., 2021; Narayanan and Georgiou, 2013) and NLP approaches such as topic modeling, multimodal models (Tavabi et al., 2020; Leeson et al., 2019). Differing from other clinical annotation tasks, assessment tasks may directly impact the diagnostic accuracy of patient care, the annotation accuracy requires psychometric standards to ensure assessment validity and reliability (Abbasi et al., 2021; Wang et al., 2020; Gorman et al., 2015; Laverghetta Jr and Licato, 2023). Applying NLP techniques to bilingual telehealth datasets remained limited due to difficulties in accessing patient data and the high cost of human annotations (Chen et al., 2022). Therefore, to improve clinician workflow and accuracy, novel approaches need to be developed with clinically informed guidelines to support clinicians.

2.3 Behavioral Coding in Clinical NLP

Behavioral coding is a common data analysis methodology in social science (Wang et al., 2022; Black et al., 2013) and has been widely adopted in health and clinical research (Tavabi et al., 2020; Cao et al., 2019; Sun et al., 2024; Mayer et al., 2024; Pellemans et al., 2024). Prior work have utilized computational behavioral recognition for coding video and audio data from parent-child interactions (Lønfeldt et al., 2023). However, conducting a child language assessment requires more fine-grained coding for linguistic features in addition to assessment behaviors (Wang et al., 2020; Gorman et al., 2015), and bilingual data may introduce additional complexity for analysis due to issues such as code-switching (Du et al., 2020).

3 Bilingual Dataset

We collaborated with bilingual (Mandarin-English) speech language pathologists (SLP) to collect and annotate a text-based dataset of bilingual child language assessment sessions. The dataset comprises 59 parent-child dyads using the Mandarin-English Receptive Language Screener (MERLS) tool. The dataset consists of 16 in-person sessions and 43 telehealth sessions, where the telehealth data was collected through Zoom recordings during COVID-19. An illustration of the clinical workflow, our dataset collection, and model development process is shown in Figure 1.

MERLS is a multi-modal web-based platform designed for assessing language comprehension skills for Mandarin-English-speaking children (Sheng et al., 2021; Du et al., 2020). The assessment consists of a Mandarin module with 44 test items and an English module with 36 test items. For each test item, MERLS plays audio instructions in one language and asks bilingual children to select a picture on the interface that matches the instruction (see Figure 2). Parents are expected to supervise and provide technical support to children during assessments.

3.1 Data Annotation Process

Raw session video recordings were transcribed for speaker utterances verbatim and documented for parents' verbal and non-verbal behaviors by two research assistants. Next, two bilingual SLPs independently annotate parental behaviors using an established video analysis codebook (Du et al., 2020) developed via Clinical Discourse Analysis (Damico, 1985) to ensure the classification of behaviors was informed based on the existing clinical framework of family-centered assessment practices (Crais et al., 2006). Disagreement between annotators was resolved via member checking method through discussions and codebook refinement until consensus (Birt et al., 2016). Interobserver agreement reached 97% (in-person split) and 86.1% (remote split) as shown in Table 3.

Parental behavior classification conducted by SLPs is an eight-class classification task, with each behavior assigned one correct label. Table 1 presents the two primary classes and four corresponding sub-categories. "***Interference***" behaviors represent when parents negatively impacted the assessment including "*Repeating Questions (RQ)*", "*Answering Questions (AQ)*", "*Analyzing Items (AI)*", and "*Judging of Correctness (JC)*"; "***Support***" behaviors represent incidents when parents positively facilitated the assessment including "*Encouragement (E)*", "*Technical Support (TS)*", "*Broadcasting (B)*", and "*Miscellaneous (M)*". For NLP models, each input includes the current test item, a description of the child's actions, and the parent's behavior.

Top-Level Category	Sub-Level Category	Definition
Interfere	Repeating Questions (**RQ**)	Repeating the <Voiceover> audio before and/or during the process of a child selecting the picture on the web.
	Answering Questions (**AQ**)	Using verbal or gestural cues to suggest or select a correct answer for the child.
	Analyzing Items (**AI**)	Elaborating on the critical linguistic components by labeling objects and actions, making emphasis via prosodic cues, or breaking down complex sentences from <Voiceover>.
	Judging of Correctness (**JC**)	Verbally evaluating the child's response as correct or incorrect.
Support	Encouragement (**E**)	Showing verbal and/physical affirmation for the child to continue, saying good job/excellent to reinforce the child's selection, expressing empathy (e.g., it's okay) on struggled items.
	Technical Support (**TS**)	Offering verbal (labeled as "Technical Support Verbal") and/or physical assistance (labeled as "Technical Support Physical") to the child related to interacting with the website and the computer.
	Broadcasting (**B**)	After the child makes a selection, describing the selection via a word, a phrase, or a sentence.
	Miscellaneous (**M**)	Initiating and/or responding to events that redirected a child's attention, sharing personal opinions about test procedures and stimuli, or other verbal and nonverbal behaviors that were out of the child's view.

Table 1: The classification categories and corresponding definitions of parent behaviors based on established clinical guidelines. Two high-level categories ("interfere" and "support") consist of four individual sub-categories.

	In-Person (n=16)			Virtual (n=43)		
	English	Mandarin	All	English	Mandarin	All
RQ	41	101	142	8	30	38
AQ	0	12	12	0	0	0
AI	18	68	86	3	3	6
JC	20	33	53	1	12	13
E	39	68	107	50	83	133
TS	14	78	92	164	206	370
B	14	42	56	4	10	14
M	12	18	30	77	75	152
Interference	79	214	293	12	45	57
Support	79	206	285	295	374	669

Table 2: MERLS dataset (in-person n=16 and virtual n=43) statistics. The top row shows the label distribution across different datasets and test languages.

3.2 Dataset Description & Statistics

The dataset is structured to include the following components: (1) **Time stamps**: Precise time stamps for each assessment item and corresponding parent-child behavior; (2) **GUI descriptions**: Textual descriptions of the graphical user interface (GUI) elements displayed on the MERLS platform; (3) **Audio transcriptions**: Transcriptions of the audio recordings, with annotations that identify different speakers in each voiceover; (4) **Behavior descriptions**: Textual descriptions of parents' verbal and non-verbal behaviors. An example of such data is illustrated in Figure 6.

Table 2 presents overall statistics for the In-person and Virtual sessions. The two datasets exhibit imbalances in their label distributions: the Virtual split contains fewer interference behaviors and more technical support behaviors. It may be due to (1) an instructional video in the MERLS system aimed to reduce interference behaviors, or (2) the use of the MERLS system during telehealth assessment awhich increased the need for technical support activities conducted by parents.

4 Behavior Classification with LLMs

Our experiments focus on zero-shot (ZS) and few-shot (FS) in-context learning (ICL) prompting strategies for LLMs to investigate whether LLMs can reliably classify parental behaviors during child language assessments.

Prompts. Our zero-shot prompt in Figure 3 provides instructions, explains the input format, and defines each of the eight labels. From the test example itself, the model is shown (1) the text of the current question, (2) a description of the child's behavior, and (3) the description of the parent's behavior. The few-shot prompt is similar but includes one demonstrative example (by a clinical expert to ensure validity) under each label definition. Figures 4 and 5 in the Appendix show the few-shot prompts, split over multiple pages. Both prompts include (1) **Voiceover**, the text of the current question; (2) **Child behavior**, a description of the child's behavior, and (3) **Parent behavior**, the utterance and/or a description of the action performed by the parent.

