Playlogue: Dataset and Benchmarks for Analyzing Adult-Child Conversations During Play

MANASA KALANADHABHATTA, Manning College of Information and Computer Sciences, University of Massachusetts Amherst, USA
MOHAMMAD MEHDI RASTIKERDAR, Manning College of Information and Computer Sciences, University of Massachusetts Amherst, USA
TAUHIDUR RAHMAN, Halıcıoğlu Data Science Institute, University of California San Diego, USA
ADAM S. GRABELL, Department of Psychological and Brain Sciences, University of Massachusetts Amherst, USA
DEEPAK GANESAN, Manning College of Information and Computer Sciences, University of Massachusetts Amherst, USA

There has been growing interest in developing ubiquitous technologies to analyze adult-child speech in naturalistic settings such as free play in order to support children's social and academic development, language acquisition, and parent-child interactions. However, these technologies often rely on off-the-shelf speech processing tools that have not been evaluated on child speech or child-directed adult speech, whose unique characteristics might result in significant performance gaps when using models trained on adult speech. This work introduces the Playlogue dataset containing over 33 hours of long-form, naturalistic, play-based adult-child conversations from three different corpora of preschool-aged children. Playlogue enables researchers to train and evaluate speaker diarization and automatic speech recognition models on child-centered speech. We demonstrate the lack of generalizability of existing state-of-the-art models when evaluated on Playlogue, and show how fine-tuning models on adult-child speech mitigates the performance gap to some extent but still leaves considerable room for improvement. We further annotate over 5 hours of the Playlogue dataset with 8668 validated adult and child speech act labels, which can be used to train and evaluate models to provide clinically relevant feedback on parent-child interactions. We investigate the performance of state-of-the-art language models at automatically predicting these speech act labels, achieving significant accuracy with simple chain-of-thought prompting or minimal fine-tuning. We use in-home pilot data to validate the generalizability of models trained on Playlogue, demonstrating its utility in improving speech and language technologies for child-centered conversations. The Playlogue dataset is available for download at https://huggingface.co/datasets/playlogue/playlogue-v1.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; **Human computer interaction (HCI)**; • **Computing methodologies → Speech recognition**; **Natural language processing**.

Additional Key Words and Phrases: Child speech, adult-child interaction, play, automatic speech recognition, speaker diarization, speech classification, large language models, audio dataset

Authors' Contact Information: Manasa Kalanadhabhatta, Manning College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, USA, manasak@cs.umass.edu; Mohammad Mehdi Rastikerdar, Manning College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, USA, mrastikerdar@cs.umass.edu; Tauhidur Rahman, Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, USA, trahman@ucsd.edu; Adam S. Grabell, Department of Psychological and Brain Sciences, University of Massachusetts Amherst, Amherst, USA, agrabell@umass.edu; Deepak Ganesan, Manning College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, USA, dganesan@cs.umass.edu.

ACM 2474-9567/2024/12-ART173
https://doi.org/10.1145/3699775

ACM Reference Format:
Manasa Kalanadhabhatta, Mohammad Mehdi Rastikerdar, Tauhidur Rahman, Adam S. Grabell, and Deepak Ganesan. 2024. Playlogue: Dataset and Benchmarks for Analyzing Adult-Child Conversations During Play. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 173 (December 2024), 34 pages. https://doi.org/10.1145/3699775

1 Introduction

Analyzing adult-child speech interactions is crucial for understanding and promoting healthy language development, school readiness, and academic performance in children [38, 42, 60, 79]. Adult-child conversations, both at home with parents or caregivers and in classroom settings with teachers, play a significant role in shaping a child's language skills and overall communication abilities [79]. For instance, the variability in the amount and diversity of talk that parents direct to their toddlers is associated with variability in their language skills at age 3 [12]. Extensive research has demonstrated that the quantity and quality of adult-child interactions are also key drivers of growth in children's prosocial behavior and mental health, with short and long-term consequences for children's ability to care, show empathy, and voluntarily assist others [48, 90]. Teacher-child interactions in the classroom environment are similarly important for early social-emotional functioning [7].

Given the importance of analyzing adult-child conversations for understanding and supporting children's social, academic, and language development, there has been strong interest in developing ubiquitous technology to automatically record and analyze adult-child speech interactions in naturalistic settings. Various sensing modalities, particularly audio, are being increasingly deployed in the home as well as in classrooms to understand young children's experiences within early childhood education [22, 96].

One setting of particular interest in studying adult-child conversations has been during play-based interactions. As humans, some of our most formative early experiences occur in the context of play-based interpersonal interactions with our caregivers, making play sessions incredibly meaningful and informative social interactions [31, 78]. Social scientists have therefore studied play sessions for decades to understand early development, parenting, and early family functioning [31, 91, 93]. Several standardized assessment tools have been developed to evaluate the quality and effectiveness of play-based conversations. For example, the Dyadic Parent-Child Interaction Coding System (DPICS) is widely used in parent-child interaction therapy (PCIT) to assess the quality of parent-child communication and guide intervention strategies where clinicians coach parents to gain new, more adaptive parenting skills across repeated play sessions [25]. Similarly, the Brief Observation of Social Communication Change (BOSCC) is an observational measure designed to capture changes in social communication skills and monitor intervention progress in children with autism spectrum disorder (ASD) during brief play-based interactions with an adult [34].

The importance of analyzing adult-child speech interactions, both in the context of play and otherwise, has been increasingly recognized in the fields of ubiquitous computing and human-computer interaction. In the UbiComp community, several studies have focused on using wearable devices and sensors to capture and analyze adult-child speech interactions in naturalistic settings. Some of the early work in this domain includes the "Human Speechome Project" to record and analyze child-adult conversations in the home environment to study child development [80] as well as the use of wearable cameras and audio recorders to capture adult-child interactions during play sessions [75]. There have also been a number of efforts focused on developing mobile and wearable technologies to support language assessment and intervention in the context of parent-child interaction (e.g. [36, 43, 88, 106]). However, most of these technologies rely on low-level vocal features, focus only on adult speech, use off-the-shelf speech recognition models that are not tested on child-centered speech, or resort to manual approaches to understand child speech, thereby extracting limited insights from an extremely rich and varied source of data. Much of the work in the UbiComp and HCI fields also does not take interaction dynamics into account, but instead focuses on utterance-level insights or conversation-level aggregates rather than moment-to-moment changes within conversations.

While an accurate and in-depth analysis of adult-child speech during play-based interactions would have major implications for early education, parenting, and clinical work, the development of ubiquitous technologies to make this possible necessitates overcoming several technical challenges. Any pipeline for analyzing such interactions must include several preliminary stages such as the separation of child and adult speech by identifying the speaker and automatic speech recognition (ASR) to accurately transcribe the spoken content. Finally, a scoring function like DPICS needs to be applied to determine the quality of the interaction. Recent advances in large deep learning-based ASR models and Large Language Models (LLMs) offer powerful tools that could accelerate our ability to automatically analyze and mediate adult-child interactions. However, several gaps exist in the current state of research:

(1) Existing speech and language models are not sufficiently tested for child speech or adult speech that is directed towards young children, which differs significantly from adult-to-adult communication. The unique characteristics of such speech can substantially impact the performance of ASR and speaker diarization models in this domain [27], but prior work has largely ignored taking into account this performance gap.
(2) There is a lack of benchmark ASR and diarization datasets containing naturalistic adult-child speech, especially with younger children whose language skills are still developing. This makes it difficult to evaluate existing models in this crucial age range.
(3) There is a lack of datasets containing real-world adult-child conversations with annotations such as DPICS, which would enable the development of automated scoring methods. While it is difficult to collect and publicly release such datasets due to privacy concerns, they are essential for realizing the vision of intelligent systems that can automatically reason about the quality of parent-child interactions and recommend appropriate interventions.

To address these gaps, we curate a dataset for child-centered speech processing during play-based interactions and evaluate baseline models on this dataset. As an example of a downstream task, we annotate a subset of the dataset with parent and child DPICS labels and investigate whether various state-of-the-art text classification models can accurately predict these labels. Our paper makes the following contributions:

(1) We curate a dataset, Playlogue, containing over 33 hours of adult-child interaction audio recorded during play sessions from three different corpora plus non-play sessions from one additional corpus. We apply extensive manual filtering and automated forced-alignment techniques to enable researchers to use the dataset for applications such as speaker diarization and ASR. We annotate a subset of this curated dataset with DPICS codes, producing 4773 labeled parent utterances and 3895 labeled child utterances along with full conversation audio and context. This provides a valuable resource for exploring diarization, ASR, and DPICS prediction using state-of-the-art large audio processing and language models, thereby enabling new research in this domain.
(2) Using this dataset, we first evaluate the performance of state-of-the-art speaker diarization models trained to segment adult-adult interactions on adult-child speech. We show that there is considerable performance degradation, and investigate whether fine-tuning these models on adult-child speech improves performance. We also train and evaluate adult/child audio classification models using both traditional audio features and speech representations obtained from pretrained deep neural networks.
(3) Second, we examine several state-of-the-art ASR models to understand the extent of performance degradation when applied to parent-child interactions compared to typical adult speech. We demonstrate that there is a substantial performance gap and assess whether applying fine-tuning techniques can help mitigate this gap. While significant improvements are observed, there remains considerable room for further enhancement.
(4) Third, we investigate the ability of state-of-the-art LLMs and sentence transformer models to predict DPICS labels. This analysis provides insights into the feasibility of future AI models that can reason about the

quality of parent-child interactions and use this information to provide feedback without disrupting the natural flow of the interaction.

