Automatic Classification of Parental Behaviors in Bilingual Datasets from In-Person and Telehealth Language Assessment

Anonymous ACL submission

Abstract

Conducting text-based behavioral coding is a labor-intensive process for clinicians, particularly when annotating complex bilingual data. This study evaluates the performance of four state-of-the-art (SOTA) large language models (LLMs) in automating the classification of parent behaviors within a bilingual dataset comprising 59 Mandarin-English child language assessment sessions (16 in-person and 43 telehealth). While the four LLMs - GPT-4, Llama-3, Qwen2, and DeepSeek-V3 - achieved notable accuracy, they still fall short of the performance of bilingual human annotators. Additional error analysis revealed that both human annotators and the generally best-performing model, GPT-4, faced challenges in classifying parental behaviors in categories involving complex task procedures, especially when analyzing bilingual code-mixed text. This study contributes to the understanding of how LLMs can be utilized to advance the automated classification of behavioral coding in bilingual child language assessments.

1 Introduction

Language assessment is a standardized clinical procedure to evaluate children's communication abilities and detect potential language delays and disorders for early intervention (Wang et al., 2020; Gorman et al., 2015; Wang et al., 2024). During assessments, clinicians administer linguistic tasks to children and gather children's language development information from parents (Klatte et al., 2020; Sheng et al., 2021; Pratt et al., 2022). However, assessing bilingual children requires conducting the procedure in two languages, increasing the workload of already scarce bilingual clinicians (Du et al., 2020).

The rise in telehealth during and after the COVID-19 pandemic gave parents and children easier access to care and diverse ways to receive language assessment (Pratt et al., 2022; Dam and Pham, 2023). However, since clinicians are not physically positioned with parent-child dyads during synchronous telehealth, parents often have to facilitate critical clinician-child interactions (Pozniak et al., 2024), such as providing encouragement and technical support for children to interact with computers (e.g., navigating the website) and videoconferencing softwares (e.g., using different controls on Zoom) (Fissel et al., 2015; Edwards-Gaither et al., 2023). However, lacking professional skills in assessment procedures (Tomlinson et al., 2018), some parents may engage in *interference behaviors* (e.g., repeating or analyzing testing questions) that affect children's performance and compromise assessment validity (Du et al., 2020). Identification of behaviors requires clinicians' manual transcription and behavioral coding from video-recorded sessions (Sun et al., 2024; Cao et al., 2019), which can be extremely time-consuming (Lønfeldt et al., 2023).

Prior NLP research has leveraged large language models (LLMs) to automate behavioral coding in tasks such as motivational interviewing for counseling (Cao et al., 2019; Tavabi et al., 2020; Mayer et al., 2024; Pellemans et al., 2024); however, majority of these studies primarily focus on monolingual adult patients during in-person contexts. Limited studies have applied NLP approaches and LLMs to both in-person and telehealth contexts with bilingual parents and children (Zhang et al., 2023a,b; Lin et al., 2022) using clinical child language assessment tasks (Karacan et al., 2024).

This paper uses LLMs to automate the classification of behavioral coding of bilingual parents as they support their children to interact with a web-based Mandarin-English language assessment. We collect and release a bilingual Mandarin-English dataset of conversational transcripts and behavior descriptions from in-person and virtual video recordings of two groups of 59 parent-child dyads. The dataset includes 1,304 total parent behaviors (In-Person dataset: 578; Virtual dataset: 726), annotated with one of eight fine-grained labels. The eight categories constituted four supportive and four interference behavioral subcategories, based on an established clinical annotation guideline developed in collaboration with domain experts.

This dataset serves as the benchmark for classifying parental behaviors using four state-of-the-art (SOTA) LLMs - GPT-4, Llama 3, Qwen2, and DeepSeek-V3 - via both zero-shot and few-shot prompting strategies (Brown et al., 2020; Lin et al., 2022). While GPT-4 and DeepSeek-V3, the stronger model, performs reasonably well on the dataset above the other models, its accuracy still falls short of human expert evaluation. For example, Mandarin utterances by parents posed significant challenges for these models (except DeepSeek-V3), highlighting the need for improved multilingual modeling to enhance LLMs' performance for bilingual datasets.

