Automatic Detection of Parental Interference Behaviors during Bilingual Child Language Assessment

Anonymous ACL submission

Abstract

Recent clinical research has developed novel protocols that enable children to participate in bilingual language assessment remotely with parents to assist in this process. However, since parents are not trained clinicians, they often perform *interference behaviors*—actions that could compromise the validity of the assessment (e.g., providing hints). In this paper, we study whether language models can help automate the detection and categorization of parental interference behaviors during bilingual English-Mandarin child language assessment. Such a system would reduce the burden on clinicians, who must otherwise rely on transcribing video recordings and checking them manually for signs of interference. We release a new, expert-annotated dataset for this task, and evaluate multiple state-of-the-art large language models. While these models achieve non-trivial accuracy, they currently lag far behind human annotators. We find that understanding Mandarin and code-mixed text are key challenges these models need to overcome. We hope that our new dataset inspires modeling advances that could improve the practice of bilingual child language assessment.

1 Introduction

Receptive and expressive language assessment is a standardized clinical procedure to evaluate children's communication abilities, detect signs of potential language delays and disorders, and facilitate early clinical interventions in a timely manner (Wang et al., 2020; Gorman et al., 2015). However, the current practice of language assessment in bilingual children is hindered by a lack of bilingual clinicians and bilingual assessment tools (Sheng et al., 2021; Wang et al., 2024; Pratt et al., 2022). This scarcity necessitates the involvement of parents, who typically speak the languages that bilingual children are exposed to at home, and can serve as key informants for clinicians during the assessment process (Klatte et al., 2020).

Prior research on using parents as "e-helpers" for telehealth in the context of bilingual assessments has drawn more attention from multidisciplinary researchers (Edwards-Gaither et al., 2023). Typically, parents are allowed to assist their children in interacting with the graphic user interface (GUI) of the assessment tool (Fissel et al., 2015). Prior work (Du et al., 2020) reported that during in-person language assessments, bilingual Mandarin-English-speaking parents often engage in verbal and physical behaviors (e.g., dyadic conversations, gestural prompting), which are considered interference behaviors. For example, providing hints or directly suggesting answers can hinder their children's assessment performance and compromise the validity of the assessment results. To ensure the integrity and fairness of the assessment, clinicians are tasked with manually annotating and evaluating parents' behaviors to determine whether they constitute interference. This clinical annotation process is not only time-consuming to complete but also challenging to reach reliable agreements between clinicians without extensive training (Yao et al., 2023).

In this paper, we investigate the potential of large language models (LLMs) to support clinicians by automatically annotating parents' behaviors during assessments, relying on LLM's abilities to learn new tasks via zero-shot or few-shot learning (Brown et al., 2020). Such an LLM system would reduce the workload of clinicians to manually identify parental interference, and could even be used to discourage interference in real time during assessments. We first collect and release a comprehensive dataset comprising conversational transcripts and behavior descriptions from recordings of two groups of parent-child dyads (in-person and virtual) undergoing bilingual English-Mandarin language assessments. This dataset, which is utilized by clinical experts for post-assessment analysis, serves as the benchmark for our model development and evaluation in classifying parental behaviors as either supportive (i.e., acceptable) or interfering (i.e., detrimental). Our dataset comprises data from 59 patients and containing 1,472 total parent behaviors annotated with one of eight fine-grained labels. This dataset will facilitate future research that could enable more efficient administration of bilingual langauge assessments.

Finally, we benchmark two state-of-the-art LLMs—Llama 3 and GPT-4—on our dataset, using both zero-shot and few-shot prompting. While GPT-4, the stronger model, performs decently well on our task, there is still considerable room for improvement compared with human expert accuracy. We observe that examples involving Mandarin utterances by parents are particularly challenging for these models, suggesting that improved multilingual modeling could make LLMs more useful for this setting. Overall, our work both sheds light on weaknesses of state-of-the-art LLMs and introduces a challenging, ecologically valid, multilingual benchmark that we hope inspires future modeling improvements.