(5) Finally, we conduct a pilot study to record in-home parent-child interactions using a smartphone app. We collect naturalistic audio data to test the generalization performance of models trained on Playlogue to real-world UbiComp settings. We demonstrate that diarization, ASR, and text classification models trained or fine-tuned on Playlogue generalize well to this unseen setting, achieving significantly better performance than state-of-the-art pretrained models.

The Playlogue dataset would enable researchers to develop more accurate and useful speech and language models for future ubiquitous technologies that support adult-child interactions. A unique feature of the Playlogue dataset is that it contains full audio and annotated transcripts from spontaneous (not enacted) conversations that are, on average, 12 minutes and 36 seconds long. This allows researchers to study adult-child interactions in context, enabling them to model interaction dynamics and state transitions from one speech act to another using a purely naturalistic dataset.

2 Related Work

2.1 Adult-Child Interaction in UbiComp and HCI Research

There has been long and sustained interest in the ubiquitous computing and human-computer interaction communities towards understanding and developing tools to support interaction between children and adults (mainly parents). Previous research has used a wide range of modalities to sense and support adult-child interactions, including wearable physiological sensing [71], instrumented interactive toys [13], wearable camera-based gazed estimation [15], social cues delivered via Google Glass [99], smartphone app-based self-reflection [46], etc. Here, we highlight prior work that specifically utilizes audio sensing to monitor parent-child interactions. For example, Hwang et al. proposed the TalkBetter system that used mobile phones to record parent-child conversations and monitor turn-taking and meta-linguistic behaviors such as excessively long parent turns, lack of child responses, syllable rate, etc. to provide parents feedback on communication strategies [44]. Similarly, TalkLIME provided parents visual feedback on their interactions with their child by tracking utterance duration and turn initiation ratios for each party [88]. Yoo et al. derived "conflict cues" based on non-verbal behaviors (e.g., crying, yelling) from child audio to capture parent behavior, enabling parents to reflect on their actions from the child's point of view [108]. MAMAS captured parent and child speech during mealtimes and performed transcription and sentiment analysis to help parents understand and improve mealtime interactions [47]. "Rosita Reads With My Family" included a bilingual conversational interface to support language learning and parent-child joint engagement during co-reading [39, 105]. Other work has focused solely on adult speech within parent-child dyadic interactions. For example, SpecialTime provided parents engaged in parent-child interaction therapy with real-time feedback by transcribing audio and classifying speech acts [43]. "Captivate!" used parent speech to sense joint attention during parent-child dyadic interactions, but did not directly analyze child speech [53].

The above examples highlight the marked interest in the UbiComp and HCI research communities in analyzing various forms of adult-child interactions. Many of the above technologies rely on various speech and language processing components, however, there is a lack of such components specifically designed with child populations in mind. "Rosita", SpecialTime, and "Captivate!" all used the Google Cloud speech-to-text service. A key limiting factor in adult-child interaction research is the lack of domain-specific open datasets for researchers to develop and evaluate machine learning models. This is especially true of datasets that contain information about adult-child interaction quality and clinically validated behaviors. Huber et al. [43] released a synthetic dataset of parent sentences labeled using the Dyadic Parent-Child Interaction Coding System (DPICS; [25]), which has been used by other work to develop text classifiers for parent speech (e.g., [55, 67]). However, this dataset contains discrete

utterances without the surrounding conversational context and lacks speech audio, both of which are important indicators for experts when assigning DPICS labels to parent-child conversations [25].

2.2 Children and Voice-based Interactions

Prior work in human-computer interaction has also focused on developing voice-based and conversational interfaces for children or using speech/vocal features for other downstream applications, which also require similar speech-processing capabilities as the work listed above. Virtual agents and AI-based storytelling applications have been utilized to create interactive technologies tailored to children that can be used as language partners or learning assistants. For example, in the education domain, MathKingdom used Baidu's speech-to-text service to record children's answers to math questions posed to them by a virtual agent as part of an educational game [104]. Mathemyths used the Google Cloud speech-to-text service and a GPT-based large language model to engage children in a conversation, co-creating storylines that expose them to mathematical language and concepts [109]. "Rosita" also uses Google's speech-to-text and DialogFlow APIs to analyze children's responses and categorize intent during a storytelling task [105]. In a similar vein, StoryCoder [20] aimed to improve children's computational thinking ability, data literacy, and creative expression through storytelling via a voice-guided smartphone application. The Spoken Impact Project used a Wizard-of-Oz system in place of automatic speech recognition to evaluate the effect of audio-visual feedback on encouraging vocalization in children with autism spectrum disorder [35]. Storytelling technologies can also exist with a physical tangible presence – for example, the Alpha Egg [94], Luka [111] and Codi [70], are commercially available AI-based robotic toys that provide platforms for interactive storytelling experiences for children. However, most of these conversational interfaces and applications rely on existing off-the-shelf services that are not evaluated in a younger population or use a Wizard-of-Oz stand-in.

Automated speech processing approaches have also been developed to use voice-based interactions to identify speech and language disorders in young children [32]. DYPA combined handwriting analysis with audio features from reading tests to screen children for dyslexia using a tablet-based application [112]. Other applications support speech therapy by utilizing speech recognition during game-based interactions to provide feedback to users or therapists (e.g., [86, 92]). Speech and vocal features have also been used to screen children for mental health issues – Cotter et al. found an association between child speech acts labeled using DPICS and parent-reported behavioral problems [17].

2.3 Gaps in Analyzing Child and Child-Centered Speech

As described above, a large body of prior work employing speech technology for children uses off-the-shelf models that may exhibit poor performance on child or child-directed speech. For example, recent work evaluating the Google speech-to-text API on 6- to 11-year-old children's speech reported a word error rate (WER) of 24% [10], compared to current state-of-the-art ASR systems that achieve a WER of approximately 6% on adult speech benchmarks [89]. While there has been some research on developing speech processing systems specifically for child speech, this has mainly focused on older children. For example, in terms of ASR technologies, child speech researchers have mainly used the MyScienceTutor dataset of spoken dialogues by children in grades 3 through 5 [98], the PF STAR dataset of children aged 4 to 14 years [81], or the CMU Kids dataset of spoken sentences from children aged 6 to 11 years [23] to train and evaluate models. We refer interested readers to recent reviews such as [6] for more details on the various ASR approaches adopted in this line of research.

There is, however, a gap in understanding the performance of ASR models on speech from younger children (i.e., preschool-aged), whose speech and language skills are less developed. Sciuto et al. found that younger children faced difficulties having their speech understood by smart speaker-based conversational agents due to their voice intonations and cadence [85]. Monarca et al. also studied children's interactions with a conversational agent and

found that children with below-average speech skills were both likely to speak less and be less well-understood by automatic speech recognition (ASR) systems [65]. This highlights the importance of improving ASR performance not just for child speech in general but also for younger children and children with language deficiencies in particular.

In a younger population, the LENA system [103] has been used commercially and in prior research to record and analyze daylong recordings of children in naturalistic environments, providing insights such as adult and child word counts, turn-taking, speech complexity, etc. in order to support the tracking of language development in young children (e.g., [30, 97]). However, recent work has found that LENA estimates of child vocalizations and adult word counts only had weak to moderate associations and large absolute discrepancies with ground-truth annotations [59], while measures such as conversational turn counts could also be inaccurate [18].

Research on adult-child speaker diarization has attempted to address these gaps using approaches such as agglomeration clustering [19, 102], probabilistic linear discriminant analysis [51, 52, 102], or end-to-end modeling with deep neural networks [50] using Mel Frequency Cepstral Coefficients (MFCC) [19, 50], i-vectors [19, 50], or x-vectors [51, 52] as input features. However, most approaches focus on pre-vocal infants (e.g., [19, 102]) or older children (e.g., [52]). Many of the datasets used in these works are also not publicly available, thereby limiting the comparison of proposed approaches across datasets. We aim to fill these gaps through our Playlogue dataset, which we hope will enable researchers to develop and evaluate both diarization and ASR technologies for preschool-aged children.

3 Benchmark Tasks

In this work, we focus on evaluating speech and language models on the following three tasks that are of particular relevance to researchers studying preschool populations: (i) adult-child speaker diarization, (ii) automatic speech recognition, and (iii) automatic coding of speech acts using the Dyadic-Parent Child Interaction Coding System (DPICS). Speaker diarization is an essential component of ubiquitous technologies to monitor children's linguistic development [103] and parent-child interaction quality (e.g., [44, 88, 47]). Automatic recognition of child and child-centered speech is central to voice-based and conversational interaction systems, including learning and entertainment systems for young children (e.g., [109, 111, 105]). Lastly, identifying DPICS speech acts from parent-child conversations has been of interest to child mental health researchers [17] as well as technology developers [43], particularly in mobile settings. While not an exhaustive list, these tasks serve as a jumping-off point to illustrate the need for a new, publicly available dataset of child-centered conversations to train and evaluate machine learning models. The tasks also reflect different levels of analyses of adult-child interactions, which enable different practical applications such as those described above for researchers and practitioners.

3.1 Adult-Child Speaker Diarization

Speaker diarization is the problem of identifying who speaks at what time in an audio recording with multiple speakers. In the context of child-centered conversations, having access to speaker identities and speech timings allows researchers to derive metrics such as response latencies, pauses, and turn-taking behaviors [29]. Prior research has demonstrated an association between these metrics and children's social and linguistic development [66], quality of interactions [8] as well as autism spectrum disorder severity [9]. The gold-standard commercial tool used for obtaining these measures in the LENA system [103], however, its limited accuracy [59] has prompted researchers to explore alternative approaches for deriving these insights.