To the best of our knowledge, this is the first comprehensive bilingual Mandarin-English code-mixing dataset for classifying parent behaviors during child language assessment. The study reveals the weaknesses of SOTA LLMs and presents a challenging, ecologically valid, bilingual benchmark to understand the application of NLP approaches for child language assessment tasks. Advancement on this topic can improve workflow efficiency for clinicians and clinical researchers to better understanding the complex parent-child dyadic interaction during in-person and telehealth settings.

2 Related Work

2.1 Multilingual LLMs for Real-World Tasks

LLMs like GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023a,b) have shown impressive performance, in tasks like question answering and logical reasoning (Wei et al., 2021; Sanh et al., 2021; Chung et al., 2022). Additionally, Qwen2 (Yang et al., 2024a) and DeepSeek (Bi et al., 2024; Peng et al., 2025; Neha and Bhati, 2025), trained in English and Mandarin multilingual data, excels in various benchmarks. Performance differences may stem from language features (e.g., morphology, grammar) (Hlavnova and Ruder, 2023; Weissweiler et al., 2023) or multimodal information during prompting (Yang et al., 2024b). Relevant to this work, in-context learning (ICL) (Brown et al., 2020; Zhang et al., 2022; Rubin et al., 2022; Li et al., 2023) is a common prompting strategy, allowing models to learn from few-shot examples. Recent studies have explored LLMs' capabilities in real-world scenarios that require domain expertise, such as children's education (Chen et al., 2023) and medical domains.

2.2 Annotating Clinical Assessment Data

Prior NLP research has focused on automating educational and clinical tasks, such as pediatric language assessment (Wang et al., 2020; Gorman et al., 2015), behavioral coding and testing for outcome prediction (Van Aken et al., 2021; Sun et al., 2024; Cao et al., 2019; Yang et al., 2023), and generating novel cognitive test items (Laverghetta Jr and Licato, 2023), and narrative tasks (Prud'hommeaux and Roark, 2015; Chen et al., 2023). However, applying NLP to telehealth encounters using bilingual datasets remained limited due to difficulties in accessing patient data and the high cost of human annotations (Chen et al., 2022). Differing from conventional annotation tasks, since clinical assessment tasks may directly impact the diagnostic accuracy of patient care, the annotation also needs to be annotated accurately following required psychometric standards (Abbasi et al., 2021) to ensure assessment validity and reliability. Therefore, to improve clinician workflow efficiency and accuracy, novel approaches need to be developed with clinically informed guidelines to address clinical needs for assessment tasks.

2.3 Behavioral Coding in Clinical NLP Research

Behavioral coding is a common data analysis methodology in social science research (Wang et al., 2022; Black et al., 2013) and has been widely adopted in public health and clinical research. A large body of NLP literature has examined ways to reduce traditional manual coding (Leeson et al., 2019) including using models such as BERT (Tavabi et al., 2020) and techniques such as topic modeling and Word2Vec (Leeson et al., 2019) to automatically classify detailed patient-provider interaction. Computational researchers also explored automatic speech recognition (Pérez-Rosas et al., 2021) and signal processing of speech data (Narayanan and Georgiou, 2013) to advance automated behavioral analysis.

While these approaches show promise in improving the efficiency of coding processes and enhancing the accuracy of behavioral predictions, they primarily focus on adult speakers during procedures such as motivation interviews in counseling and psychotherapy (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, results may not be easily generalized to language-dependent tasks. Conducting a child language assessment requires more fine-grained coding for analyzing linguistic features in addition to assessment behaviors (Wang et al., 2020; Gorman et al., 2015), which increased the level of complexity fo analysis. Additionally, prior studies also overlooked the complexity of bilingual interactions, particularly in telehealth settings. This present study examines behavioral coding using a bilingual dataset with both parent-child dyadic interaction and a comparison of in-person and virtual telehealth sessions, adding a unique use case to this body of literature.

3 Bilingual Dataset

We collaborated with bilingual Mandarin-English speaking speech language pathologists (SLP) and researchers to obtain an IRB-approved text-based dataset of child language assessment sessions containing 59 parent-child dyads using the Mandarin-English Receptive Language Screener (MERLS). MERLS is an audiovisual 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. The MERLS interface plays audio instructions for bilingual children to select pictures that match the instructions (see Figure 1). Children can independently complete MERLS with minimal parental assistance.