Models. To evaluate NLP performance on our tasks, we experiment with four LLMs: the open-weight models Llama-3-8B-Instruct, DeepSeek-V3, and Qwen2-7B, and the closed-source model GPT-4 (Turbo-2024-04-09). While Llama-3 is primarily English-based, its pre-training data includes data from 30 other languages.[1]

[1] https://ai.meta.com/blog/meta-llama-3/

Micro F1 / Macro F1 (%) on:	In-Person (n=16)			Virtual (n=43)		
	English	Mandarin	Overall	English	Mandarin	Overall
Llama3 ZS	48.7/37.4	43.6/33.7	45.0/34.6	29.6/21.9	31.5/20.8	30.7/21.0
Llama3 FS	45.6/31.8	39.8/26.2	41.3/27.6	22.1/9.3	23.6/10.1	23.0/9.9
GPT-4 ZS	65.8/58.5	60.5/55.7	61.9/**57.2**	45.3/22.3	52.7/38.2	49.6/33.8
GPT-4 FS	66.5/**61.1**	55.0/49.1	58.1/52.4	50.2/**28.8**	53.9/38.2	52.3/**36.2**
DeepSeek-V3 ZS	61.7/53.5	58.5/51.2	59.4/52.4	47.6/23.6	53.2/**38.9**	50.8/34.3
DeepSeek-V3 FS	64.6/54.9	63.5/**56.2**	63.8/56.3	52.4/25.2	53.7/36.2	53.2/33.3
Qwen2 ZS	31.4/19.6	28.3/19.6	29.2/20.0	38.8/13.9	40.7/20.3	39.9/17.8
Qwen2 FS	20.1/15.2	25.6/21.5	24.1/19.7	41.6/17.0	42.1/24.2	41.9/22.1
Human Experts	96.84	96.43	97.0	86.93	81.82	86.1

Table 3: Micro/Macro F1 results on assessment language for In-Person and Virtual dataset. ZS = zero-shot, FS = few-shot. The best-performing values for each metric are highlighted.

Macro F1 on:	In-Person (n=16)		Virtual (n=43)	
	Interference	Support	Interference	Support
Llama3 ZS	75.2	70.5	18.0	76.4
Llama3 FS	73.5	64.6	20.3	69.3
GPT-4 FS	**87.6**	**86.7**	**37.4**	90.1
DeepSeek-V3 ZS	78.0	81.4	20.4	86.1
DeepSeek-V3 FS	85.0	85.9	29.2	**90.6**
Qwen2 ZS	54.7	66.0	25.5	87.0
Qwen2 FS	56.3	57.4	28.2	84.7

Table 4: Macro F1 on the binary classification version ("Interference" vs. "Support" behaviors) of In-Person and Virtual dataset. ZS = zero-shot, FS = few-shot.

Qwen2 (Yang et al., 2024b) achieved strong benchmarks across approximately 30 languages. GPT-4 has also demonstrated strong performance on Chinese language understanding benchmarks (Xu et al., 2023; Zhu et al., 2024). Deepseek-V3 is optimized for computational efficiency and excels in complex linguistic and reasoning tasks with minimal supervised data (Liu et al., 2024).

Evaluation metrics. To accurately evaluate parent behaviors across different test items in Mandarin and English, we compute three metrics: (1) Macro F1 score (MACRO): prediction performance addressing the effects of dataset imbalance; (2) Micro F1 score (MICRO): precision and recall computed over all prediction instances; (3) Item-level Accuracy (ITEMACC): the proportion of items with at least one behavior where all behaviors are predicted correctly. ITEMACC is calculated as the number of items with at least one behavior that is predicted entirely correctly, divided by the total number of items that contained at least one behavior in each item.

5 Experimental Results

5.1 Main Results

Table 3 presents the micro- and macro-F1 scores for each model in the assessment items of English and Mandarin in both the in-person and virtual settings. Overall, GPT-4 and DeepSeek-V3 consistently outperform Llama3 and Qwen2 across all settings. Their stronger performance is in line with established scaling laws in large language models (LLMs), rather than indicating domain-specific adaptation (Wei et al., 2021)

The consistent gap between macro and micro F1 across settings reflects the severe class imbalance in our dataset. Since macro F1 gives equal weight to each class, even infrequent behaviors such as Analyzing Items, which appears only three times in the Virtual dataset (Table 2), can significantly impact the overall score. In the following interpretations, we emphasize macro F1 because it provides a more rigorous and fair evaluation across all behavior categories, especially in imbalanced datasets where some behaviors are rare but clinically meaningful.

On the In-Person dataset, GPT-4 achieves the highest macro F1 score on English assessment items (61.1%), while DeepSeek-V3 leads in Mandarin (56.3%). The strong Mandarin performance of DeepSeek-V3 likely reflects its exposure to high-quality Chinese-language sources (Guo et al., 2024). Llama3 shows moderate performance in both languages, consistently lagging behind GPT-4 and DeepSeek-V3 but outperforming Qwen2, which remains the weakest model, especially in English, even with few-shot support.

On the Virtual dataset, all models perform worse than in the In-Person setting. GPT-4 and DeepSeek-V3 remain the top performers, while Llama3 shows moderate performance, and Qwen2 ranks lowest. Interestingly, this performance drop occurs despite a higher number of annotated parental behaviors in Virtual sessions. This suggests that the challenge lies not in annotation sparsity but in the increased variability and complexity of behaviors exhibited during virtual interactions, which may be harder for models to learn and classify accurately. These findings highlight the unique difficulty of modeling parental behaviors in remote formats, where behavior patterns may be different from those in in-person settings.

In both in-person and virtual settings, model performance varied across clinical contexts, and the paired-setting analysis (Table 9), based on few-shot results, provides clearer insight into how these setting differences influence classification. GPT-4 exhibited the most pronounced drop, with macro F1 scores decreasing from 53.4% in the in-person setting to 27.7% in the virtual setting, particularly struggling with English in the virtual context. DeepSeek-V3, while still achieving the highest overall scores, also showed a substantial decline from 56.3% to 28.1%, indicating that it too was affected by setting-related challenges. Qwen2 was the only model to improve in the Virtual setting, rising from 19.7% to 22.6%, with its best performance observed in virtual Mandarin sessions (25.9%). In contrast, Llama3 performed poorly across all conditions, with macro F1 scores falling from 27.6% in-person to just 8.8% in virtual sessions, and without a consistent language-specific trend. These findings emphasize that the clinical setting (in-person or virtual) has a substantial impact on model performance, and that the ability to generalize across settings varies widely by model, language-specific pretraining, and robustness to behavioral variability.

5.2 Effects of Parent Language

Table 6 presents micro and macro F1 scores broken down by the language used to describe parent behaviors in the transcriptsEnglish, Mandarin, or a mix of both. In our dataset, non-verbal parent behaviors are consistently described in English. Code-mixing occurs when parents code-switch during speech or when Mandarin utterances are paired with English descriptions of non-verbal actions. Across most models, macro F1 scores are highest on English or Mixed-language transcripts and lowest on Mandarin-only transcripts, especially in the Virtual setting. GPT-4 demonstrates the most consistent performance across languages, with balanced macro F1 scores for English and Mandarin transcriptions in both settings (e.g., 51.3% vs. 46.7% in-person; 35.4% vs. 34.7% virtual). It achieves the best result in all models on Mandarin-only data in the Virtual setting.

In contrast, DeepSeek-V3 performs best on English and Mixed-language transcripts, but its macro F1 on Virtual Mandarin (29.9%) is lower than GPT-4s. Qwen2 shows a relatively small performance gap between English and Mandarin, although its overall accuracy remains low, with its highest macro F1 score (27.1%) occurring on Mixed-language Virtual transcripts. Llama3 performs poorly across all conditions, with its lowest macro F1 on Virtual Mandarin transcripts (6.9%). These results demonstrate that transcription language significantly influences model performance, with Mandarin-only transcripts, particularly in Virtual sessions, posing the greatest challenge.

Interestingly, while DeepSeek-V3 outperforms GPT-4 in overall Mandarin session performance (Table 9), GPT-4 surpasses DeepSeek when classifying Mandarin-only behavior transcriptions (Table 6), especially in the Virtual setting (34.7% vs. 29.9%). This discrepancy highlights the distinction between *session language* (i.e., whether the child was tested in Mandarin) and the *transcription language* (i.e., whether the parents behavior was described using Mandarin). DeepSeek-V3 may be more effective at capturing contextual patterns in full Mandarin-language sessions, whereas GPT-4 appears better at parsing Mandarin within isolated transcription entries. These findings underscore the importance of distinguishing between session language and transcription language when evaluating multilingual performance in LLM-based behavioral classification.

5.3 Binary Classification Results

Table 4 presents macro F1 scores for the binary classification task distinguishing Interference and Support behaviors across In-Person and Virtual sessions. This analysis enables a clearer understanding of how LLMs perform when behavior categories are simplified, reducing the impact of class imbalance present in the 8-way classification task.