Several approaches for adult-child speech diarization have been proposed in prior work across different recording settings, number of speakers, child age ranges, and child clinical phenotypes (e.g., [19, 50, 51, 102]). However, a fair comparison of existing approaches for adult-child diarization is made difficult by the lack of publicly available benchmark datasets similar to those of adult speakers – to the best of our knowledge, only the

SEEDLingS corpus of infant (6 to 18 month) speech is available to researchers for a fee through the DIHARD challenge [83]. We aim to fill this gap by curating audio clips of one-on-one conversations between an adult and a child aged 3 to 5 years along with forced-aligned segment timestamps for each speaker.

3.2 Automatic Speech Recognition

Automatically transcribing children's speech has been an active area of research over the last few decades, with diverse applications including understanding children's language development [82], developing educational tools [101], and supporting communication with smart speakers and other virtual agents [49]. As described in Section 2.3, the most commonly used evaluation datasets for child speech recognition ([81, 23, 98]) encompass child speech from older children (4 to 14 years) rather than the preschool age. Additionally, they include short segments of speech narrated by children, rather than longer, more naturalistic conversations between children or child-adult pairs/groups. We aim to address this gap by compiling a dataset of adult-child conversations involving preschool-aged children (3 to 5 years) from the publicly available CHILDES corpora [64]. While CHILDES data has been previously used to *train* ASR systems (e.g., [21]), we contribute a manually filtered and curated dataset with train/validation/test splits for future work to systematically evaluate ASR performance on adult-child speech.

3.3 Automatic Coding of Speech Acts Using DPICS

The Dyadic Parent-Child Interaction Coding System (DPICS [25]) is the formal coding system used by clinicians certified to conduct Parent-Child Interaction Therapy (PCIT). Briefly, PCIT is a type of behavioral parent training intervention specifically for preschool-age children with disruptive behaviors, in which the therapist coaches the parent on a series of parenting skills to increase the quality of positive attention paid to their child [63]. The DPICS coding system is used by the PCIT therapist to assess progress toward these skills in real time as parents engage in a play session with their child. Specific positive parenting skills measured and counted by the DPCIS system include labeled praise, reflecting the child's verbalizations, and narrating or describing the child's behavior [63].

However, DPICS is both a highly detailed and labor-intensive system – it requires therapists to undergo extensive training as well as to juggle accurately recording DPICS labels with other tasks during the session, such as making clinical observations or responding to the parent or child [63]. Therefore, there has been an increasing interest in recent years to automate DPICS coding using speech and text AI-based approaches [43, 67]. Although this has led to the creation of a public dataset with DPICS labeled sentences [43], there is a lack of datasets with conversational context and audio data corresponding to these DPICS classes. This presents a challenge in terms of training models that can utilize vocal features (such as rising tone inflections that signify questions) and prior conversational context (to detect reflections of child verbalization or behavior description), which are important considerations for PCIT therapists while manually assigning DPICS labels. We address this gap by creating a dataset of 4770 annotated parent utterances and parent-child audio during naturalistic interactions reflective of real-world scenarios.

4 Dataset

4.1 Corpus Selection

To create Playlogue, we identified three corpora from CHILDES [64], the child language component of the TalkBank system [58], that contain audio recordings and transcriptions of one-on-one play-based conversations in North-American English between adults and children aged 3 to 5 years. These included the following corpora:

(1) *EllisWeismer* [40]: This corpus contains language samples recorded between 2.5 years to 5.5 years of age from children with and without language delays. All children were from monolingual English-speaking families in the US Midwest. We only included children without language delays in our dataset. We selected

play-based examiner-child conversations at 3.5 and 4.5 years and parent-child conversations at 3.5 years, which were recorded using a standard set of toys as props for play-based conversations.

(2) *Gleason* [61]: This corpus contains recordings of children between 25 and 62 months of age in one-on-one play-based conversations with their father and mother as well as at the dinner table. All participants were from White, middle-class families in the greater Boston area in the US and spoke English as a first language. We selected all play-based recordings where the child's age was between 3 years 0 months and 5 years 11 months for inclusion in Playlogue and discarded all dinner conversations.

(3) *VanHouten* [95]:This corpus includes child-centered conversations with US-based children aged two or three years. Playlogue includes conversations from the three-year subset that were recorded during free play between an examiner and the child using a park set as props.

In addition to the above play-based corpora, we also selected a small, curated subset of storytelling narratives from CHILDES as non-play-based training data:

(1) *Cameron* [11]: This corpus includes audio recordings from three different activities completed by children aged 4 to 5 years residing in upstate New York in the US who are speakers of Standard American English (14 participants) or African American English (15 participants). We included one conversation between each child participant and an examiner, where the child came up with and told their own story based on a wordless picturebook or toy characters from a playset.

4.2 Data Filtering and Curation

After selecting relevant corpora for Playlogue, we filtered out participants from each corpus who were out of the 3- to 5-year-old age range or who were missing audio recordings or human-annotated transcripts. We manually checked each transcript and identified several recordings that were either partially transcribed, had transcription errors, or had incorrect time synchronization between audio and transcript. We followed a two-stage process to filter and fix these errors: first, we manually identified the start and end times for a correctly transcribed segment in each audio file. Segments were included even if time synchronization was missing or erroneous, as long as the speech was correctly transcribed. We used these start and end timestamps to trim each audio and transcript file, generating clips that would be included in Playlogue.

In the second step, we used the NeMo Forced Aligner (NFA) tool from the NVIDIA NeMo Framework [37] to generate token-, word- and segment-level timestamps for the selected audio clips. NFA uses Viterbi decoding [28] with an automatic speech recognition (ASR) model based on Connectionist Temporal Classification (CTC; [33]) to generate alignments. We used the Parakeet CTC 0.6B version of the FastConformer model [77] from NeMo in our NFA implementation. The reliability of the output alignments was manually verified using the Gecko tool [54] and was found to be more accurate than the original time synchronization data obtained from CHILDES. We include the generated alignment information in Playlogue and use it for all experiments described in this paper.

4.3 Dataset Statistics

Following the data selection, filtering, and alignment process described above, Playlogue includes **over 33 hours of adult-child audio with word-aligned transcripts**. This includes 158 audio recordings from 110 unique participants aged 3 to 5.5 years. Playlogue can be utilized by researchers to develop and evaluate machine learning models for automatically analyzing conversations of preschool-aged children with an adult, enabling applications such as automatic speech recognition and conversation analysis in an early childhood population. The mean duration of clips in the dataset is 12 minutes and 36 seconds, allowing researchers to analyze interaction dynamics in long-form conversations by going beyond sentence-level or few-second-long analyses. More information about the Playlogue dataset is presented in Table 1, along with per-corpus statistics.

Table 1. Descriptive statistics for the Playlogue dataset.

Corpus	Corpus Type	No. of Clips	No. of Participants	Participant Age Range (months)	Participant Sex (M: Male, F: Female)	Total Duration	Duration of Adult Speech	Duration of Child Speech
EllisWeismer	Play	85	44	42–54	M: 57, F: 28	18:22:31	07:02:08	05:33:22
Gleason	Play	23	16	36–62	M: 12, F: 11	09:20:47	04:12:47	02:21:52
VanHouten	Play	21	21	38–42	M: 12, F: 9	02:54:49	01:07:10	00:55:10
Cameron	Narrative	29	29	47–66	M: 15, F: 13, Unknown: 1	02:34:20	00:44:39	01:03:18
TOTAL		**158**	**110**	**36–66 months**	**M: 96, F: 61, Unknown: 1**	**33:12:29**	**13:06:46**	**09:53:43**

4.4 Speech Act Labeling Using DPICS

To demonstrate the utility of the Playlogue dataset on downstream applications of interest to researchers and practitioners, we labeled a portion of Playlogue using the DPICS parent and child labels. We used the play-based parent-child conversations recorded at the age of 4.5 years from the *EllisWeismer* corpus for this task since DPICS labels are mainly used in the context of conversations between these parties during parent-child interaction therapy. This resulted in **27 clips** with unique participants, or **5 hours and 31 minutes** of data, that were annotated with DPICS labels.

Two authors served as coders and used the DPICS manual [25] to familiarize themselves with the ten parent and four child verbalization categories. To maintain consistency with prior work such as [43], the standard parent categories of "Direct Command" and "Indirect Command" were condensed into the "Command" label. Similarly, "Information Question" and "Descriptive/Reflective Question" were combined into a single "Question" label. This resulted in eight DPICS parent labels including "Negative Talk", "Command", "Labeled Praise", "Unlabeled Praise", "Question", "Reflective Statement", "Behavior Description", and "Neutral Talk". DPICS labels were assigned to child verbalizations based on the standard categories, i.e., "Negative Talk", "Command", "Question", and "Prosocial Talk". To the best of our knowledge, there are no existing datasets of child DPICS codes that are publicly available for use by researchers interested in studying child speech and language.

Coders studied the DPICS manual and discussed label definitions as well as examples before starting the annotation process. Labels were assigned through a custom interface implemented on the Qualtrics platform[1] that allowed coders to simultaneously listen to the audio and scroll through the transcript. Coders actively considered vocal characteristics (such as changes in tone and pitch), content, and context (preceding parent and child conversation) to assign one of the eight parent DPICS labels to each parent sentence or one of the four child labels to each child sentence. Parent and child DPICS codes were annotated in two separate passes over each clip to minimize errors and ensure higher data quality. Six of the 27 clips (approximately 20%) were independently rated by both coders. Inter-rater reliability across the six clips was found to be almost perfect [62] with a Cohen's kappa [16] value of 0.93 and a percentage agreement of 95.3% over 886 parent utterances and a kappa of 0.98 and agreement of 98.6% over 716 child sentences. Disagreements in labels were resolved via discussion and the remaining files were annotated by one of the two coders.