3.1 Data Collection

The MERLS dataset comprises an in-person dataset (n=16) collected in person via a video recorder placed behind the parent-child dyads, and a telehealth dataset (n=43) collected virtually via webcams through Zoom during the COVID-19 pandemic. Combining the two datasets, there is a total of 32 parent-child pairs (16 in-person and 16 virtual) that are matched in parent education and childrens age within 6 months of differences. This enables researchers to conduct further analysis to compare in-person and telehealth efficacy using two matching groups of participants, a gold standard practice common in clinical research.

3.2 Annotation Process

To annotate the dataset, two bilingual research assistants first transcribed all the parent-child interaction videos for speaker utterances verbatim and then annotated parents' verbal and non-verbal behaviors (e.g., gestures) for all English and Mandarin items. Next, two bilingual clinical experts independently coded all parent behaviors using an established video analysis codebook (Du et al., 2020) developed via Clinical Discourse Analysis (Damico, 1985). To reach a consensus between annotators, disagreement between annotators was resolved via the member checking method (Birt et al., 2016) by meeting and discussing disagreement coding, resolving the disagreement, identifying the accurate categories, and then refining the codebook with a better definition. Interobserver agreement (IOA) was calculated between the two annotators by comparing coding agreements over all behaviors per transcript, then averaged across all transcripts. To calculate IOA agreement for performing behavioral coding tasks, extensive training are required for human experts. IOA reached 97% (in-person dataset) and 86.1% (virtual dataset) illustrated in Table 3.

The NLP task is an eight-class classification problem which aims to categorize all types of parental involvement during assessments, with each behavior assigned one correct label. Each input includes the current test item, a description of the child's actions, and a description of the parent's behavior. Table 1 illustrated the two primary classes and four corresponding sub-categories for parent behaviors. "Interference" behaviors represent incidents when parents negatively impacted the assessment including "Repeating Questions (RQ)", "Answering Questions (AQ)", "Analyzing Items (AI)", and "Judging of Correctness (JC)", whereas "Support" behaviors represent incidents when parents positively facilitated the assessment including "Encouragement (E)", "Technical Support (TS)", "Broadcasting (B)", and "Miscellaneous (M)".

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 behaviors	79	214	293	12	45	57
# Support behaviors	79	206	285	295	374	669
# Behaviors	158	420	578	307	419	726
# Items ≥ 1 behaviors	113	250	363	187	243	430

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.

3.3 Dataset Description & Statistics

The dataset is structured to include the following components in English and Mandarin tests:

1. **Time stamps**: Precise time stamps for each assessment item and corresponding parent-child behavior from each audio recording.
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 each voiceover.
4. **Behavior descriptions**: Textual descriptions of parents' verbal and non-verbal behaviors.

Table 2 presents overall statistics for the In-person and Virtual datasets partitioned by classes. The two datasets exhibit imbalances in their label distributions: the Virtual dataset contains fewer interference behaviors and more technical support behaviors. It may be due to (1) the system redesign of the MERLS website before the collection of the Virtual dataset (e.g., adding an instructional video about prohibited interference behaviors), or (2) the use of Zoom for Virtual data collection added additional technical behaviors from parents who took the majority of technical support activities to help their children without direct in-person help from clinicians.

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 compared to clinical expert.

Prompts. Our zero-shot prompt in Figure 2 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 3 and 4 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

BehavAcc(%)/F1(%) on:	In-Person (n=16) English	Mandarin	Overall	Virtual (n=43) English	Mandarin	Overall
Llama3 ZS	48.7/37.4	43.6/33.7	45.0/34.6	29.3/21.9	31.3/20.8	30.4/21.0
Llama3 FS	45.6/31.8	39.8/26.2	41.3/27.6	22.1/9.32	23.6/10.1	23.0/9.90
GPT-4 ZS	**65.8**/58.5	60.5/55.7	61.9/**57.2**	45.3/22.3	51.6/**38.2**	48.9/33.8
GPT-4 FS	65.5/**61.1**	55.0/49.1	58.1/52.4	48.5/**28.8**	52.5/**38.2**	50.8/**36.2**
DeepSeek-V3 ZS	61.4/53.5	58.3/51.2	59.1/52.4	47.6/23.6	53.2/38.1	50.8/34.3
DeepSeek-V3 FS	64.2/54.9	**63.2/56.2**	**63.5**/56.3	**52.4**/25.3	**53.7**/36.2	**53.2**/33.3
Qwen2 ZS	31.0/19.6	27.1/19.6	28.2/20.0	36.5/13.9	38.7/20.3	37.7/17.8
Qwen2 FS	17.7/15.2	22.6/21.5	21.3/19.7	36.2/17.0	37.9/24.2	37.2/22.1
Human Experts	96.84	96.43	97.0	86.93	81.82	86.1

Table 3: BehavAcc/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.