2 Related Work

2.1 Clinical NLP Research on Bilingual Language Assessment in Telehealth

Prior NLP research has examined the automation of various educational and clinical tasks, such as automated scoring and analysis of pediatric language assessment (Wang et al., 2020; Gorman et al., 2015), behavioral testing for clinical outcome prediction (Van Aken et al., 2021), novel test item generation in clinical assessments (Laverghetta Jr and Licato, 2023), and narrative tasks (Prud'hommeaux and Roark, 2015; Chen et al., 2023). However, prior work related to utilizing NLP approaches to analyze clinical encounters in the context of telehealth using a bilingual dataset has been very limited. In clinical language assessment, gathering and analyzing data for clinician-led language assessment task can be challenging to obtain and time-consuming to analyze. Therefore, investigating how to use NLP approaches to classify and annotate these behaviors brings significant contributions to improving efficiency in clinical workflow and increasing the quality of bilingual assessment.

2.2 LLMs for Real-World Domains

Large language models (LLMs) like GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023a,b) have shown impressive task-solving performance off-the-shelf, such as question answering and logical reasoning (Wei et al., 2021; Sanh et al., 2021; Chung et al., 2022). 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 to teach LLMs to solve a particular task. ICL enables models to learn from few-shot examples provided in the input context. Many recent work have also explored LLMs' capabilities in real-world scenarios, which generally require significant domain expertise, such as children education (Chen et al., 2023) and medical domains (Xu et al., 2024; M. Bran et al., 2024; Jablonka et al., 2024).

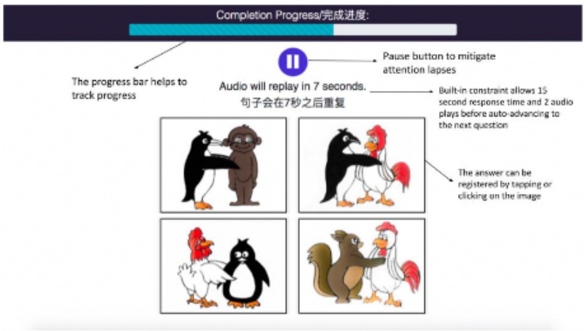
Completion Progress/完成进度:
The progress bar helps to track progress
Pause button to mitigate attention lapses
Audio will replay in 7 seconds.
句子会在7秒之后重复
Built-in constraint allows 15 second response time and 2 audio plays before auto-advancing to the next question
The answer can be registered by tapping or clicking on the image

Figure 1: MERLS 1.0 web interface English test item: "The monkey is hugged by the penguin."

3 Dataset

We collaborated with bilingual Mandarin-English speaking SLP practitioners and researchers to obtain two comprehensive text-based datasets derived from the Mandarin-English Receptive Language Screener (MERLS). MERLS is a web-based receptive sentence comprehension assessment designed for Mandarin-English-speaking children between 4 and 8 years old (Sheng et al., 2021; Du et al., 2020). The assessment consists of 44 Mandarin test items and 36 English test items. During assessment, the MERLS web interface (see Figure 1) plays Mandarin and English audio instructions for bilingual children to select pictures that match the instructions. Children do not require parental assistance to complete tasks except for technical support.

3.1 Data Collection

Our dataset comprises two distinct studies using MERLS. The In-Person dataset (MERLS 1.0) was collected in-person with video recording, while the Virtual dataset (MERLS 2.0) was collected remotely during the COVID-19 pandemic via Zoom. Each dataset includes 16 pairs of parents and children, resulting in a total of 32 parent-child pairs that are matched in parent education and children's age within 6 months differences. An additional 27 parent-child dyads were collected virtually, for a total of 43 parent-child dyads in the Virtual dataset.

3.2 Annotation Process

For each test item collected in each parent-child video recording, a clinical annotation team identified, transcribed, and categorized all child and parent behaviors performed while the child answered that item. The transcription of each behavior describes both verbal and non-verbal actions. Following guidelines by(Du et al., 2020), each parent behavior was first categorized into one of two primary classes: "interference" and "support." Each primary class was further subdivided into four subclasses to capture more detailed behavior types, as listed in Table 4 in Appendix A.1. Namely, sub-categories for "interference" are: "repeat questions", "answer questions", "analyze items", and "judging of correctness"; sub-categories for "support" include "encouragement", "analyze items", "broadcasting", "miscellaneous".