Across all models, performance is substantially higher in the In-Person setting than in the Virtual setting for both behavior types. GPT-4 achieves the highest overall performance, with macro F1 scores of 87.6% for Interference and 86.7% for Support in the In-Person setting. While its performance drops in the Virtual setting, it still maintains relatively strong accuracy, especially on Support behaviors (90.1%). DeepSeek-V3 closely follows, with similar In-Person scores (85.0% and

In-Person English				In-Person Mandarin			
Label pair	Annotator 1 Acc	Annotator 2 Acc	Overall	Label pair	Annotator 1 Acc	Annotator 2 Acc	Overall
RQ-AI	100.0	75.0	75.0	RQ-AI	60.0	70.0	60.0
AI-RQ	0.0	100.0	**0.0**	AI-RQ	69.0	100.0	69.0
JC-E	71.4	71.4	**57.1**	**E-RQ**	22.2	0.0	**0.0**
				E-M	40.0	16.6	**16.6**
				TS-B	100.0	100.0	100.0
Virtual English				Virtual Mandarin			
Label pair	Annotator 1 Acc	Annotator 2 Acc	Overall	Label pair	Annotator 1 Acc	Annotator 2 Acc	Overall
TS-M	82.2	53.3	**53.3**	TS-M	85.7	90.2	74.5
M-E	75	92.3	68.8	**M-E**	68.8	61.5	**61.5**

Table 5: Clinician annotation accuracy based on the misclassified pairs from Figure 14. A-B denotes that A is the true clinician-annotated label, while B represents the GPT prediction. The overall accuracy is calculated as the number of correctly classified behaviors for both clinicians divided by the total number of misclassified behavior pairs. These low values (highlighted in the table) suggest that these misclassified pairs are also somewhat challenging for clinicians to classify accurately.

85.9%) and slightly lower Virtual performance on Interference (29.2%) but the highest score on Support (90.6%). Llama3 and Qwen2 perform considerably worse on Interference behaviors in Virtual settings, with macro F1 scores below 30%. However, their performance on Support behaviors remains relatively strong in Virtual contexts (e.g., 84.7% for Qwen2 FS and 76.4% for Llama3 ZS). This suggests that Support behaviors are more consistently recognized across models, whereas Interference behaviors are more difficult to detect, especially in Virtual sessions where contextual cues may be limited or harder to interpret.

These findings reinforce that LLM performance degrades in virtual environments, particularly for subtle or ambiguous behavioral categories like Interference. However, simplifying the task to binary classification improves overall accuracy and highlights meaningful variation across model architectures and prompting strategies.

5.4 Error Analysis with Human Annotators

While our primary evaluation focused on macro and micro F1 scores (Tables 3 and 4), we conducted a detailed error analysis using item-level accuracy to identify challenging behavior pairs, focusing on one of the top-performing models: GPT-4 ZS for the In-Person dataset, and GPT-4 FS for the Virtual dataset. The confusion matrices in the Appendix Figures 14 identified the misclassified pairs. To examine these errors, we selected the most frequently misclassified pairs for each behavioral category within each dataset (Table 5). These misclassified data were selected based on two criteria: 1) the number of instances in the class is no less than the average of that dataset, and 2) the prediction accuracy for that class is below 80%. Two novel clinical expert annotators were trained using the same clinical protocol in Table 3 for error analysis. Novice human annotators were not utilized due to the specialized training required for behavioral coding, without such training, they have variability and reliability issues. Two annotators independently selected which option they believe is the correct answer without knowing which class was machine or human annotation. By comparing their responses to the true labels, we identified challenging pairs the clinical experts also struggled to perform classification.

Errors in In-Person Dataset. Several commonly misclassified pairs by GPT-4 can be identified in the top rows from Table 5. Overall, the LLM struggles to distinguish "Repeating Questions (RQ)" from "Analyzing Items" in both English and Mandarin tests. Notably, this distinction is also challenging for clinicians, as the AI-RQ category for the In-Person dataset (English) shows 0.0% accuracy, with significant disagreement among experts. Figure 6 illustrates a sample parent-child interaction transcript for this disagreement between the best-performing model (GPT4-ZS) and two human annotators.

A key insight from our analysis is that LLM errors reveal potential *linguistic ambiguities* in the clinical definition of parent behaviors. GPT-4 mislabels “Analyzing Items (AI)" as “Repeating Questions (RQ)" when parents repeat only key components of a question (e.g., "wash the cat") for their children across both in-person and virtual datasets.

This is likely because GPT-4 lacks the specific details to differentiate by repeating how much of the partial question is considered "Analyzing Items." The errors also appeared in code-mix utterances and could be due to inadequate translation from word-level lexicon to sentence-level utterances. Through the disagreements between human annotators and predictions from LLMs, we gain a deeper insights onimproving both the development of a clinical annotation codebook (Leeson et al., 2019) as well as the prompting strategies for LLMs (Lin et al., 2022; Ranaldi and Pucci, 2023; Lønfeldt et al., 2023).

Errors in the Virtual Dataset. In the virtual dataset, the common misclassifications are demonstrated in the bottom rows from Table 5. We found the challenging pairs for clinicians to classify are "TS-M" ("Technical Support" vs. "Miscellaneous") for Virtual English and "M-E" ("Miscellaneous" vs. "Encouragement") for Virtual dataset in Mandarin. Figure 7 in the appendix illustrates a sample transcript, showing disagreement across two human annotators and the best-performing model (GPT-4 FS). GPT-4 appears to overgeneralize utterances that do not contain words related to "website" or "computer" as "Technical Support" behaviors. This is indeed due to *behavioral definition ambiguities* originated from the clinician's codebook which informed the prompting. After two annotators conducted the member-checking procedure (Birt et al., 2016) to discuss inconsistent annotations, they identified additional utterances (e.g., parents monitoring children's needs for breaks or snacks) as a new potential category of support behavior in the virtual dataset. This error analysis process highlights the models limitations in analyzing novel information, suggesting that future use of more aligned behavioral coding definitions to inform prompting for LLMs to accurately identify existing categories accurately and recognize novel patterns to enhance behavioral coding process. In a deeper error analysis, we conducted an ablation study to assess the impact of various components of GPT-4's performance, see results in Appendix B.

6 Conclusion

This paper introduces a bilingual dataset for classifying parental behaviors during English-Mandarin child language assessments. This study highlights the potential of LLMs to support bilingual language assessment in clinical and school settings. By identifying supportive and interfering parental behaviors, automated classification can inform improvements in family-centered assessment practices with enhanced clinical validity. Additionally, LLMs may help alleviate the shortage of bilingual SLPs by supporting screening processes, expanding access to linguistically appropriate evaluations, and reducing the risk of over-diagnosing language disorders in emergent bilingual children.

While current SOTA LLMs show moderate accuracy, they struggle with Mandarin data, a challenge also faced by human annotators, particularly with virtual data. These difficulties point to potential ambiguities in behavioral definitions, especially in cases involving code-switching and nuanced parental language. This highlights the need to refine the clinical annotation codebook, which may improve consistency across both human and automated coding efforts. This dataset promotes further NLP research for multilingual clinical tasks, advancing the analysis of using multimodal behavioral coding (Yang et al., 2024c) of bilingual datasets (Hlavnova and Ruder, 2023; Weissweiler et al., 2023) during child language assessment in in-person and telehealth contexts.

7 Future Work

Our bilingual dataset also contained prosodic information (e.g., parents emphasize words when analyzing items with different stress patterns) which has been shown to introduce additional complexity in clinical NLP (Black et al., 2013), especially in the bilingual context (Pattichis et al., 2023). Manual transcription by clinicians introduces potential for error and inconsistency. To reduce such errors, future work should continue exploring available LLMs to achieve better performance or integration of multimodal speech (OpenAI Whisper, Llama-Omni, Qwen2-Audio) that can automate audio-to-text transcription for additional technical novelty in this research (Tavabi et al., 2020). By aligning our methodology with clinical annotation guidelines and error analysis, future work may include collaborating with clinicians to refine annotation guidelines and additional error analysis, and aligning model outputs with assessment protocols, and piloting LLM-supported language assessments in clinical settings for bilingual children.

8 Limitations

Our study is constrained by the imbalance between the in-person and virtual datasets, as well as a relatively small sample size with (1,304 parent behavior annotations from 59 sessions), which further limited underrepresented classification categories (e.g., broadcasting). Such a small sample size further affect generalizability and classification reliability; additionally, the imbalance in label distribution across the eight categories contributes to this limitation, affecting overall model performance. For instance, smaller subcategories like "Broadcasting" make up less than 1% compared to larger categories such as "Technical Support" impacting overall model evaluation. Despite the small sample size, our clinically informed annotation guideline could be extended to the ongoing data collection with our clinical partners; it can also inform clinical research in other bilingual populations, such as Spanish-English and Vietnamese-English speaking children and parents (Dam and Pham, 2023; Pratt et al., 2022).