BehavAcc(%)/F1(%) on:	In-Person (n=16) English	Mandarin	Mixed	Virtual (n=43) English	Mandarin	Mixed
# Examples	233	167	178	226	389	111
Llama3 ZS	39.5/31.2	42.5/22.0	53.4/39.6	44.2/27.5	21.1/18.1	35.1/16.5
Llama3 FS	34.8/25.7	40.1/20.8	49.4/33.1	32.7/12.1	14.7/6.91	32.4/14.6
GPT-4 ZS	**64.8/54.4**	55.1/**45.9**	62.9/**59.0**	56.6/34.4	43.2/**33.2**	**53.2/32.0**
GPT-4 FS	57.1/53.3	52.1/38.9	63.5/56.6	60.2/**41.6**	45.5/**33.2**	50.5/31.8
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	33.5/18.6	28.7/17.0	20.8/16.6	38.5/15.2	37.3/18.8	37.8/12.9
Qwen2 FS	23.6/17.9	21.6/18.0	18.0/16.6	39.4/21.9	32.9/19.7	47.7/27.1

Table 4: BehavAcc/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.

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 and Qwen2-7B, the closed-source model GPT-4 (Turbo-2024-04-09), and the open-source model DeepSeek-V3. While Llama-3 is primarily English-based, its pre-training data includes data from 30 other languages.[1] Qwen2 (Yang et al., 2024a) improves upon Qwen1.5, achieving 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 an open-source model optimized for computational efficiency and excels in complex linguistic and reasoning tasks with minimal supervised data (Neha and Bhati, 2025).

Evaluation metrics. To accurately evaluate parent behaviors across different test items in Mandarin and English, we compute three metrics: (1) Behavior-level Accuracy (BehavAcc): the fraction of correctly predicted behaviors; (2) Macro F1 score (F1): prediction performance addressing the effects of dataset imbalance; (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 overall accuracies and macro F1 scores of all models. DeepSeek-V3 and GPT-4 significantly outperforms Llama3 and Qwen2 across all datasets. The performance of GPT-4 and DeepSeek-V3 is highly comparable across both the In-Person and Virtual datasets. For the In-Person dataset, GPT-4 outperforms the other models in English assessment items, achieving a BehavAcc of 65.8%, while DeepSeek-V3 deliv-

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

BehavAcc(%)/F1 on:	In-Person (n=16) English	Mandarin	Overall	Virtual (n=16) English	Mandarin	Overall
Llama3 ZS	48.7/37.4	43.6/33.7	45.0/34.6	30.2/21.7	33.1/16.6	31.7/18.7
Llama3 FS	45.6/31.8	39.8/26.2	41.3/27.6	21.4/8.70	19.8/8.51	20.5/8.82
GPT-4 ZS	**65.8**/58.5	60.5/55.7	61.9/**57.2**	46.5/19.0	44.8/32.5	45.6/28.6
GPT-4 FS	65.5/**61.1**	55.0/49.0	58.1/52.4	47.8/**25.5**	45.9/33.6	46.8/**31.0**
DeepSeek-V3 ZS	61.4/53.5	58.3/51.2	59.1/52.4	44.0/19.5	**51.7/34.5**	48.0/28.9
DeepSeek-V3 FS	64.2/54.9	**63.2/56.2**	**63.5**/56.3	**50.9**/21.1	49.4/30.7	**50.2**/28.1
Qwen2 ZS	31.0/19.6	27.1/19.6	28.2/20.0	38.4/15.9	34.3/19.0	36.3/17.3
Qwen2 FS	17.7/15.2	22.6/21.5	21.3/19.7	40.3/16.8	34.3/25.9	37.2/22.6

Table 5: 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.