Table 2 shows the number of instances of each DPICS parent label within all 27 parent-child conversations. Note that the DPICS labels are heavily imbalanced – parents mostly engage in "Neutral Talk" or "Questions"

[1] https://www.qualtrics.com/

Table 2. Number of instances of each DPICS parent and child code in the labeled parent-child conversations from the Playlogue dataset.

DPICS Label	No. of sentences	% of total sentences
Parent Verbalizations		
Negative Talk	37	0.8%
Command	417	8.7%
Labeled Praise	12	0.3%
Unlabeled Praise	173	3.6%
Question	1648	34.5%
Reflective Statement	277	5.8%
Behavior Description	35	0.7%
Neutral Talk	2174	45.5%
Total	**4773**	
Child Verbalizations		
Negative Talk	56	1.4%
Command	207	5.3%
Question	556	14.3%
Prosocial Talk	3076	79.0%
Total	**3895**	

Table 3. Number of instances of each DPICS parent and child code in the parent-child conversations from the in-home pilot data.

DPICS Label	No. of sentences	% of total sentences
Parent Verbalizations		
Negative Talk	10	1.0%
Command	146	15.1%
Labeled Praise	16	1.7%
Unlabeled Praise	61	6.3%
Question	274	28.3%
Reflective Statement	29	3.0%
Behavior Description	23	2.4%
Neutral Talk	409	42.3%
Total	**968**	
Child Verbalizations		
Negative Talk	19	3.3%
Command	64	11.1%
Question	55	9.5%
Prosocial Talk	440	76.1%
Total	**578**	

while hardly using "Labeled Praise". A similar imbalance is also seen in child labels, which are dominated by "Prosocial Talk". While this can be considered a limitation of the Playlogue dataset, we argue that this is, in fact, representative of the distribution of DPICS labels that would be expected in a real-life parent-child conversation. The presence of audio and annotated transcripts from the full conversation also allows researchers to study transitions from one DPICS label to another across participants, which is not possible using existing datasets that only provide example sentences from each DPICS class without additional context (e.g., [25, 43]).

4.5 Dataset Validation: Pilot Study of In-Home Parent-Child Conversations During Play

In order to demonstrate the utility of Playlogue in training and validating adult-child speech and language models in UbiComp settings, we conducted a brief, in-home pilot study recording play-based parent-child interactions using a smartphone application. We developed a cross-platform app using the Flutter framework[2] to record parent and child audio during a structured play session. The session was based on the SpecialTime activity used in parent-child interaction therapy [26] and consisted of three phases: (i) ten minutes of child-led play, where the child chooses any activity and the parent plays with them following their lead, (ii) ten minutes of parent-led play, where the parent leads the child in playing according to their rules, and (iii) five minutes of clean-up, where the parent directs the child to put away all toys on their own.

We recruited three parent-child dyads to participate in our pilot study through community-based advertisements. Table 4 shows the participant demographics. All participants were based in the US with the child participants speaking English as their first language. All adult participants had earned a college degree or completed post-graduate training. To ensure naturalistic conditions, parents were asked to install the data collection app on their

[2]https://flutter.dev/

Table 4. Participant demographics for the in-home pilot study.

ID	Child's Age	Child's Sex	Child's Race and Ethnicity	Parent	Parent's Race and Ethnicity
P1	5 years, 4 months	Male	White; Not Hispanic/Latino	Mother	White; Not Hispanic/Latino
P2	4 years, 1 month	Female	Asian, White; Not Hispanic/Latino	Father	White; Hispanic/Latino
P3	3 years, 1 month	Female	Asian; Not Hispanic/Latino	Mother	Asian; Not Hispanic/Latino

phones and complete the 25-minute session at home without the researchers present. The app provided audio instructions at the beginning of each phase and recorded stereo audio at a sampling rate of 44.1 kHz and bitrate of 128 Kbps. Overall, the pilot study allowed us to collect a total of 75 minutes of parent-child interaction audio in diverse real-world settings using ubiquitous devices.

The pilot audio data was manually transcribed and each speech segment was assigned a speaker label (adult/child). Segment-level timestamps were obtained using the forced-alignment technique described in Section 4.2. Parent and child speech acts were labeled using DPICS in a manner similar to that described in Section 4.4 for the Playlogue data.

We first use this at-home pilot data to validate the real-world representativeness of the labeled speech acts in the Playlogue dataset. Table 3 shows the number and percentage of different DPICS parent and child codes observed within the in-home pilot data. Comparing this with the DPICS labels in Playlogue (Table 2), we see that classes such as "Negative Talk" remain underrepresented and "Neutral Talk"/"Prosocial Talk" remain the majority classes in the naturalistic, in-home data.

We later use this in-home validation dataset to evaluate the generalizability of speech and language models trained on Playlogue to other UbiComp settings. Note that due to privacy restrictions, the in-home data will not be part of the Playlogue data release and is only meant to illustrate the representativeness and generality of the Playlogue dataset to other real-world contexts of adult-child play interactions.

4.6 Ethics Statement

The audio data in Playlogue was collected by various researchers following their own institutional guidelines and contributed to the CHILDES system [64]. The audio and anonymized transcripts are available to authorized researchers through TalkBank [58]. The individual participants in each study consented to data-sharing through the TalkBank system. This research follows TalkBank's ground rules for data usage and sharing. As requested, the senior author applied to become a member of the TalkBank system and described the intended research to the TalkBank maintainers. The authors took measures to ensure that no sensitive information was revealed when curating data for Playlogue. The dataset creation procedure reported on in this work did not involve any additional human subjects research and therefore did not require any institutional approval; however, the two authors who performed DPICS coding completed human subjects research training requirements at their institution. All models tested using the data were either run on institutional clusters or via APIs that do not retain data. The in-home pilot study to validate Playlogue was separately approved by the Institutional Review Board at the authors' institution.

5 Analysis

In this section, we utilize the Playlogue dataset to evaluate the performance of several state-of-the-art speech and language models developed for adult speech on each of the benchmark tasks described in Section 3. We also

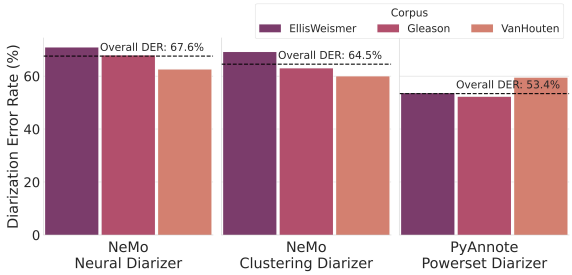

Fig. 1. Performance of pretrained diarization models on the Playlogue dataset.

Table 5. Diarization performance of PyAnnote powerset diarization model on the Playlogue dataset after fine-tuning.

Model Type	Diarization Error Rate
Pretrained Model	53.4%
Fine-tuned on play-based corpora	40.3%
Fine-tuned on all corpora	33.2%

conduct preliminary analyses to investigate whether training or fine-tuning models using child-centered speech from Playlogue can improve model performance.

5.1 Adult-Child Speaker Diarization

5.1.1 Evaluating Pretrained Models. To evaluate the feasibility of utilizing pretrained models to perform adult-child diarization, we tested three speaker diarization models trained on adult speech on the child-centered conversations in Playlogue. Models are compared using Diarization Error Rate (DER) as the metric, which is calculated as the duration of false alarms, missed detection, and speaker confusion errors divided by the total duration across all audio clips in the test set. A lower DER value implies better performance, with a DER of zero indicating perfect diarization. For all models, DER was calculated with no forgiveness collar and taking into account overlapped speech.

We randomly assigned 20% of the participants in each play-based corpus in Playlogue into test and validation sets, with the remaining 60% of the play-based corpora and all of the *Cameron* narrative corpus used for training/fine-tuning. This resulted in training, validation, and test sets of durations 19 hours and 6 mins, 5 hours and 59 mins, and 8 hours and 6 mins respectively. We report the performance of models on the test set in all our experiments henceforth. The pretrained models evaluated included the PyAnnote powerset speaker diarization model [72] and the NeMo clustering and neural diarizers [37, 69], which report state-of-the-art diarization error rates (DERs) on various (adult) speaker diarization benchmarks.

Figure 1 shows the DER of the pretrained models on each play-based corpus in Playlogue. We see that the PyAnnote model obtained the lowest overall DER of 53.4%. To contextualize this result, the model reports DERs ranging from 7.8% to 50% on benchmark diarization datasets containing adult speech. Our analysis demonstrates significant gaps in the generalizability of state-of-the-art diarization models at separating adult vs. child speech.

5.1.2 Fine-tuning on Playlogue. In order to investigate whether fine-tuning using adult-child speech improves diarization performance, we fine-tuned the speaker segmentation model in the best-performing PyAnnote diarization model from Figure 1 using the train set from Playlogue. The segmentation model was fine-tuned on either the play-based corpora only or the play-based+narrative corpora for 30 epochs with 16-bit mixed-precision training. The segmentation model with the highest performance on the validation set was then used to optimize the clustering threshold for agglomerative clustering (see [72]) by optimizing for 10 iterations. The performance of these adapted pipelines is reported in Table 5 – fine-tuning on the play-based corpora (*EllisWeismer*, *Gleason*, and *VanHouten*) alone improves DER from 53.4% to 40.3%. The performance improvement is statistically significant as indicated be a paired t-test of DERs across the 33 test files ($t(33) = 7.50, p < 0.001$). Including training data from the *Cameron* narrative corpus brings the DER further down to 33.2% ($t(33) = 2.98, p = 0.005$). This demonstrates

Table 6. Adult-child speaker diarization performance of speech classification models on the Playlogue dataset.