Similar to other closed-source models, our best-performing model GPT-4 poses challenges for replication. One challenge is the semantic alignment at the word and utterance level and how this impacts behavioral classification, especially in bilingual datasets (Cao et al., 2019; Huzaifah et al., 2024). Although our behavioral classifications included categories such as "Encouragement," due to the scope of this paper, we did not explore the use of sentiment analysis (Zhang et al., 2023a) on specific linguistic features in the transcripts to improve accuracy for individual subcategories.

Additionally, error analysis revealed challenges in classifying behaviors related to the "Miscellaneous" category for the best performing model GPT-4 and human annotators. Due to the scope of this paper, we did not conduct additional ablation studies to evaluate more specifically defined behavioral categories. Since this is a bilingual dataset from a novel clinical pilot study, ongoing collaborations between NLP researchers and clinicians may continue to improve the behavioral coding protocol as well as the performance of LLMs.

Furthermore, the lack of improvement in model performance with ICL suggests that the primary bottleneck in this classification task may not arise from challenges in semantic understanding of the text, but rather from capturing the deeper, underlying intent embedded in parental language.

9 Ethical Considerations

Data collection and analysis. Our dataset was collected and processed under university human subject research approval and data sharing agreements. The de-identified text transcripts from the clinical video analysis being used for model evaluation contain no sensitive information about parent-child pairs.

Bias mitigation. When comparing the 16 virtual and 16 in-person parent-child pairs, we considered the effects of childrens age and parents education level, which can influence language abilities and parental behavior during assessments. A detailed parent-child demographic table for the 16 In-person and Virtual pairs are provided in Table 11.

Privacy and trust. Privacy issues are paramount when annotating parent behaviors and utterances. Annotators anonymized sensitive information in the transcripts (e.g., using boys name). This is crucial when applying LLMs to clinical data to protect patient privacy. Implementing LLMs in evaluating clinical data could lead to biases that affect clinical decision-making. Clinicians may rely on models without fully understanding their limitations. We also evaluated GPT-4S misclassifications alongside two human annotators to compare decision-making and address biases, which is essential for building trust and explainable AI in clinical settings.

References

Ahmed Abbasi, David Dobolyi, John P Lalor, Richard G Netemeyer, Kendall Smith, and Yi Yang. 2021. Constructing a psychometric testbed for fair natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Linda Birt, Suzanne Scott, Debbie Cavers, Christine Campbell, and Fiona Walter. 2016. Member checking: a tool to enhance trustworthiness or merely a nod to validation? *Qualitative health research*, 26(13):1802–1811.

Matthew P Black, Athanasios Katsamanis, Brian R Baucom, Chi-Chun Lee, Adam C Lammert, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2013. Toward automating a human behavioral coding system for married couples interactions using speech acoustic features. *Speech communication*, 55(1):1–21.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jie Cao, Michael Tanana, Zac E Imel, Eric Poitras, David C Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. *arXiv preprint arXiv:1907.00326*.

Jiaju Chen, Yuxuan Lu, Shao Zhang, Bingsheng Yao, Yuanzhe Dong, Ying Xu, Yunyao Li, Qianwen Wang, Dakuo Wang, and Yuling Sun. 2023. Fairytalecqa: Integrating a commonsense knowledge graph into children's storybook narratives. *arXiv preprint arXiv:2311.09756*.

Zhuohao Chen, Nikolaos Flemotomos, Zac E Imel, David C Atkins, and Shrikanth Narayanan. 2022. Leveraging open data and task augmentation to automated behavioral coding of psychotherapy conversations in low-resource scenarios. *arXiv preprint arXiv:2210.14254*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Elizabeth R Crais, Vicky Poston Roy, and Karen Free. 2006. Parents' and professionals' perceptions of the implementation of family-centered practices in child assessments.

Quynh Diem Dam and Giang T Pham. 2023. Remote first-language assessment: Feasibility study with vietnamese bilingual children and their caregivers. *Language, Speech, and Hearing Services in Schools*, 54(2):618–635.

J Damico. 1985. Clinical discouse analysis: a functional approach to language assessment. *Communication skills and classroom success*.

Yao Du, Li Sheng, and Katie Salen Tekinbas. 2020. " try your best" parent behaviors during administration of an online language assessment tool for bilingual mandarin-english children. In *proceedings of the interaction design and children conference*, pages 409–420.

Lesley Edwards-Gaither, Ovetta Harris, and Valencia Perry. 2023. Viewpoint telepractice 2025: Exploring telepractice service delivery during covid-19 and beyond. *Perspectives of the ASHA Special Interest Groups*, 8(2):412–417.

Mohammad Amin Ghanizadeh and Mohammad Javad Dousti. 2025. Towards data-efficient language models: A child-inspired approach to language learning. *arXiv preprint arXiv:2503.04611*.

Kyle Gorman, Steven Bedrick, Géza Kiss, Eric Morley, Rosemary Ingham, Metrah Mohammad, Katina Papadakis, and Jan PH van Santen. 2015. Automated morphological analysis of clinical language samples. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2015, page 108. NIH Public Access.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. 2024. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Ester Hlavnova and Sebastian Ruder. 2023. Empowering cross-lingual behavioral testing of nlp models with typological features. *arXiv preprint arXiv:2307.05454*.

Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. Evaluating code-switching translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394.

Baris Karacan, Ankit Aich, Avery Quynh, Amy Pinkham, Philip Harvey, Colin Depp, and Natalie Parde. 2024. Towards comprehensive language analysis for clinically enriched spontaneous dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16457–16472.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Antonio Laverghetta Jr and John Licato. 2023. Generating better items for cognitive assessments using large language models. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 414–428.

William Leeson, Adam Resnick, Daniel Alexander, and John Rovers. 2019. Natural language processing (nlp) in qualitative public health research: a proof of concept study. *International Journal of Qualitative Methods*, 18:1609406919887021.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified Demonstration Retriever for In-Context Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Nicole Nadine Lønfeldt, Sneha Das, Flavia Dalia Frumosu, Anna-Rosa Cecilie Mora-Jensen, Anne Katrine Pagsberg, and Line Clemmensen. 2023. Scaling-up behavioral observation with computational behavior recognition.

Tobias Mayer, Neha Warikoo, Amir Eliassaf, Dana Atzil-Slonim, and Iryna Gurevych. 2024. Predicting client emotions and therapist interventions in psychotherapy dialogues. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1463–1477.

Shrikanth Narayanan and Panayiotis G Georgiou. 2013. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE*, 101(5):1203–1233.

Mir Tafseer Nayeem and Davood Rafiei. 2024. Kidlm: Advancing language models for children–early insights and future directions. *arXiv preprint arXiv:2410.03884*.

Fnu Neha and Deepshikha Bhati. A survey of deepseek models. *Authorea Preprints*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Rebecca Pattichis, Dora LaCasse, Sonya Trawick, and Rena Cacoullos. 2023. Code-switching metrics using intonation units. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16840–16849.

Mathijs Pellemans, Salim Salmi, Saskia Mérelle, Wilco Janssen, and Rob van der Mei. 2024. Automated behavioral coding to enhance the effectiveness of motivational interviewing in a chat-based suicide prevention helpline: secondary analysis of a clinical trial. *Journal of medical internet research*, 26:e53562.

Yifan Peng, Bradley A Malin, Justin F Rousseau, Yanshang Wang, Zihan Xu, Xuhai Xu, Chunhua Weng, and Jiang Bian. 2025. From gpt to deepseek: Significant gaps remains in realizing ai in healthcare.

Verónica Pérez-Rosas, Rada Mihalcea, et al. 2021. Evaluating automatic speech recognition quality and its impact on counselor utterance coding. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 159–168.

Kinga Pozniak, Peter Rosenbaum, and Elaine Yuen Ling Kwok. 2024. Tasks performed by parents to enable telepractice for children with communication disorders: an interview study with clinicians and parents. *Disability and Rehabilitation*, 46(8):1547–1558.

Amy S Pratt, Jissel B Anaya, Michelle N Ramos, Giang Pham, Miriam Muñoz, Lisa M Bedore, and Elizabeth D Peña. 2022. From a distance: Comparison of in-person and virtual assessments with adult–child dyads from linguistically diverse backgrounds. *Language, Speech, and Hearing Services in Schools*, 53(2):360–375.