Overall, this dataset is therefore an eight-way classification task, which tests nuanced understanding of parental participation during the assessments. Each input consists of the text of the current test item, a description of the child's actions (if any), and the description of the parent's behavior. More details are provided in Appendix A.2

To validate the annotations, interobserver agreement (IOA) was calculated between the two clinical experts using member checking (Birt et al., 2016). IOA is 97% (In-person dataset) (Du et al., 2020) and 86.1% (Virtual dataset). We derive human accuracy numbers in Table 1 from these IOA numbers.

3.3 Dataset Statistics

Table 4 presents overall statistics for the In-person and Virtual datasets. The two datasets have different label distributions: the Virtual dataset contains fewer interference behaviors and more technical support behaviors. It may be due to the (1) the system redesign of the MERLS website prior to 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 which added additional technical behaviors in parent-child dyads in the Virtual dataset.

4 Behavior Classification with LLMs

We aim to investigate whether LLMs can reliably classify parental behaviors during child language assessments, in comparison with clinical experts' performance. We focus on the zero-shot (ZS) and few-shot (FS) In-Context Learning prompting strategies for LLMs throughout our experiment.

Prompts. Our zero-shot prompt 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 (if any), and (3) the description of the parent's behavior. The few-shot prompt is similar but includes one demonstration under each label definition. The demonstrations were written by a clinical expert, to be separate from all dataset examples. The full prompts are in Appendix A.5.

Models. We experiment with two LLMs: the open-weight model Llama-3-8B-Instruct and the closed-source model GPT-4. While Llama-3 is a primarily English model, its pre-training dataset does include data from 30 other languages[1]. GPT-4 has also been shown to perform well on Chinese language understanding benchmarks (Xu et al., 2023; Zhu et al., 2024).

Evaluation metrics. We compute two metrics: Behavior-level Accuracy (BEHAVACC) measures the fraction of behaviors that are predicted correctly, while Item-level Accuracy (ITEMACC) measures the fraction of all test items containing at least one behavior for which *all* behaviors done during that item were predicted correctly.

5 Experimental Results

5.1 Main Results

Table 1 shows the overall accuracies of all models on our dataset. GPT-4 greatly outperforms Llama3 in all data subsets, with the best GPT-4 variant achieving 61.2 BEHAVACC on the In-person dataset and 49.3 BEHAVACC on the Virtual dataset. This is still far below the human expert accuracy of 98.5 and 93.0, respectively, indicating significant room for improvement. Appendix A.3 shows results on the subset of 16 Virtual dataset who were paired to match the In-person dataset; we see similar trends as on the full Virtual dataset.

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

	In-Person			Virtual		
Question Language	English	Mandarin	All	English	Mandarin	All
Llama3 ZS	48.1 (38.9)	43.3 (28.0)	44.6 (31.4)	29.3 (26.7)	31.3 (24.3)	30.4 (25.3)
Llama3 FS	44.9 (33.6)	39.3 (27.6)	40.8 (29.5)	22.1 (20.3)	23.6 (19.8)	23.0 (20.0)
GPT-4 ZS	65.2 (**61.1**)	**60.0** (**49.2**)	**61.4** (**52.9**)	45.3 (36.9)	51.6 (45.3)	48.9 (41.6)
GPT-4 FS	**65.8** (60.2)	54.5 (40.8)	57.6 (46.8)	**48.5** (**43.3**)	**52.5** (**46.5**)	**50.8** (**45.1**)
Human Experts	-	-	98.5	-	-	93.0

Table 1: Main results. Each cell shows BehavAcc with ItemAcc in parentheses. ZS = zero-shot, FS = few-shot.

	In-Person			Virtual		
Parent Language	English	Mandarin	Mixed	English	Mandarin	Mixed
# Examples	233	167	178	226	389	111
Llama3 ZS	39.5	42.5	53.4	44.2	21.1	35.1
Llama3 FS	34.8	40.1	49.4	32.7	14.7	32.4
GPT-4 ZS	64.8	55.1	62.9	56.6	43.2	53.2
GPT-4 FS	57.1	52.1	63.5	60.2	45.5	50.5

Table 2: BehavAcc 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.