Model	Diarization Error Rate			Overall DER
	EllisWeismer	*Gleason*	*VanHouten*	
Fine-tuned PyAnnote diarization model (Table 5)	37.4%	24.3%	32.2%	33.2%
Classification with openSMILE features	38.1%	37.8%	44.8%	38.5%
Classification with WavLM features	33.8%	27.3%	32.2%	31.8%

the utility of adult-child speech datasets in improving the diarization performance of state-of-the-art models when used for inference in preschool-aged populations.

5.1.3 Training Adult-Child Speech Classification Models. While speaker diarization models segment speech from an arbitrary number of speakers and assign speaker labels in a permutation-invariant manner (i.e., Speaker A -> Speaker B -> Speaker A is treated as equivalent to Speaker B -> Speaker A -> Speaker B as long as speaker labels are consistent and start/end times of each speaker segment are accurate), processing adult-child conversations sometimes calls for accurately classifying which of the speakers are adults/children. One approach for obtaining these labels is to apply a diarization pipeline that outputs permutation-invariant speaker clusters and then detect whether each output cluster represents an adult or child. Another approach followed by prior work such as [43] involves training a classifier directly using audio features that can distinguish between adult and child speech.

To compare against end-to-end diarization approaches, we trained two adult-child speech classification models using different speech representation features. First, we used the openSMILE toolkit [24] to extract pitch, voice quality, energy, spectral, and cepstral features defined in the ComParE 2016 computational paralinguistics challenge [84] from the raw audio signal using a window length of 60ms. Second, we used the pretrained Microsoft WavLM-large model [14], which has been trained using large scale-self supervision and shown to learn useful representations for a range of downstream tasks, to extract hidden state representations from raw audio at a frequency of approximately 50 Hz. We then trained a simple multi-layer perceptron model with two hidden layers to classify adult/child speech using the openSMILE or WavLM features as inputs. Each model was trained for up to 30 epochs with a learning rate of 0.001 and an L2 weight decay of 0.01 to minimize multi-label binary cross entropy loss to independently predict the presence/absence of adult and child speech in each window. Binary classification thresholds for each label were selected by maximizing the F1 score on the validation set. Frame-level labels were combined to reconstruct speaker activity segments, merging segments with the same speaker label and a gap of less than 1 second between them.

Table 6 shows the DERs achieved by the classification models using openSMILE and WavLM features. While the openSMILE model performs worse than the fine-tuned PyAnnote model with a DER of 38.5% ($t(33) = -1.27, p = 0.212$), the classification model using WavLM hidden states as features outperforms the fine-tuned model and achieves an overall DER of 31.8% ($t(33) = 1.31, p = 0.198$).

5.1.4 Validating Models on In-Home Pilot Data. In order to validate whether speaker diarization models fine-tuned/trained on Playlogue generalize to real-world UbiComp settings, we evaluate their performance on the in-home pilot dataset described in Section 4.5. Table 7 shows that the fine-tuned PyAnnote model, as well as the feature-based classification models, generalize reasonably well to this unseen context, with all models achieving significantly lower DERs than the pretrained model. This demonstrates the utility of Playlogue for training adult-child speaker diarization models for other research settings.

5.1.5 Implications. Our analysis shows that existing speaker diarization models that achieve state-of-the-art DERs on adult speaker benchmarks fail to generalize out-of-the-box to adult-child speech. While fine-tuning with

Table 7. Adult-child speaker diarization performance on in-home pilot data.

Model	Diarization Error Rate
Pretrained PyAnnote diarization model	53.9%
PyAnnote diarization model fine-tuned on Playlogue	39.4%
Classification model trained on Playlogue with openSMILE features	43.1%
Classification model trained on Playlogue with WavLM features	33.0%

as little as 20 hours of adult-child data results in a significant performance improvement, we demonstrate that some applications might benefit even more from a simple, highly efficient classification approach. Specifically, extracting audio embeddings from a large pretrained model such as WavLM and training a much smaller classification model using these features achieves the lowest DER when separating adult-child audio. However, the performance of both fine-tuned and trained models may vary significantly across corpora, highlighting the need for future adult-child diarization models to be evaluated on diverse datasets such as Playlogue. We illustrate this point by evaluating these models on our in-home pilot data, where the performance trends match that of the evaluations on Playlogue.

5.2 Automatic Speech Recognition

5.2.1 Evaluating Pretrained Models. We evaluated six pretrained (on adult speech) ASR models on the Playlogue dataset to benchmark their performance on child-centered speech. For our experiments, we selected the base/large versions of Facebook's wav2vec 2.0 [5] and the medium/large versions of OpenAI Whisper [73] based on prior work demonstrating the favorable fine-tuning performance of these models on older children's speech [45]. We also evaluated the large version of Facebook's data2vec-audio [4] and the NVIDIA Canary-1B model [37, 77], which achieved state-of-the-art performance on the SUPERB benchmark [107] and the Hugging Face Open ASR leaderboard [89] respectively (as of July 2024).

We used the same participants as in Section 5.1 as our train, validation, and test subsets. We formulate our problem as long-form ASR, where we transcribe several minutes of audio (entire conversations) instead of a few-second-long clip. We performed strided, chunked inference using a chunk size of 20 seconds for wav2vec 2.0 and data2vec models, 30 seconds for Whisper models, and 40 seconds for Canary-1B. We used the official NeMo toolkit [37] for inference on Canary-1B and Hugging Face pipelines with the official releases for all other models. Performance was evaluated in terms of Word Error Rate (WER), calculated as the ratio of the total number of additions, deletions, and substitutions to the total number of words in the transcript. We calculated WERs for each model after applying Whisper's basic text normalizer [73] on reference and predicted transcriptions to remove symbols, punctuation, and extra whitespace, convert all text to lowercase, and remove words within parentheses.

Figure 2 shows the performance of each pretrained model on the test participants in different corpora within Playlogue. Whisper Large v3 (1.5 billion parameters) and Whisper Medium (769 million parameters) demonstrated the two lowest WERs of 40.84% and 49.44%, respectively. To put these error rates in context, the original version of Whisper Large reported WERs ranging from 3.52% to 19.60% on long-form transcription of adult speech (see Figure 6 in [73]).

5.2.2 Fine-tuning on Playlogue. We then fine-tuned Whisper Large v3 on the train split from Playlogue. Models were fine-tuned for 10 epochs with 16-bit mixed precision training and gradient checkpointing. We used a single GPU with a train and validation batch size of 32 and the AdamW optimizer [57] with a linear warmup to a learning rate of 1e-5. As shown in Table 8, fine-tuning on only the play-based corpora (*EllisWeismer*, *Gleason*,

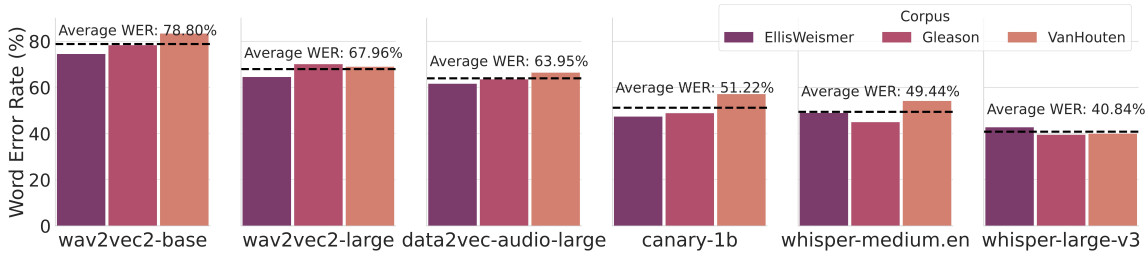

Fig. 2. Performance of pretrained ASR models on the Playlogue dataset.

Table 8. ASR performance on the Playlogue dataset after fine-tuning. ↓ implies a lower score is better while ↑ indicates that a higher score is better.

Model	Type	Average Word Error Rate ↓	Average BERT F1 Score ↑
whisper-large-v3	Pretrained Model	40.84%	91.40%
	Fine-tuned on play-based corpora from Playlogue	31.43%	92.84%
	Fine-tuned on all corpora from Playlogue	29.37%	92.58%

and *VanHouten*) led to a reduction in average WER from 40.84% to 31.43% ($t(33) = 6.22, p < 0.001$). Including the narrative *Cameron* corpus in the fine-tuning data improved the performance marginally, achieving a WER of 29.37% ($t(33) = 0.91, p = 0.368$). We also report the performance of the pretrained and fine-tuned models in terms of the BERTScore, which evaluates text generation by measuring the cosine similarity between words in the reference and predicted transcriptions using pretrained embeddings from a BERT model [110]. BERTScore has been shown to correlate with human judgment on sentence-level evaluation. Table 8 shows that while all models achieve high BERT F1 scores, there is a small improvement after fine-tuning. This is likely due to the fact that our test audio clips are several minutes long, which means the semantic meaning of a conversation can still be captured fairly accurately even from an erroneous transcription. However, it is important to note that several applications in child-centered language processing (e.g., morphological/phonetic analysis, measuring lexical diversity during language acquisition, etc.) still require accurate transcripts with a low WER. In this regard, it is clear that fine-tuning with in-domain adult-child speech leads to a significant performance improvement.