Emily Prud'hommeaux and Brian Roark. 2015. Graph-based word alignment for clinical language evaluation. *Computational Linguistics*, 41(4):549–578.

Leonardo Ranaldi and Giulia Pucci. 2023. Does the english matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183.

Nicholas Riina, Likhitha Patlolla, Camilo Hernandez Joya, Roger Bautista, Melissa Olivar-Villanueva, and Anish Kumar. 2024. An evaluation of english to spanish medical translation by large language models. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 222–236.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning To Retrieve Prompts for In-Context Learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Li Sheng, Danyang Wang, Caila Walsh, Leah Heisler, Xin Li, and Pumpki Lei Su. 2021. The bilingual home language boost through the lens of the covid-19 pandemic. *Frontiers in Psychology*, 12:667836.

Xin Sun, Jiahuan Pei, Jan de Wit, Mohammad Aliannejadi, Emiel Krahmer, Jos TP Dobber, and Jos A Bosch. 2024. Eliciting motivational interviewing skill codes in psychotherapy with llms: A bilingual dataset and analytical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5609–5621.

Leili Tavabi, Kalin Stefanov, Larry Zhang, Brian Borsari, Joshua D Woolley, Stefan Scherer, and Mohammad Soleymani. 2020. Multimodal automatic coding of client behavior in motivational interviewing. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 406–413.

Serena RL Tomlinson, Nick Gore, and Peter McGill. 2018. Training individuals to implement applied behavior analytic procedures via telehealth: A systematic review of the literature. *Journal of Behavioral Education*, 27:172–222.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arxiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arxiv:2307.09288.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina Kann. 2023. On the automatic generation and simplification of children's stories. *arXiv preprint arXiv:2310.18502*.

Betty Van Aken, Sebastian Herrmann, and Alexander Löser. 2021. What do you see in this patient? behavioral testing of clinical nlp models. *arXiv preprint arXiv:2111.15512*.

Danyang Wang, Alexander Choi-Tucci, Anita Mendez-Perez, Ronald B Gillam, Lisa M Bedore, and Elizabeth D Peña. 2024. Where to start: Use of the bilingual multidimensional ability scale (b-mas) to identify developmental language disorder (dld) in bilingual children. *International Journal of Speech-Language Pathology*, pages 1–17.

Yilei Wang, Jingyuan Tian, Deniz S Ones, Richard N Landers, et al. 2022. Using natural language processing and machine learning to replace human content coders. *Psychological Methods*.

Yiyi Wang, Emily PrudHommeaux, Meysam Asgari, and Jill Dolata. 2020. Automated scoring of clinical expressive language evaluation tasks. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 177. NIH Public Access.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, et al. 2023. Counting the bugs in chatgpt's wugs: A multilingual investigation into the morphological capabilities of a large language model. *arXiv preprint arXiv:2310.15113*.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark. *Preprint*, arXiv:2307.15020.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze

Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jian Yang, Hongcheng Guo, Yuwei Yin, Jiaqi Bai, Bing Wang, Jiaheng Liu, Xinnian Liang, Linzheng Cahi, Liqun Yang, and Zhoujun Li. 2024c. m3p: Towards multimodal multilingual translation with multimodal prompt. *arXiv preprint arXiv:2403.17556*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Indra Winata, and Alham Fikri Aji. 2023a. Multilingual large language models are not (yet) code-switchers. *arXiv preprint arXiv:2305.14235*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023b. Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. *arXiv preprint arXiv:2305.16339*.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active Example Selection for In-Context Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jie Zhu, Junhui Li, Yalong Wen, and Lifan Guo. 2024. Benchmarking large language models on cflue – a chinese financial language understanding evaluation dataset. *Preprint*, arXiv:2405.10542.

A Appendix A

A.1 User Interface of MERLS

Figure 2 illustrates the interaction of a sample test item in the English subtest. When children interact with MERLS, they first hear an audio recording (e.g., "The chicken is hugged by the penguin.") and then select the corresponding picture that matches the audio to demonstrate their understanding of the sentence in English. Parents may or may not demonstrate a behavior depending on their observation of the item and how children respond to the item.

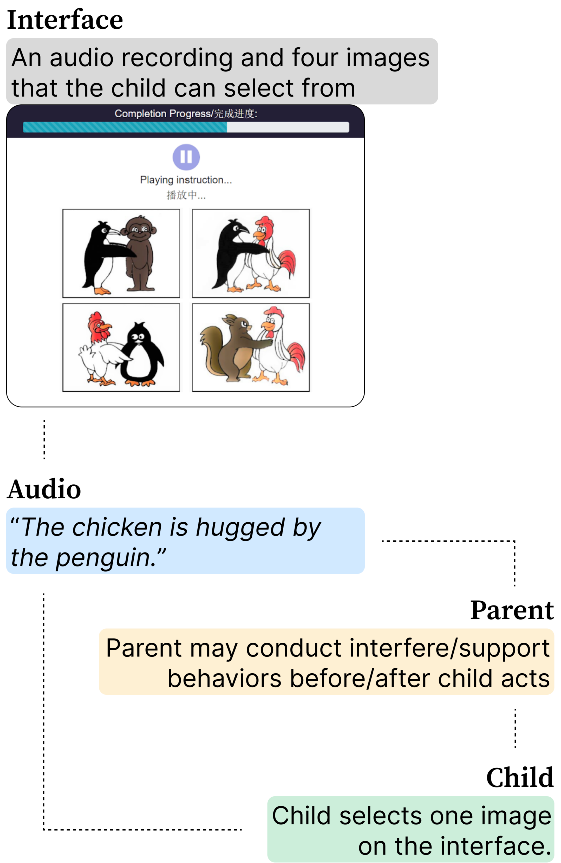


Figure 2: MERLS English test item “The chicken is hugged by the penguin.”

A.2 Results of item level accuracy

Here we also provide the accuracy in item level. Table 7 shows the ItemAcc for the entire In-Person dataset (n=16) and the entire Virtual dataset (n=43) partitioned upon question languages, whereas Table 8 shows the ItemAcc for the entire In-person dataset (n=16) and the subset of the Virtual dataset (n=16) that is matched in terms of child age and parent education level.

B Appendix B: Ablation Study

B.1 Removal of role play description in prompting

This experiment tests whether understanding the role and scenario is necessary, or if simple instructions based on ‘annotation of parental behavior’ suffice, by removing the ‘role play’ description and retaining only the basic instructions to evaluate both datasets. We tested GPT-4’s performance by using a prompt based on our original zero-shot test in Figure 3, but without the role description (e.g., "Assume you are a video analyst classifying transcribed text conversation...") (see prompt in Figure 12). The corresponding results are summarized in Table 10. The maximum variation in accuracy was approximately 3%, and the maximum variation in F1 score was about 4%. Therefore, we conclude that the ‘role play’ descriptor does not significantly impact the overall prediction results.

B.2 Chain of Thoughts Experiment

To further evaluate the factors contributing to performance variance, we then conducted a Chain of Thoughts (CoT) experiment with GPT-4 to assess the best model performance for in-person data (using zero-shot prompts) and virtual data (using few-shot prompts) after retaining "role play" scenario. We then conducted an error analysis across all categories to compare GPT-4’s reasoning with that of human expert annotators. We designed the prompts following a similar two-step procedure as outlined in (Kojima et al., 2022). In the first prompt, we described the classification task as in the original paper, with the addition of a trigger sentence: Lets think step by step. This encourages GPT-4 to generate a step-by-step reasoning process as output (see Figure 8, 9 and 10 for the first prompt input in the Appendix). In the second prompt (see Figure 11 in the Appendix), we combined the original task description with GPT-4’s analysis from the first step and include an answer extraction instruction, such as: Therefore, among all the categories, please respond with the category name only. After running the experiment with prompts that include CoT, we then analyzed the initial responses in greater depth by examining the step-by-step reasoning provided. Here we provide one example where GPT-4 makes a false prediction as demonstrated in Figure 13. In this example, GPT-4 falsely predicts the parent behavior to be "Repeat Questions", however the human expert annotates "Technical Support" because the parent guides the child to make a selection and request a system repetition independently. GPT-4 incorrectly interpreted the parent behavior verbatim as "Repeat Questions" because the utterance itself seems to request repetition; in reality, the parent did not repeat any of the test items and did not interfere with the child. This difference in interpretation provides evidence that GPT models lack specific knowledge of this clinical context, and adding more examples (e.g., few-shot) may not necessarily help model performance.