BehavAcc(%)/F1 on:	In-Person (n=16) Interference	Support	Overall	Virtual (n=43) Interference	Support	Overall
Llama3 ZS	80.5/75.2	65.3/70.5	73.0/72.8	53.3/18.0	64.4/76.4	63.5/47.2
Llama3 FS	82.9/73.5	56.1/64.6	69.7/69.1	70.9/20.3	54.3/69.3	55.6/44.8
GPT-4 ZS	**91.8/87.9**	82.5/**86.4**	**87.2/87.2**	71.9/32.5	77.0/85.8	76.6/59.2
GPT-4 FS	91.1/86.7	80.4/84.8	85.8/85.8	71.9/**40.0**	84.0/90.1	83.1/**65.1**
DeepSeek-V3 ZS	73.3/78.0	86.4/81.4	80.0/79.7	38.6/20.4	79.5/86.1	76.3/53.2
DeepSeek-V3 FS	84.4/85.0	**86.8**/85.9	85.6/85.5	43.9/29.2	**86.7/90.6**	**83.3**/59.9
Qwen2 ZS	50.3/54.7	75.4/66.0	62.5/60.4	69.1/25.5	80.5/87.0	79.1/56.2
Qwen2 FS	64.2/56.3	59.0/57.4	61.9/56.9	**86.3**/28.2	75.6/84.7	77.5/56.4

Table 6: BehavAcc/F1 on the binary classification version ("Interference" vs. "Support" behaviors) of In-Person and Virtual dataset. ZS = zero-shot, FS = few-shot.

ers the best performance in Mandarin assessment items with a BehavAcc of 63.2%. In contrast, on the Virtual dataset, GPT-4 surpasses DeepSeek-V3 in all F1 scores, although it demonstrates a lower BehavAcc. Here DeepSeek-V3's impressive performance in Mandarin is highly likely attributed to its training on high-quality articles, which has greatly enhanced its understanding of the Chinese language (Guo et al., 2024). However, this still remains well below human expert accuracy (97.0% and 86.1% for In-Person and Virtual dataset, respectively), indicating substantial room for improvement. Table 5 shows similar trends on a subset of 16 matched Virtual and In-Person pairs: DeepSeek-V3 consistently demonstrates its remarkable capabilities in Mandarin assessment items. On the other hand, Qwen2 lags behind, with its FS performance achieving merely 17.7% BehavAcc on English assessment items just above the random chance baseline, whereas its performance in Mandarin is slightly stronger.

Overall, the Virtual dataset is consistently more difficult to classify than the In-Person dataset, as demonstrated in Table 3 and 5. This is likely due to limitations information captured via the Zoom recording for virtual sessions compared to the more comprehensive in-person sessions. These differences also influenced human annotation during transcription.

In addition, a clear trend shows that the macro F1 score is consistently much lower than BehavAcc, highlighting the imbalanced distribution across the eight categories. As shown in Table 2, categories such as "Encouragement" and "Technical Support" are overrepresented, while the "Analyzing Items" behavior appears only three times in the Virtual dataset. The results of ItemAcc (Table 8 and Table 9 in Appendix A.2) indicate that the best prediction results for ItemAcc are also achieved by GPT-4 and DeepSeek-V3.

5.2 Effects of Parent Language

We examine whether the language used to describe parents behavior impacts the LLM's accuracy. Descriptions can be in English, Mandarin, or a mix of both. In our dataset, non-verbal actions are always described in English, while many parent speech acts are in Mandarin. Code-mixing occurs when parents code-switch, or when Mandarin speech is paired with an English description of a non-verbal action.

Table 4 shows model accuracies and F1 score broken down by the language describing the parent behavior. The F1 score is notably lower than

BEHAVACC, largely due to the imbalanced distribution of parent language occurrences. All models generally perform worst on Mandarin-only transcriptions, except for Llama3 on the In-Person dataset. Qwen2 shows a smaller performance gap between English and Mandarin transcriptions, as expected due to its focus on non-English performance. Deepseek-V3 shows the best performance for Mandarin-only transcriptions in both In-Person and Virtual datasets, indicating its key strength in Chinese language processing. The increased difficulty of classification in the Virtual dataset is explained primarily due to behaviors involving Mandarin. For DeepSeek-V3, the English Virtual dataset is easier than the English In-Person dataset to predict, whereas the other three models show the opposite pattern.

5.3 Binary Classification Results

We also evaluate models on the binary classification task to distinguish interference from support behaviors. Identifying interference can help alert clinicians to potential issues, even if the model cannot identify the individual type of interference. As shown in Table 6, GPT-4 and DeepSeek-V3 substantially outperform Llama3 and Qwen2, with GPT-4 exhibiting slightly superior performance compared to DeepSeek-V3. In contrast, Qwen2 struggles, with approximately 50% BEHAVACC in the ZS setting for the In-Person dataset when predicting interference behaviors. For the Virtual dataset, Qwen2 FS performs best at predicting "Interference" behaviors. This likely due to the imbalanced data between "Interference" and "Support" behaviors as reflected in the low F1 score.