Across all models and language subsets, the Virtual dataset is much harder to classify than the In-person dataset. This is likely due to the fact that the In-person dataset was collected with a more comprehensive view of the parent-child interaction with the screen, whereas the Virtual dataset was collected via Zoom with limited camera view for capturing detailed parent-child interaction. These environmental factors influenced human annotation during initial text transcription. We also observe that the zero-shot setting leads to better performance with GPT-4 on In-person dataset whereas the few-shot setting performs best across all models on Virtual dataset. This could also be contributed to the different behaviors in In-person and Virtual datasets that the experts-annotated demonstrations are more alike to the ones in the Virtual dataset.

5.2 Effects of Parent Language

We now examine whether the language that describes the parent's behavior impacts the LLM's accuracy. This description could be either English-only, Mandarin-only, or a mix. In our dataset, all descriptions of non-verbal actions are in English; on the other hand, many parent speech acts are in Mandarin. A mix of languages can occur when the parent code-switches, or when the parent's Mandarin speech act is accompanied by an English description of a non-verbal action.

Table 2 shows model accuracies broken down by the language describing the parent behavior. All models generally perform worst on the Mandarin-only items, with the exception of Llama3 on the In-person dataset. Moreover, the higher difficulty of the Virtual dataset relative to the In-person dataset is explained primarily by the increase in difficulty of behaviors that involve Mandarin. Averaging across models, the English Virtual dataset is only 0.6 BehavAcc harder than the English In-person dataset; however, this number jumps to 16.3 for Mandarin and 14.5 for the Mixed data.

5.3 Binary Classification Results

Finally, we evaluate models on the binary classification task to distinguish interference from support behaviors. Detecting interference could be useful to alert a clinical expert to potential issues, even if the model cannot identify the type of interference. As shown in Appendix A.4, GPT-4 outperforms Llama3 and can achieve an accuracy greater than 83% in both the In-person and Virtual datasets.

6 Conclusion

This paper introduces a new dataset for fine-grained classification of parental behaviors during bilingual English-Mandarin child language assessment. Automating this task would enable more efficient and reliable language assessments. Current state-of-the-art LLMs perform somewhat accurately on this task, but still have considerable room for improvement, especially on examples involving Mandarin parental speech. We hope our dataset encourages further NLP research to support clinical tasks that are specifically designed for multilingual speakers.

7 Limitations

In our current setup, a model must classify parent behavior given the text of the parent's speech and/or a textual description of their actions. This text was transcribed manually from the video recording of the language assessment sessions by a human expert. The use of human transcription ensures high quality text data inputs, and thus our results represent a likely upper bound for model performance. In a real application, it would be ideal to avoid this manual transcription step and make predictions from the raw video directly. This could be done by using a separate video transcription model to first transcribe the video to text, then running an LLM. It could also be done by using a multimodal foundation model to process the video directly. We leave exploration of handling this more challenging multimodal version of the problem to future work.

Our dataset focuses exclusively on one language pair, English and Mandarin. This choice was based on the expertise of the clinical authors of the paper and their access to English-Mandarin bilingual participants. We believe that the framework presented could be extended to other language pairs, though validating this would require new collaborations to collect the required data.

8 Ethical considerations

Data collection and analysis. Our dataset was originally collected via university human subject research approval and data sharing agreements. The clinician annotated text transcripts for In-person and Virtual dataset were generated as a part of the clinical video analysis, which are all de-identified behaviors without any sensitive information from parent-child pairs.

Bias mitigation. In order to ensure a representative dataset, when comparing the 16 virtual and 16 in-person parent-child pairs, we carefully considered the impact of children's age and parents' education. Children's ages can be directly correlated to their language ability; parent education level can results in different abilities to interpret the testing procedures, leading to various levels of parent behaviors.

References

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.

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.

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*.

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*.

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.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. 2024. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, pages 1–9.

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.

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.

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.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11.

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

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.

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.

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.

Yiyi Wang, Emily Prud'Hommeaux, 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*.

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.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Bingsheng Yao, Ishan Jindal, Lucian Popa, Yannis Katsis, Sayan Ghosh, Lihong He, Yuxuan Lu, Shashank

Srivastava, Yunyao Li, James Hendler, et al. 2023. Beyond labels: Empowering human annotators with natural language explanations through a novel active-learning architecture. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11629–11643.