Micro F1 / Macro F1 (%) on:	In-Person (n=16) English	In-Person (n=16) Mandarin	In-Person (n=16) Mixed	Virtual (n=43) English	Virtual (n=43) Mandarin	Virtual (n=43) Mixed
# Examples	233	167	178	226	389	111
Llama3 ZS	39.9/31.2	42.5/22.0	53.9/39.6	44.8/27.5	21.1/18.1	36.0/16.5
Llama3 FS	35.6/25.7	40.1/20.8	50.0/33.1	32.7/12.1	14.7/6.91	32.4/14.6
GPT-4 ZS	54.2/47.1	49.1/45.9	59.0/52.1	55.3/35.1	41.9/30.8	55.0/**32.5**
GPT-4 FS	58.8/51.3	56.9/**46.7**	57.9/52.4	56.2/35.4	43.7/**34.7**	47.7/29.7
DeepSeek-V3 ZS	59.2/49.4	58.7/45.7	59.0/50.6	65.9/**40.0**	42.4/30.7	49.5/30.1
DeepSeek-V3 FS	63.9/**52.9**	61.1/45.5	65.7/**56.6**	66.4/39.9	46.3/29.9	50.5/27.1
Qwen2 ZS	34.5/18.6	30.2/17.0	20.4/16.6	41.0/15.2	39.7/18.8	39.6/12.9
Qwen2 FS	27.4/17.9	23.8/18.0	20.6/16.6	44.5/21.9	37.3/19.7	53.0/27.1

Table 6: Micro/Macro F1 results broken down by the language in the transcript of the parent's behavior (either English, Mandarin, or a mix of both languages). ZS = zero-shot, FS = few-shot. The best-performing values for each metric are highlighted.

ItemAcc(%) on:	In-Person (n=16) English	In-Person (n=16) Mandarin	In-Person (n=16) All	Virtual (n=43) English	Virtual (n=43) Mandarin	Virtual (n=43) All
Llama3 ZS	39.8	28.0	31.7	26.7	24.3	25.3
Llama3 FS	34.5	27.6	29.8	20.3	19.8	20.0
GPT-4 ZS	**61.9**	49.2	**53.2**	36.9	45.3	41.6
GPT-4 FS	61.1	40.8	47.1	43.3	**46.5**	**45.1**
DeepSeek-V3 ZS	54.2	44.5	47.5	42.8	43.6	43.3
DeepSeek-V3 FS	57.5	**50.9**	53.0	**45.5**	44.9	**45.1**
Qwen2 ZS	20.4	16.0	17.4	29.4	27.2	28.1
Qwen2 FS	11.5	15.6	14.3	28.9	28.0	28.4

Table 7: Main results for item level accuracy. ZS = zero-shot, FS = few-shot. The best-performing values for each metric are highlighted.

ItemAcc(%) on:	In-Person (n=16) English	In-Person (n=16) Mandarin	In-Person (n=16) All	Virtual (n=16) English	Virtual (n=16) Mandarin	Virtual (n=16) All
Llama3 ZS	39.8	28.0	31.7	23.2	23.7	23.4
Llama3 FS	34.5	27.6	29.8	15.9	12.9	14.3
GPT-4 ZS	**61.9**	49.2	**53.2**	34.1	37.6	36.0
GPT-4 FS	61.1	40.8	47.1	**40.2**	**40.9**	**40.6**
DeepSeek-V3 ZS	54.2	44.5	47.5	34.1	39.8	37.1
DeepSeek-V3 FS	57.5	**50.9**	53.0	39.0	38.7	38.9
Qwen2 ZS	20.4	16.0	17.4	28.0	20.4	24.0
Qwen2 FS	11.5	15.6	14.3	28.0	23.7	25.7

Table 8: Results with the entire In-person dataset and the subset of the Virtual dataset consisting of 16 patients who are matched with the 16 In-person patients in terms of child age and parent education level in item accuracy. The data layout follows the same procedure described in Table 7.

Micro F1 / Macro F1 (%) on:	In-Person (n=16) English	In-Person (n=16) Mandarin	In-Person (n=16) Overall	Virtual (n=16) English	Virtual (n=16) Mandarin	Virtual (n=16) Overall
Llama3 ZS	48.7/37.4	43.6/33.7	45.0/34.6	30.4/**21.7**	33.7/16.6	32.1/18.7
Llama3 FS	45.6/31.8	39.8/26.2	41.3/27.6	21.4/8.7	19.8/8.5	20.5/8.8
GPT-4 ZS	68.4/**61.6**	50.0/43.8	55.0/50.9	44.7/19.8	44.2/26.4	44.4/24.3
GPT-4 FS	63.3/58.6	56.0/51.1	58.0/53.4	43.4/19.0	47.7/31.4	45.6/27.7
DeepSeek-V3 ZS	61.7/53.5	58.5/51.2	59.4/52.4	44.0/19.5	51.7/**34.5**	48.0/**28.9**
DeepSeek-V3 FS	64.6/54.9	63.5/**56.2**	63.8/**56.3**	50.9/21.1	49.4/30.7	50.2/28.1
Qwen ZS	31.4/19.6	28.3/19.6	29.2/20.0	41.2/15.9	36.2/19.0	38.6/17.3
Qwen FS	20.1/15.2	25.6/21.5	24.1/19.7	45.7/16.8	38.6/25.9	42.0/22.6

Table 9: Results with the entire In-Person dataset and the subset of the Virtual dataset consisting of 16 patients who are matched with the 16 In-Person patients in terms of child age and parent education level. The data layout follows the same procedure described in Table 3.

```
Assume you are video analyst classifying transcribed text conversation shown in <Parent Behavior>
    from parents, who are supervising their bilingual children to answer language comprehension
    tests in English and Mandarin. A <Voiceover> is the system output sound which reads the picture
    on the website that the child has to choose. A <Child Behavior> is the children's behavior. You
    need to classify <Parent Behavior> as described in <task>.

Please respond the category name only.

<Task>
Based on <Parent Behavior>, please determine which type of behavior it is: 'Repeat Questions', '
    Answer Questions', 'Analyze Items', 'Judging', 'Encouragement', 'Technical Support', '
    Broadcasting', 'Miscellaneous'. Definition for each category is shown in <Definitions>

<Definitions>
- Repeat Questions: Repeating the <Voiceover> audio before and/or during the process of a child
    selecting the picture on the web.
- Answer Questions: Using verbal or gestural cues to suggest or select a correct answer for the child
    .
- Analyze Items: Elaborating on the critical linguistic components by labeling objects and actions,
    making emphasis via prosodic cues, or breaking down complex sentences from <Voiceover>.
- Judging of Correctness: Verbally evaluating the child's response as "correct" or "incorrect".

- Encouragement: Showing verbal and/physical affirmation for the child to continue, saying "good job/
    excellent" to reinforce the child's selection, expressing empathy (e.g., "it's okay") on
    struggled items.
- Technical Support: Offering verbal and/or physical assistance to the child related to interacting
    with the website and the computer.
- Broadcasting: After the child makes a selection, describing the selection via a word, a phrase, or
    a sentence.
- Miscellaneous: Initiating and/or responding to events that redirected a child's attention, sharing
    personal opinions about test procedures and stimuli, or other verbal and nonverbal behaviors
    that were out of the child's view.
```

Figure 3: The full zero-shot prompt used in our experiments.

	In-Person (n=16)			Virtual (n=43)		
BEHAVACC(%)/F1(%) on:	English	Mandarin	Overall	English	Mandarin	Overall
GPT-4 ZS	67.1/62.1	59.8/56.3	61.8/58.3	41.7/22.8	46.3/34.3	44.4/31.5

Table 10: Prediction results without role play description: BEHAVACC/F1 on assessment language for In-Person and Virtual dataset.