We further investigate the effect of fine-tuning corpus size, using 5, 10, or 15 hours of training data to fine-tune the model. Fine-tuning with as little as 5 hours of data reduces the average WER from 40.84% to 30.26%, demonstrating significant performance improvement across test corpora (Figure 3). However, additional training data beyond 5 hours does not improve performance on all corpora – while the WER on the *EllisWeismer* corpus continues decreasing, WERs on *Gleason* and *VanHouten* start to increase. This suggests that the model might be overfitting to the *EllisWeismer* corpus, which is over-represented in the training data, thus hurting generalizability.

To further test generalizability in low training-data regimes, we fine-tuned models with a fixed amount (5 hours) of training data from the Playlogue dataset. Figure 4 shows the performance of the models on each corpus when fine-tuned with and without data from the same corpus. As expected, including fine-tuning data from non-overlapping participants in the test corpus generally results in lower WERs. However, there is variability across corpora in terms of performance gains – while the *Gleason* corpus benefits from fine-tuning on other corpora, there is a slight regression in WER when fine-tuned using participants from the same corpus. On the

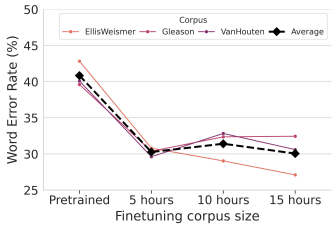

Fig. 3. Performance of ASR models fine-tuned with varying amounts of data.

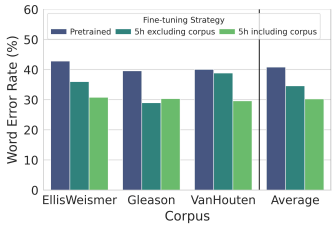

Fig. 4. Generalizability of fine-tuned ASR models to unseen corpora.

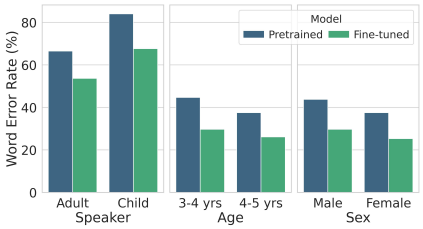

Fig. 5. Performance of ASR models by speaker attributes.

Table 9. Performance of ASR models on in-home pilot data after fine-tuning on Playlogue. ↓ implies a lower score is better while ↑ indicates that a higher score is better.

Model	Type	Word Error Rate ↓	BERT F1 Score↑
whisper-large-v3	Pretrained Model	40.22%	94.00%
	Fine-tuned on play-based corpora from Playlogue	19.09%	94.20%
	Fine-tuned on all corpora from Playlogue	11.80%	96.60%

other hand, the WER reduction on the *VanHouten* corpus is marginal when fine-tuned on other corpora, but substantial when fine-tuned with data from the same corpus.

5.2.3 ASR Performance by Attribute. Finally, we evaluated the WER of the pretrained and fine-tuned Whisper Large models across speaker characteristics (see Figure 5. First, we computed WERs separately for speech segments from adults and children. Note that this leads to an inflated WER estimate since we count erroneous word insertions made by the model between adult and child segments twice (once toward the computation of adult WER and once toward that of child WER). Nevertheless, we observe that, as expected, both pretrained and fine-tuned models perform better on adult speech than child speech. However, fine-tuning the model still results in improved performance for adult speech in addition to child speech. We hypothesize that this is because fine-tuning helps the model learn to better recognize the characteristic "baby talk" voice that adults tend to employ when speaking with children. From Figure 5, we also note that the WERs of both models are lower for older children (4-5 years) than younger children (3-4 years) and for female children compared to male children. This highlights the need for testing and improving the performance of ASR models across subgroups even within an early-childhood population.

5.2.4 Validating Models on In-Home Pilot Data. Following our analysis on Playlogue, we evaluate both pretrained and fine-tuned ASR models on our in-home adult-child conversations. Table 9 demonstrates the utility of fine-tuning the Whisper Large v3 on Playlogue, showing a significant drop in WER from 40.22% to 19.09% and 11.80% after fine-tuning with play-based and mixed corpora respectively. We also see a slight improvement in the model's BERT F1 scores after fine-tuning. These results highlight the need to adapt state-of-the-art speech processing models before using them in child-centered UbiComp settings.

5.2.5 Implications. We show that state-of-the-art ASR models exhibit significant performance gaps when tested on both child speech and child-directed adult speech. Our findings suggest that researchers and practitioners should exercise caution when using off-the-shelf models and APIs in voice-based applications that interface

Table 10. Accuracy and F1 scores of the TF-IDF SVM baselines (reimplementation from [43] trained on both the SpecialTime [43] dataset and training data from Playlogue) at parent DPICS classification on our test set.

Models	Accuracy	F1 Score
TF-IDF SVM model [43]		
- Trained on SpecialTime dataset [43]	54.0%	25.31%
- Trained on Playlogue	68.0%	21.58%

with children, and, at minimum, should evaluate ASR components on child speech datasets such as Playlogue before deploying them in the wild. We show that fine-tuning with as little as 5 hours of data shows a significant reduction in WER for Whisper, and that fine-tuning on the Playlogue dataset greatly improves ASR performance on child-centered speech from an unseen corpus.

5.3 Automatic Coding of Speech Acts Using DPICS

5.3.1 Evaluating Baseline Text Classification Models for Parent Speech Acts. We now examine the feasibility of using text classification models on transcriptions of parent-child conversations to identify DPICS parent codes, evaluating the approach proposed by Huber et al. [43]. For this evaluation, we used the validation and test participants in *EllisWeismer* who completed the parent-child session from among the train/val/test splits defined in Section 5.1 as our test set for DPICS labeling. This resulted in 13 out of 27 clips in the test set and the rest in the train set. The test set contained 2176 labeled utterances ("Negative Talk": 24, "Command": 187, "Labeled Praise": 5, "Unlabeled Praise": 81, "Question": 738, "Reflective Statement": 133, "Behavior Description": 20, "Neutral Talk": 988).

The proposed approach for detecting DPICS parent labels in Huber et al. [43] involved extracting a feature vector for each parent dialogue (i.e., single sentence) using the information-theoretic measure term frequency-inverse document frequency (TF-IDF; [74]). TF-IDF vectors are derived from unigram and bigram features of parent dialogues in the training set using both words as well as part-of-speech (POS) tags. The authors used these features to train a linear support vector machine (SVM) with $C = 0.1$ as the text classifier predicting DPICS labels. While this classifier was used to label 7 of the 8 DPICS parent classes in [43], "Question" was labeled using audio features.

We first trained a TF-IDF SVM model using a similar approach and with the same dataset as used in Huber et al. [43] and evaluated it on our test set. Diverging slightly from their approach, we used this text classification model to also detect the "Question" labels. The TF-IDF model achieved an overall accuracy of 54% on our test set, compared to the 78.3% accuracy reported on the SpecialTime dataset [43]. This demonstrates that the TF-IDF approach does not generalize well to unseen datasets and contexts and underscores the importance of training the TF-IDF model on a dataset that closely mirrors the distribution of the test set.

We then trained a similar TF-IDF SVM model on our train set (14 clips) and evaluated it on our test set (13 clips). While this resulted in a significant improvement in accuracy from 54% to 68%, the performance was further skewed, with a drop in F1 score from 25.31% to 21.58%. This could be due to the severely low representation of some DPICS classes in our training set – for example, the training set contained only 7 examples of "Labeled Praise" and 15 examples of "Behavior Description", but 1183 examples of "Neutral Talk". While this is a limiting factor, it is nevertheless representative of the distribution of class labels that could be expected in naturalistic conversations.

Table 11. Accuracy and F1 scores of large language models (without fine-tuning) and sentence transformer models (with fine-tuning) at parent and child DPICS classification on our test set.

Models	Parent DPICS (8-class classification)		Child DPICS (4-class classification)	
	Accuracy	F1 Score	Accuracy	F1 Score
Single Sentence Classification				
Large Language Models (no fine-tuning)				
- Llama3 (8B)	64.0%	29.66%	82.95%	44.0%
- GPT-3.5	72.20%	33.92%	78.51%	44.94%
- GPT-4	77.70%	48.09%	87.35%	52.87%
Sentence Transformer Model (fine-tuned on Playlogue)				
- MPNet-base sentence embedding model	81.80%	58.49%	91.91%	72.17%
- RoBERTa-large sentence embedding model	82.31%	61.04%	92.02%	72.46%
Classification with Context				
Large Language Models (no fine-tuning)				
- GPT-4 (variable window)	77.0%	44.20%	86.70%	52.29%
- GPT-4 (fixed window, 10 sentences)	78.60%	49.26%	86.49%	43.23%
Sentence Transformer Model (fine-tuned on Playlogue)				
- RoBERTa-large (fixed window, 5 sentences)	85.16%	66.55%	91.85%	72.31%

5.3.2 Evaluating Transformer-based Language Models for Parent and Child Speech Acts. Next, we investigate whether some of the limitations of baseline models such as TF-IDF can be overcome using recent transformer-based language models by leveraging their large-scale pretrained capabilities. In addition to evaluating models for classifying parent DPICS labels as described above, we also investigate classification performance on child DPICS labels from the 13 test clips. The test set contained 1866 labeled child utterances ("Negative Talk": 27, "Command": 77, "Question": 249, "Prosocial Talk": 1513).