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 Example Appendix

A.1 Description of Labels

Table 3 provides detailed descriptions of the eight labels used in our dataset.

A.2 Dataset details

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

1. Time Stamps: Precise time stamps for each assessment item and corresponding parent-child behavior during the assessment sessions.

2. GUI Descriptions: Textual descriptions of the graphical user interface (GUI) elements displayed during the assessment.

3. Audio transcriptions: Transcriptions of the audio recordings from the assessment sessions, with annotations that identify the speaker for each voiceover.

4. Behavior Descriptions: Textual descriptions of the parents' behaviors during the assessments, detailing both verbal and non-verbal actions.

The initial data cleaning and annotation were conducted by a clinical team consisting of two graduate research assistants, with inter-rater reliability verified by a licensed SLP for MERLS 1.0 Dataset and SLP assistant for MERLS 2.0 Dataset. The statistics of our dataset is reported in Table 4.

A.3 Results with paired cohort

Table 5 shows similar results as our main results table, Table 1, but with a paired subset of the virtual dataset. In particular, we use a subset of 16 patients from the virtual dataset who were chosen to match the 16 in-person patients in terms of child age and parent education level. Overall, we observe similar trends to those in Table 1.

A.4 Binary Classification Results

Table 6 shows the binary BEHAVACC results on our dataset. To evaluate binary classification accuracy, we simply map the predictions from the model prompted for 8-way classification to either interference or support. We see that GPT-4 is able to achieve at least 83% binary classification accuracy on both the in-person and virtual datasets.

A.5 Prompts

Figure 2 shows the zero-shot prompt used in our experiments. Figures 3 and 4 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 (if any), and (3) **Parent behavior**, the utterance and/or a description of the action performed by the parent.

Top-Level Category	**Sub-Level Category**	**Definition**
Interfere	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".
Support	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.
	Analyze Items	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.

Table 3: The classification categories and corresponding definitions of parent behaviors based on established clinical guidelines. Two high-level categories ("interfere" and "support") each consists of four sub-categories.

```
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.

```
- 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.

	In-Person	Virtual
# Parents	16	43
# Behaviors	677	795
# Items with $\geq$ 1 behavior	363	430
# Repeat Questions	144	24
# Answer Questions	12	11
# Analyze Items	89	3
# Judging of Correctness	51	12
# Encouragement	136	281
# Technical Support	148	291
# Broadcasting	60	17
# Miscellaneous	37	156

Table 4: MERLS dataset statistics. Bottom shows the label distribution, with interference behaviors shaded.

	In-Person			Virtual (paired 16)		
Question Language	English	Mandarin	All	English	Mandarin	All
Llama3 ZS	48.1 (38.9)	43.3 (28.0)	44.6 (31.4)	30.2 (23.2)	33.1 (23.7)	31.7 (23.4)
Llama3 FS	44.9 (33.6)	39.3 (27.6)	40.8 (29.5)	21.4 (15.9)	19.8 (12.9)	20.5 (14.3)
GPT-4 ZS	65.2 (61.1)	60.0 (49.2)	61.4 (52.9)	46.5 (34.1)	44.8 (37.6)	45.6 (36.0)
GPT-4 FS	65.8 (60.2)	54.5 (40.8)	57.6 (46.8)	47.8 (40.2)	45.9 (40.9)	46.8 (40.6)

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. Each cell has BEHAVACC with ITEMACC in parentheses afterwards. ZS = zero-shot, FS = few-shot.

	In-Person			Virtual		
Binary accuracy on:	Interference	Support	All	Interference	Support	All
Llama3 ZS	80.5	65.3	73.0	50.9	64.1	63.1
Llama3 FS	82.9	56.1	69.7	71.9	54.3	55.6
GPT-4 ZS	91.8	82.5	87.2	71.9	77.0	76.6
GPT-4 FS	91.1	80.4	85.8	71.9	84.0	83.1

Table 6: BEHAVACC on the binary classification version of our dataset. We include the overall accuracy as well as accuracy on interference only or support only. ZS = zero-shot, FS = few-shot.