Child ID (In Person)	**Age (Year; Month)**	**Gender**	**Parent Education**	**Child ID (Virtual)**	**Age**	**Gender**	**Parent Education**
P1	3;8	F	Master	V1	4;0	M	Master
P2	3;8	M	Master	V2	4;1	F	Master
P3	4;1	F	PhD	V3	4;2	M	Master
P4	5;5	F	Master	V4	5;5	F	Master
P5	5;9	M	Master	V5	5;6	M	PhD
P6	5;11	F	Master	V6	5;9	F	Master
P7	6;1	M	PhD	V7	5;9	F	PhD
P8	6;1	M	Master	V8	5;10	F	Master
P9	6;3	M	PhD	V9	5;11	F	PhD
P10	6;4	F	Master	V10	6;3	F	PhD
P11	6;6	M	PhD	V11	6;5	M	Bachelor
P12	6;7	M	PhD	V12	6;11	M	High School
P13	6;7	M	PhD	V13	7;0	M	Bachelor
P14	7;9	M	Vocational School	V14	7;0	F	Bachelor School
P15	8;2	M	Bachelor	V15	8;2	M	Vocational
P16	8;6	M	PhD	V16	8;2	F	PhD

Table 11: Demographic information of gender, age, and parent education for in-person (P1–P16) and virtual study participants (V1–V16). These two groups of children were matched with the comparable level of parent education (except P12 with a parental education of PhD degree, and V12 with a parental education for high school). Childrens ages were also matched with no more than 9 months differences to ensure they are comparable in age for similar language abilities.

```
Assume you are video analyst classifying transcribed text conversation shown in <Parent Behavior> from parents, who are supervising their bilingual children to answer language comprehension tests in English and Mandarin. A <Voiceover> is the system output sound which reads the picture on the website that the child has to choose. A <Child Behavior> is the children's behavior. You need to classify <Parent Behavior> as described in <task>.

Please respond the category name only.

<Task>
Based on <Parent Behavior>, please determine which type of behavior it is: 'Repeat Questions', 'Answer Questions', 'Analyze Items', 'Judging', 'Encouragement', 'Technical Support', 'Broadcasting', 'Miscellaneous'.
Definition for each category is shown in <Definitions>

<Definitions>
- Repeat Questions: Repeating the <Voiceover> audio before and/or during the process of a child selecting the picture on the web. For example,

<Voiceover>
"the black cat is drinking water"

<Parent Behavior>
"the black cat is drinking water"

<Classification>
Repeat Questions


- Answer Questions: Using verbal or gestural cues to suggest or select a correct answer for the child. For example,

<Voiceover>
"What is the cat drinking?"

<Parent Behavior>
"Drinking water."

<Classification>
Answer Questions


- Analyze Items: Elaborating on the critical linguistic components by labeling objects and actions, making emphasis via prosodic cues, or breaking down complex sentences from <Voiceover>. For example,

<Voiceover>
"the black cat is drinking water"

<Parent Behavior>
"This is the one with a black cat."

<Classification>
Analyze Items
```

Figure 4: The few-shot prompt used in our experiments, part 1 of 2.

```
- Judging of Correctness: Verbally evaluating the child's response as "correct" or "incorrect". For
    example,

<Voiceover>


<Parent Behavior>
"This is not right."

<Classification>
Judging of Correctness


- Encouragement: Showing verbal and/physical affirmation for the child to continue, saying "good job/
    excellent" to reinforce the child's selection, expressing empathy (e.g., "it's okay") on
    struggled items. For example,

<Voiceover>


<Parent Behavior>
"it's fine you are trying your best."

<Classification>
Encouragement


- Technical Support: Offering verbal and/or physical assistance to the child related to interacting
    with the website and the computer. For example,

<Voiceover>


<Parent Behavior>
"Select the picture to continue."

<Classification>
Technical Support


- Broadcasting: After the child makes a selection, describing the selection via a word, a phrase, or
    a sentence. For example,

<Voiceover>


<Parent Behavior>
"I selected the picture."

<Classification>
Broadcasting


- Miscellaneous: Initiating and/or responding to events that redirected a child's attention, sharing
    personal opinions about test procedures and stimuli, or other verbal and nonverbal behaviors
    that were out of the child's view. For example,

<Voiceover>
<Parent Behavior>
"My child needs to use the bathroom."

<Classification>
Miscellaneous
```

Figure 5: The few-shot prompt used in our experiments, part 2 of 2.

Timestamp	Speaker	Test Item	Annotator (YD)	Annotator (YY)	GPT4-ZS	Final Agreement
0:20:48	Voiceover	The rabbit that is washing the cat is wearing a ribbon.	n/a	n/a	n/a	n/a
	Child	*<attempted to make a selection>*	n/a	n/a	n/a	n/a
	Parent	The rabbit washing a cat is wearing a robin.	RQ	RQ	RQ	RQ
	Parent	哪个 rabbit 在 wash the cat? Rabbit 在 wash the cat, 然后它还有 robin. Rabbit.	AI	AI	RQ	AI
	Child	*<moved her hand on top of another answer>*	n/a	n/a	n/a	n/a
0:21:18	Voiceover	The bird that is singing is pushing the turtle.	n/a	n/a	n/a	n/a
	Child	*<attempted to make a selection>*	n/a	n/a	n/a	n/a
	Parent	"The bird that is singing is pushing a turtle."	RQ	RQ	RQ	RQ
	Parent	哪个 bird 在 push turtle? 哪个 bird 在 singing?	AI	AI	RQ	AI
	Child	*<moved her hand on top of another answer>*	n/a	n/a	n/a	n/a
	Parent	" 哎对了。"	JC	JC	JC	JC

Figure 6: Sample annotation 1 from the In-person dataset (English item). N/A is used for transcript content that did not have a behavioral coding.

Timestamp	Speaker	Test Item	Annotator (YD)	Annotator (YY)	GPT4-FS	Final Agreement
20:11	Voiceover	"The fox points at the tree."	n/a	n/a	n/a	n/a
	Child	*<points to answer>*	n/a	n/a	n/a	n/a
	Parent	Wow! Good job!	E	E	E	E
		你要吃个 snack 吗还是你 okay 我們 keep going?	M	M	TS	M
		我們 keep going 了，好嗎?	TS	E	E	TS
		okay, let's keep going	TS	E	E	TS
20:41	Voiceover	"The corn is between the bottles."	n/a	n/a	n/a	n/a
	Child	"That one."	n/a	n/a	n/a	n/a
	Parent	" 一二還是三。"	TS	TS	TS	TS
	Child	" 一。"	n/a	n/a	n/a	n/a

Figure 7: Sample annotation 2 from the Virtual dataset (English item). N/A is used for transcript content that did not have a behavioral coding.

```
Assume you are a video analyst classifying transcribed text conversation shown in <Parent Behavior>
    from parents, who are supervising their bilingual children to answer language comprehension
    tests in English and Mandarin. A <Voiceover> is the system output sound which reads the picture
    on the website that the child has to choose. A <Child Behavior> is the children's behavior. You
    need to classify <Parent Behavior> as described in <task>. Let's think stey by step.

<Task>
Based on <Parent Behavior>, please determine which type of behavior it is: 'Repeat Questions', '
    Answer Questions', 'Analyze Items', 'Judging', 'Encouragement', 'Technical Support', '
    Broadcasting', 'Miscellaneous'. Definition for each category is shown in <Definitions>.

<Definitions>
- Repeat Questions: Repeating the <Voiceover> audio before and/or during the process of a child
    selecting the picture on the web.
- Answer Questions: Using verbal or gestural cues to suggest or select a correct answer for the child
    .
- Analyze Items: Elaborating on the critical linguistic components by labeling objects and actions,
    making emphasis via prosodic cues, or breaking down complex sentences from <Voiceover>.
- Judging of Correctness: Verbally evaluating the child's response as correct or incorrect.
- Encouragement: Showing verbal and/physical affirmation for the child to continue, saying good job/
    excellent to reinforce the child's selection, expressing empathy (e.g., it'sokay) on struggled
    items.
- Technical Support: Offering verbal and/or physical assistance to the child related to interacting
    with the website and the computer.
- Broadcasting: After the child makes a selection, describing the selection via a word, a phrase, or
    a sentence.
- Miscellaneous: Initiating and/or responding to events that redirected a child's attention, sharing
    personal opinions about test procedures and stimuli, or other verbal and nonverbal behaviors
    that were out of the child's view.
```