We consider two different approaches toward labeling parent and child utterances: (i) classification using chain-of-thought (CoT) prompting with state-of-the-art pretrained large language models (LLMs) and (ii) fine-tuning pretrained sentence transformers with a limited amount of labeled data. With the former approach, we aim to evaluate whether LLMs can achieve high levels of accuracy in classifying parent and child speech acts without requiring fine-tuning on our specific context or dataset. This is highly desirable because while LLMs have been shown to achieve impressive performance at language classification tasks, fine-tuning them is often prohibitively time-consuming and resource-intensive. CoT prompting, on the other hand, has previously been shown to be an effective approach for text classification [100] . With the latter approach, we intend to find a middle ground through sentence transformer fine-tuning. Specifically, we use the SetFit framework for efficient training of text classification models with limited training data. This allows us to adapt pretrained models to our specific domain with less than 5 minutes of training on a single GPU with 80GB of VRAM. We now describe both these approaches and the resulting performance of each in more detail.

For our first approach, we considered state-of-the-art **large language models** including OpenAI's GPT-4 [2] and GPT-3.5 [68] as well as Meta Llama3 (8B) [3] (representing large, medium, and small LLM categories respectively). The evaluation of GPT models was done using the OpenAI API The Llama3 (8B) model was loaded locally using the official HuggingFace pipeline on a single GPU. We used default hyperparameters for all models and provided them with the same prompt to maintain consistency. Each query consisted of two parts, a "fixed"'

prompt and a "test" prompt. The fixed prompt introduces PCIT and/or DPICS, describes the task we expect the target LLM to perform, explains the data that will be provided as input, and defines the candidate classes with a few examples of each class from the DPICS manual [25]. Note that the fixed prompt does not include any examples from our training set in order to estimate performance on any arbitrary parent-child interaction data. We reproduce the full fixed prompt for both the parent and child DPICS classification tasks in Appendix A. The test prompt then provides the parent or child utterance (as a single sentence) to be classified by the LLM.

Table 11 shows the performance of each of the LLMsfor both parent and child DPICS classification on our test set. Relative to the baseline models in Table 10, LLMs trained on a huge corpus of data are less sensitive to variations in test distribution. As expected, bigger LLMs have higher performance, with GPT-4, GPT-3.5, and Llama3 (8B) achieving 77.7%, 72.2%, and 64% accuracy respectively at parent DPICS classification, with F1 scores also increasing with model size. The per-class recall of GPT-4, the best performing LLM, was 54.17%, 74.33%, 60.00%, 58.02%, 90.24%, 2.26%, 60.00%, and 81.78% for "Negative Talk", "Command", "Labeled Praise", "Unlabeled Praise", "Question", "Reflective Statement", "Behavior Description", and "Neutral Talk", respectively. As expected, the model is especially poor at detecting "Reflective Statements", which, by definition, can only be identified using knowledge of the previous child dialogue. However, GPT-4 is reasonably accurate at detecting most other DPICS parent classes even without seeing any examples from our dataset. For child DPICS labeling, while Llama3 outperforms GPT-3.5 in terms of accuracy, we see a consistent improvement in F1 scores with increasing model size. GPT-4 again emerges as the best-performing LLM, with an accuracy of 87.35% and F1 score of 52.8%. The per-class recall was 96.29%, 58.44%, 95.18%, and 87.37% for "Negative Talk", "Command", "Question", and "Prosocial Talk" respectively, demonstrating highly accurate performance for three of the four categories with no in-domain training data. Given that data collection and annotation for applications such as DPICS labeling are both highly complex and time-consuming, the generalizability of LLMs can be invaluable in offering a way to circumvent these extensive processes.

Next, we turn to our second approach of **fine-tuning sentence transformer models** using the training data from Playlogue and evaluating their performance on parent and child DPICS labeling. Briefly, SentenceTransformers [76] are text embedding models that have been used for a range of applications such as text classification, semantic similarity, semantic search and retrieval etc. These sentence embedding models have been developed by fine-tuning pretrained transformer models on over 1 billion sentence pairs using contrastive objectives, and are designed to be general-purpose models that can be efficiently adapted for different text-based tasks. We use the sentence transformer models based on MPNet-base [87] and RoBERTa-large [56] in our experiments, representing small- and medium-sized models that achieve high performance in the SBERT sentence embedding benchmark [1] as of July 2024. To adapt these models to the DPICS classification task, we fine-tune them using (sentence, label) pairs from the Playlogue train set using the SetFit framework. This involves fine-tuning the embedding model using a triplet loss [41] followed by training a simple logistic regression classification head that utilizes the fine-tuned embeddings. We separately train and evaluate models for both parent and child DPICS codes.

As shown in Table 11, both the fine-tuned sentence transformer models outperform the pretrained LLMs at parent and child DPICS classification. The larger RoBERTa models show slight improvement over the smaller MPNet models, especially in terms of parent F1 scores. Figures 6 and 7 show the confusion matrices for the parent and child models respectively. We see that both models are highly accurate at identifying the "Question" and "Neutral Talk"/"Prosocial Talk" categories. Similar to LLMs, the sentence transformer-based parent DPICS model is least accurate at identifying the "Reflective Statement" label. Below, we attempt to address this performance gap by including contextual information during classification.

5.3.3 Model Performance with Additional Context. While our evaluations of DPICS classification models have so far considered the problem of labeling each parent or child utterance on its own, human raters labeling speech acts often take into account the underlying context of a conversation when assigning labels. This is

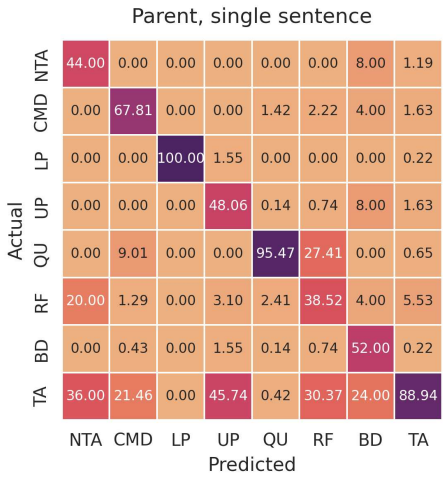

Fig. 6. Confusion matrix showing the predictive performance of sentence transformer model for parent DPICS labeling using a single sentence. The average classification accuracy is 82.31%.

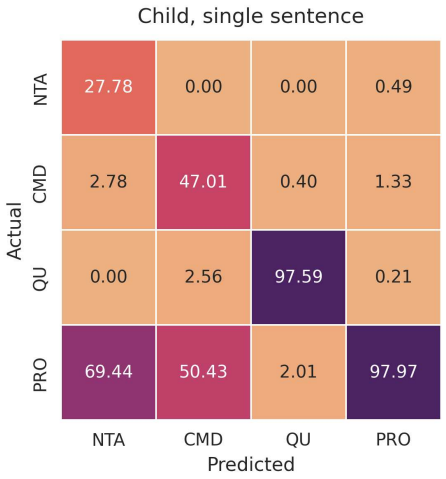

Fig. 7. Confusion matrix showing the predictive performance of sentence transformer model for child DPICS labeling using a single sentence. The average classification accuracy is 92.02%.

especially important for some labels – for example, identifying a "Reflective Statement" requires knowledge of child dialogue(s) that came before the parent dialogue. We hypothesize that providing such interaction context would therefore improve DPICS classification performance of both LLMs and sentence transformer models.

We first evaluate the **effect of additional context on LLM performance** by segmenting each clip in our test set into non-overlapping "windows". In our "variable window" approach, we segment the clip such that each window contains one conversation turn ending with the target speaker. Concretely, for *parent* DPICS classification, a "variable window" would include one child turn (any number of consecutive child sentences) followed by one parent turn (consecutive parent sentences). Conversely, for *child* DPICS classification, the "variable window" consists of one parent turn followed by one child turn. We also consider another approach, called the "fixed window" approach, which groups 10 consecutive sentences irrespective of the speaker. This is equivalent to a non-overlapping sliding window with a size of 10 sentences. Figure 10 illustrates these approaches with examples for the case of parent DPICS labeling. We include these "variable" and "fixed" windows in our test prompts to GPT-4 (the best-performing LLM from the previous evaluation) instead of a single parent or child sentence.

Table 11 shows the performance of GPT-4 under both the "variable" and "fixed window" settings. While the "variable window" approach leads to worse performance compared to the single sentence setting for both parent and child DPICS classification, we see a small improvement in parent accuracy and F1 score using the "fixed window" approach. The per-class recall scores for parent DPICS using a 10-sentence window were 66.67%, 80.21%, 40.00%, 53.09%, 79.40%, 54.14%, 60.00%, and 83.91% for "Negative Talk", "Command", "Labeled Praise", "Unlabeled Praise", "Question", "Reflective Statement", "Behavior Description", and "Neutral Talk" respectively. As expected, we see the largest improvement in identifying "Reflective Statement", with a jump from 2.26% to 54.14%. We also see an improvement in the classification of "Negative Talk", "Command", "Behavior Description" and "Neutral Talk", but more confusion between "Labeled Praise" and "Unlabeled Praise" when using the "fixed window" approach. Overall, we find that test prompts with multiple dialogues provide a more balanced performance across all classes for parent labels compared to the single sentence mode, emphasizing the value of DPICS datasets

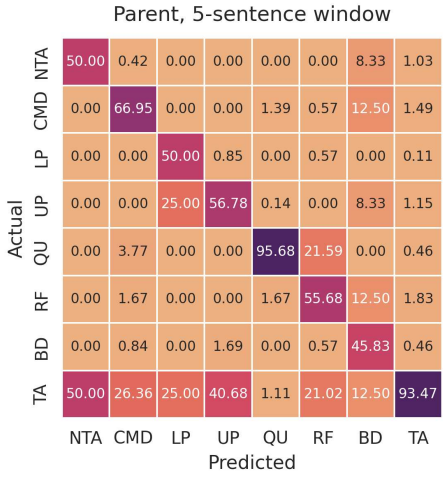

Fig. 8. Confusion matrix showing the predictive performance of sentence transformer model for parent DPICS labeling using a 5-sentence window. The average classification accuracy is 85.16%.