5.4 Error Analysis with Human Annotators

Based on the overall BEHAVACC and F1 score in Table 3 and Table 6, we conducted a detailed error analysis 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 13 identified the misclassified pairs. To examine these errors, we selected the most frequently misclassified pairs for each behavioral category within each dataset (Table 7). 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 7. 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 5 in the appendix 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 happened in code-mix utterances and could be due to inadequate translation from word-level lexicon to sentence-level utterances. By analyzing the disagreements between human coders and predictions from LLMs, we gain a deeper understanding of these distinctions to improve 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 7. We found the challenging pairs for clinicians to classify are “TS-M” (“Technical Support" vs. “Miscel-

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 7: Clinician annotation accuracy based on the misclassified pairs from Figure 13. 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.

laneous") for Virtual English and “M-E” (“Miscellaneous" vs. “Encouragement") for Virtual dataset in Mandarin. Figure 6 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 the inconsistencies in annotation, they identified additional utterances (e.g., parents monitoring children’s needs to take a break or eat a snack) 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 not only identify existing categories accurately but also recognize novel patterns to enhance clinicians’ manual 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. The detailed results are provided in Appendix B.

6 Conclusion

This paper introduces a bilingual dataset for fine-grained classification of parental behaviors during bilingual English-Mandarin child language assessment. Automating this task could increase clinicians’ workflow efficiency and expand the use of LLMs for behavioral coding in clinical settings. While current SOTA LLMs show moderate accuracy, they struggle with Mandarin data, a challenge also faced by human annotators, particularly with virtual data. This dataset promotes further NLP research for multilingual clinical tasks, advancing the analysis of using multimodal behavioral coding (Yang et al., 2024b) of bilingual datasets (Hlavnova and Ruder, 2023; Weissweiler et al., 2023) during child language assessment in complex 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 the clinical team could be prone to errors and inconsistency. To reduce such errors, future work should continue exploring fine-tuning our 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 with clinicians, future work will continue a more comprehensive quantitative analysis of this dataset.

8 Limitations

Our study is constrained by the imbalance between the in-person and virtual datasets, as well as a relatively small sample size, which is further limited by the data provided by our clinical partners. 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 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 under university human subject research approval and data sharing agreements. The de-identified text transcripts from the clinical video analysis 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.

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*.

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*.

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.

Schea N Fissel, Pamela R Mitchell, and Robin L Alvares. 2015. An adapted assessment model for emergent literacy conducted via telepractice. *Perspectives on Telepractice*, 5(2):48–56.

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.

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.

Inge S Klatte, Rena Lyons, Karen Davies, Sam Harding, Julie Marshall, Cristina McKean, and Sue Roulstone. 2020. Collaboration between parents and slts produces optimal outcomes for children attending speech and language therapy: Gathering the evidence. *International Journal of Language & Communication Disorders*, 55(4):618–628.

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.

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.

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.

Fnu Neha and Deepshikha Bhati. 2025. A survey of deepseek models. TechRxiv. Accessed: 2025-02-07.

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.

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.

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, et al. 2024a. 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. 2024b. 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 1 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.

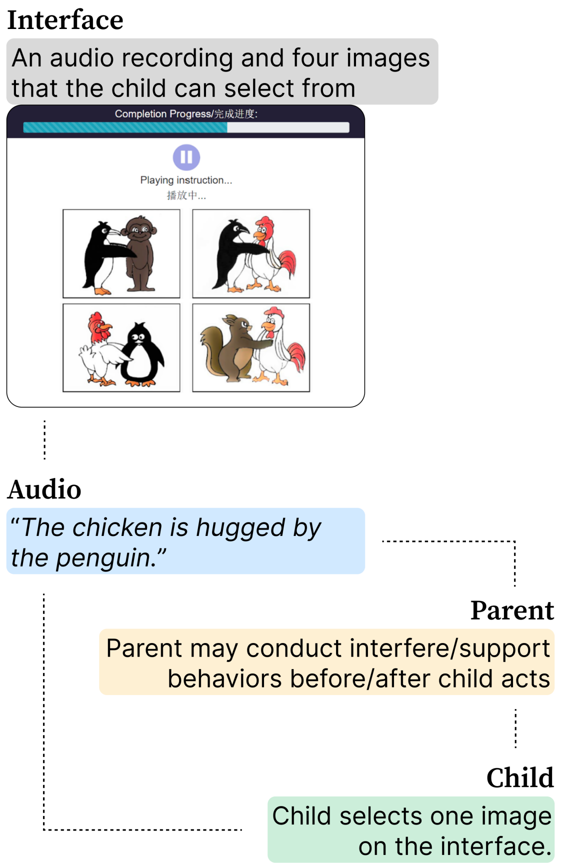


Figure 1: MERLS English test item “The chicken is hugged by the penguin.”