Figure 8: The first zero-shot prompt used in Chain of Thoughts experiment.

```
Assume you are video analyst classifying transcribed text conversation shown in <Parent Behavior>
    from parents, who are supervising their bilingual children to answer language comprehension
    tests in English and Mandarin. A <Voiceover> is the system output sound which reads the picture
    on the website that the child has to choose. A <Child Behavior> is the children's behavior. You
    need to classify <Parent Behavior> as described in <task>. Let's think step by step.

<Task>
Based on <Parent Behavior>, please determine which type of behavior it is: 'Repeat Questions', '
    Answer Questions', 'Analyze Items', 'Judging', 'Encouragement', 'Technical Support', '
    Broadcasting', 'Miscellaneous'.
Definition for each category is shown in <Definitions>

<Definitions>
- Repeat Questions: Repeating the <Voiceover> audio before and/or during the process of a child
    selecting the picture on the web. For example,

<Voiceover>
"the black cat is drinking water"

<Parent Behavior>
"the black cat is drinking water"

<Classification>
Repeat Questions


- Answer Questions: Using verbal or gestural cues to suggest or select a correct answer for the child
    . For example,

<Voiceover>
"What is the cat drinking?"

<Parent Behavior>
"Drinking water."

<Classification>
Answer Questions


- Analyze Items: Elaborating on the critical linguistic components by labeling objects and actions,
    making emphasis via prosodic cues, or breaking down complex sentences from <Voiceover>. For
    example,

<Voiceover>
"the black cat is drinking water"

<Parent Behavior>
"This is the one with a black cat."

<Classification>
Analyze Items
```

Figure 9: The first few-shot prompt used in Chain of Thoughts experiment, part 1 of 2.

```
- Judging of Correctness: Verbally evaluating the child's response as "correct" or "incorrect". For
    example,

<Voiceover>


<Parent Behavior>
"This is not right."

<Classification>
Judging of Correctness


- Encouragement: Showing verbal and/physical affirmation for the child to continue, saying "good job/
    excellent" to reinforce the child's selection, expressing empathy (e.g., "it's okay") on
    struggled items. For example,

<Voiceover>


<Parent Behavior>
"it's fine you are trying your best."

<Classification>
Encouragement


- Technical Support: Offering verbal and/or physical assistance to the child related to interacting
    with the website and the computer. For example,

<Voiceover>


<Parent Behavior>
"Select the picture to continue."

<Classification>
Technical Support


- Broadcasting: After the child makes a selection, describing the selection via a word, a phrase, or
    a sentence. For example,

<Voiceover>


<Parent Behavior>
"I selected the picture."

<Classification>
Broadcasting


- Miscellaneous: Initiating and/or responding to events that redirected a child's attention, sharing
    personal opinions about test procedures and stimuli, or other verbal and nonverbal behaviors
    that were out of the child's view. For example,

<Voiceover>
<Parent Behavior>
"My child needs to use the bathroom."

<Classification>
Miscellaneous
```

Figure 10: The first few-shot prompt used in Chain of Thoughts experiment, part 2 of 2.

```
Therefore, among all the categories, please respond the category name only.
```

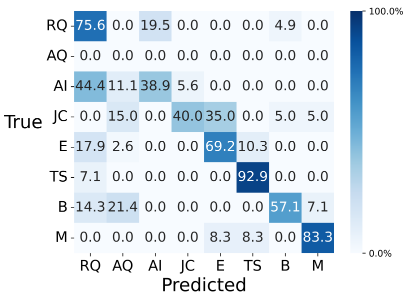
Figure 11: The second prompt used in Chain of Thoughts experiment for both zero-shot and few-shot.

```
This is a text transcript consisting of <Parent Behavior> from bilingual parents who are supervising their bilingual children to answer language assessment questions in English and Mandarin. A <Voiceover> is the system output sound which reads the picture for the accurate answer the child has to choose. A <Child Behavior> is the children's behavior. Please classify <Parent Behavior> as described in <Task> and respond to the category name only.
<Task>
Based on <Parent Behavior>, please determine which type of behavior it is: 'Repeat Questions', 'Answer Questions', 'Analyze Items', 'Judging', 'Encouragement', 'Technical Support', 'Broadcasting', 'Miscellaneous'. Definition for each category is shown in <Definitions>.

<Definitions>
- Repeat Questions: Repeating the <Voiceover> audio before and/or during the process of a child selecting the picture on the web.
- Answer Questions: Using verbal or gestural cues to suggest or select a correct answer for the child.
- Analyze Items: Elaborating on the critical linguistic components by labeling objects and actions, making emphasis via prosodic cues, or breaking down complex sentences from <Voiceover>.
- Judging of Correctness: Verbally evaluating the child's response as correct or incorrect.
- Encouragement: Showing verbal and/physical affirmation for the child to continue, saying good job/excellent to reinforce the child's selection, expressing empathy (e.g., it's okay) on struggling items.
- Technical Support: Offering verbal and/or physical assistance to the child related to interacting with the website and the computer.
- Broadcasting: After the child makes a selection, describing the selection via a word, a phrase, or a sentence.
- Miscellaneous: Initiating and/or responding to events that redirected a child's attention, sharing personal opinions about test procedures and stimuli, or other verbal and nonverbal behaviors that were out of the child's view.
```

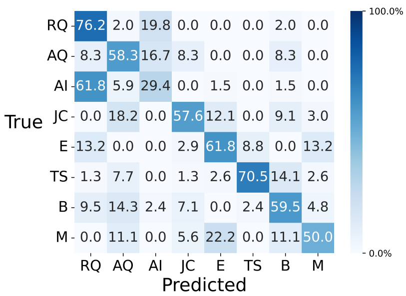
Figure 12: The full zero-shot prompt used in non role play test.

```
Parent behavior: <in response to child> "" <speaking towards the screen> "

GPT-4 with CoT response:
"let's analyze the provided <parent behavior> step by step in order to classify it:

1. the first part of the <parent behavior> is: ""<in response to child> ''"". this is in response to the child's action of moving the parent's hand to the screen. the parent's response, """" or "" you click it yourself"", seems to encourage the child to perform the action by themselves instead of relying on the parent. this can be classified under 'encouragement', as it encourages independence in the child.
2. the second part of the <parent behavior> is: ""<speaking towards the screen> ''"", which translates to ""can you say that again, please?"" directed towards the screen. this part indicates that the parent is requesting the system to repeat the <voiceover> audio. this action fits the 'repeat questions' category, as it involves asking for the question or statement from <voiceover> to be repeated. given the two different actions, if we consider the more prominent behavior from the parent's actions, the emphasis appears to be on encouraging the child to engage with the task independently. however, the explicit category from the definitions provided that matches an exact behavior is 'repeat questions', due to the direct request made to the system. <classification> repeat questions"
```

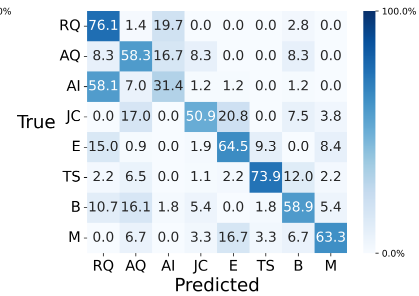
Figure 13: The first zero-shot prompt used in Chain of Thoughts experiment.



English

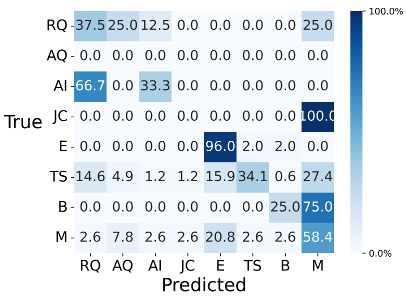


Mandarin

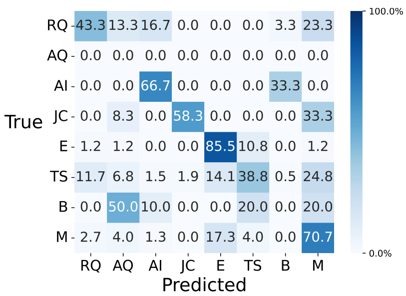


Overall

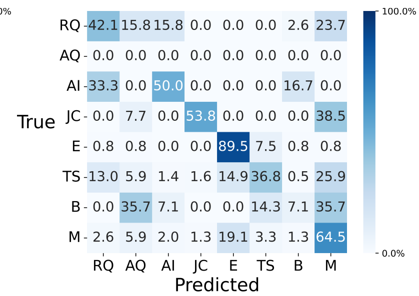
GPT-4 ZS: In-Person dataset.



English



Mandarin



Overall

GPT-4 FS: Virtual dataset.

Figure 14: Normalized confusion matrices (in percentages) for GPT-4 ZS (In-Person) and GPT-4 FS (Virtual) datasets. Each row is normalized to sum to 100% within each matrix, representing the percentage distribution of predictions across classes.