A.2 Results of item level accuracy

Here we also provide the accuracy in item level. Table 8 shows the ItemAcc for the entire In-Person dataset (n=16) and the entire Virtual dataset (n=43) partitioned upon question languages, whereas Table 9 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 2, but without the role description (e.g., "Assume you are a video analyst classifying transcribed text conversation...") (see prompt in Figure 11). 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 7, 8 and 9 for the first prompt input in the Appendix). In the second prompt (see Figure 10 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 12. 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.

	In-Person (n=16)			Virtual (n=43)		
ItemAcc(%) on:	English	Mandarin	All	English	Mandarin	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 8: Main results for item level accuracy. ZS = zero-shot, FS = few-shot. The best-performing values for each metric are highlighted.

	In-Person (n=16)			Virtual (n=16)		
ItemAcc(%) on:	English	Mandarin	All	English	Mandarin	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 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 in item accuracy. The data layout follows the same procedure described in Table 8.

```
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 2: The full zero-shot prompt used in our experiments.

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

	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.

```
- 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 4: 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 5: 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 6: 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 7: 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 8: 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 9: 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.
```

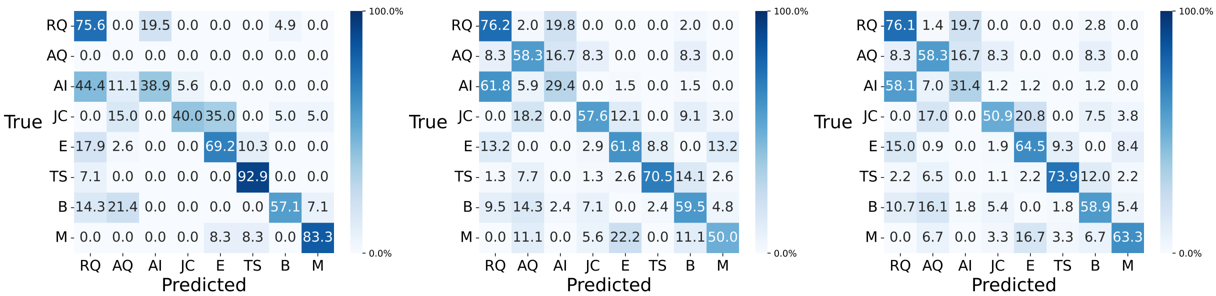
Figure 10: The second prompt used in Chain of Thoughts experiment for both zero-shot and few-shot.

```
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 11: The full zero-shot prompt used in non role play test.

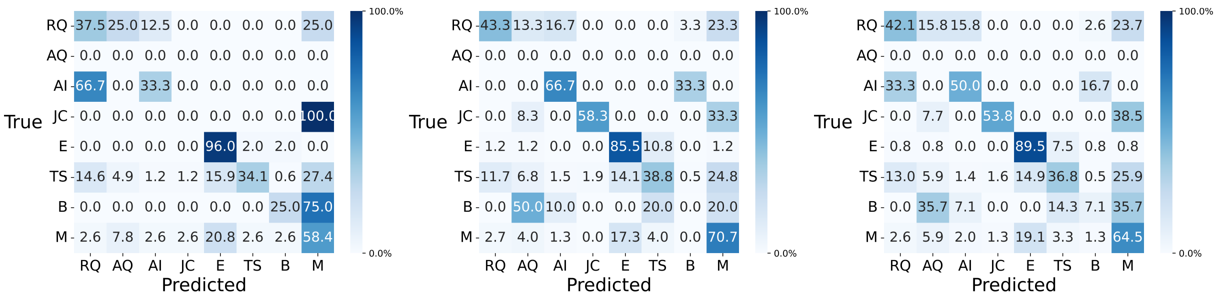
```
Parent behavior: <in response to child> "" <speaking towards the screen> "

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 12: 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 13: 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.