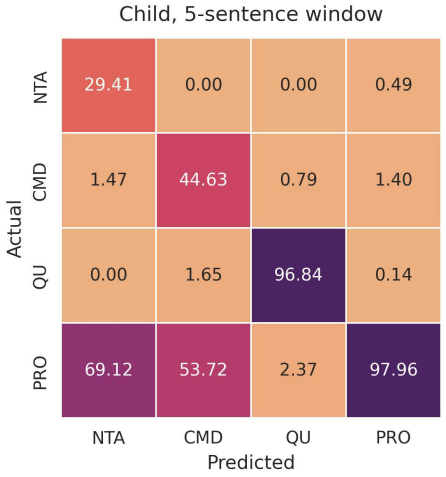

Fig. 9. Confusion matrix showing the predictive performance of sentence transformer model for child DPICS labeling using a 5-sentence window. The average classification accuracy is 91.85%.

that provide as much conversation context as possible rather than single/pairs of sentences. However, for child DPICS classification, we see that GPT-4 performs best when using a single sentence as input, as opposed to a "variable" or "fixed" window. This is likely because the child DPICS classes rely less on contextual information and can be identified based on the structure and semantics of an individual sentence.

Next, we evaluate the **effect of contextual information on fine-tuned sentence transformer models**. We select the RoBERTa model since it achieved the highest classification performance in the single-sentence scenario. Instead of training the classification head with a (sentence, label) pair, we use (window, label) pairs with windows containing a fixed number of sentences and ending with the target sentence that is to be classified. Since the maximum sequence length of the sentence embedding model is only 256 words, we use a "fixed window" of 5 sentences (4 preceding sentences + target sentence). We use the same training paradigm to obtain both parent and child DPICS classification models separately.

As seen in Table 11 we see a substantial improvement in parent accuracy (from 82.31% to 85.16%) and F1 score (from 61.04% to 66.55%) when using a sentence transformer model trained with the "fixed window" approach. Figure 8 shows the confusion matrix for the parent sentence transformer model using 5-sentence windows. Comparing the per-class recall to that in Figure 6 we see an improvement in prediction performance for "Negative Talk", "Question", "Reflective Statement", and "Neutral Talk", but a regression for "Labeled Praise", "Unlabeled Praise", and "Behavior Description". Similar to the context-aware LLM scenario, this improvement is not seen in the child DPICS classification model, where the single-sentence model (Figure 7) outperforms the fixed-window model (Figure 9).

5.3.4 Validating Models on In-Home Pilot Data. We now evaluate the performance of both the best-performing LLM (GPT-4) and sentence transformer model (RoBERTa fine-tuned on Playlogue) on parent and child DPICS labels from our in-home pilot data. Table 12 shows the performance of these models both when using a single sentence or a fixed window as input. First, we observe that the same CoT prompting approach on GPT-4 produces slightly worse results on the in-home data compared to the Playlogue test set. While the pretrained GPT-4 model

Table 12. Accuracy and F1 scores of GPT-4 (without fine-tuning) and RoBERTa sentence transformer model (fine-tuned on Playlogue) at parent and child DPICS classification on the in-home pilot data.

Models	Parent DPICS (8-class classification)		Child DPICS (4-class classification)	
	Accuracy	F1 Score	Accuracy	F1 Score
Single Sentence Classification				
- GPT-4	74.8%	54.12%	82.35%	54.40%
- RoBERTa-large sentence transformer model	72.93%	52.34%	88.24%	74.17%
Classification with Context				
- GPT-4 (fixed window, 10 sentences)	71.80%	59.85%	82.69%	54.22%
- RoBERTa-large sentence transformer model (fixed window, 5 sentences)	76.34%	55.94%	89.27%	75.60%

has not been fine-tuned for the DPICS classification task using data from Playlogue, it is evident that LLMs such as GPT may not achieve consistent performance across datasets and should therefore be carefully evaluated on a target dataset. Second, we also observe a drop in performance for the sentence transformer model when evaluated on the in-home pilot data. However, the sentence transformer model trained on Playlogue with a 5-sentence window still achieves the best performance on both parent and child DPICS classification in the in-home setting. This demonstrates the utility of Playlogue's continuously labeled conversations in training context-aware and generalizable speech act classification models for adult-child dialogues.

5.3.5 Implications. Operating under the assumption that we are able to accurately diarize and transcribe parent and child speech, we investigated the ability of various text classification models such as TF-IDF retrieval models, sentence transformer models, and emerging LLMs to automatically predict parent and child DPICS labels. We found that LLMs and sentence transformer models offer a plausible alternative to traditional TF-IDF models that do not generalize well to new settings and datasets. Pretrained LLMs can achieve reasonable DPICS classification accuracy across datasets even without fine-tuning, which can be invaluable in scenarios where data collection and annotation are challenging and expensive. However, sentence transformer models offer a powerful alternative in scenarios where small to moderate amounts of labeled data are available. We demonstrate that less than 5 minutes of fine-tuning with labeled data from Playlogue can result in fairly generalizable DPICS classification models. We also show that conversational context leads to significant performance improvements in parent DPICS classification, highlighting the utility of datasets such as Playlogue.

6 Discussion

6.1 Rationale for a New Dataset of Adult-Child Conversations During Play

As discussed in Section 2, there is significant interest in the UbiComp community to develop technologies to support adult-child (especially parent-child) interactions. Play sessions are an important context for many of these interactions both within UbiComp [43, 44, 88] and in the broader scientific community [31, 91]. However, there is a lack of publicly available datasets of naturalistic, play-based adult-child conversations that could accelerate the development and evaluation of ubiquitous technologies in this domain.

There have been several attempts to bridge this gap over the past few decades, ranging from the development of the CHILDES system [64] to the release of example parent-child interaction labels in SpecialTime [43]. Although CHILDES contains an unprecedented collection of open speech and language data, it has seen limited adoption

in the machine learning and speech processing research communities due to the relatively capricious quality of audio, transcripts, and time synchronization data across its constituent corpora [21]. In this work, we utilize a hybrid approach consisting of manual filtering and automated readjustments that allows us to leverage a significant portion of CHILDES data to create Playlogue.

In terms of identifying speech acts during adult-child interactions, the two sources of DPICS parent codes available to the research community – SpecialTime [43] and the DPICS coding manual [25] – both contain synthetic examples without contextualizing them within a longer conversation between the parent and child (SpecialTime further only contains parent labels). As a result, models trained on these datasets cannot learn to take previous utterances from the conversation into account even though this would improve classification accuracy (as shown in Section 5.3). Similarly, the lack of audio data in existing sources precludes researchers from training audio or multimodal models for DPICS classification. The Playlogue dataset addresses these two limitations of existing datasets by providing full conversation audio and sentence-level DPICS codes for both parent and child utterances.

The lack of existing datasets of preschool-aged child and child-directed adult speech has also prevented a thorough evaluation of existing speech processing approaches on this population. As a result, researchers and practitioners continue to use off-the-shelf models that exhibit state-of-the-art performance on adult speech in applications that interface with children, without accounting for potential performance gaps and biases. As shown in Section 5, there is indeed a significant loss of accuracy when following this approach. Datasets such as Playlogue can enable further evaluation of existing models prior to deployment with child populations. As demonstrated by the encouraging generalization performance on our in-home pilot data, datasets like Playlogue can also support researchers in training or fine-tuning custom models for this purpose.

6.2 Opportunities for Future Research

As described above, the Playlogue dataset will enable researchers working on child-centered speech and language applications to develop and evaluate machine learning models for analyzing various levels of adult-child interactions during play. Researchers and practitioners can utilize Playlogue to model low-level vocal behaviors (e.g., tone, pitch), diarization-based metrics such as turn durations and overlaps, ASR-based measures like word complexity and diversity, and higher-level interaction markers such as DPICS codes. The Playlogue dataset contains naturalistic audio recorded in various home and lab-based settings with a variety of microphones and noise conditions and contains 110 unique participants in the age range of 3 to 5 years, making it especially suited for building and evaluating ubiquitous technologies for this population. The availability of full conversation audio, timestamped and speaker-attributed transcripts, and DPICS codes enables researchers to study conversation dynamics in addition to sentence-level measures. For instance, future work could examine transitions from one DPICS label to another or investigate which child utterances are most likely to follow a given parent DPICS label.

The performance gaps in using state-of-the-art models on child-centered speech that have been highlighted in Section 5 also demonstrate how there is substantial work still to be done in order to build automated adult-child conversation analysis pipelines that are accurate and generalizable. The cascading errors from each step in the pipeline could amount to alarmingly poor accuracy in predicting higher-level insights like DPICS codes. While holistic performance improvement and system building are beyond the scope of the current work, future research can investigate ways to mitigate some of the challenges associated with these goals. For example, prior work has utilized LLMs to correct transcription errors in ASR or used multimodal audio-text models for zero-shot audio classification into relevant categories. The Playlogue can be used to develop and evaluate such end-to-end approaches for DPICS labeling. It can also be used to assess smaller speech and language models (or distilled/quantized versions of models) that are suitable for on-device deployment.

Finally, future research could build on Playlogue by adding annotations for other behavioral markers, such as perceived emotion or engagement levels, that are of interest within the context of play-based interactions. Playlogue also only focuses on English-speaking children in North America with typically developing speech. As such, our analysis and results only apply to this target demographic and should not be assumed to generalize to other contexts. We encourage future researchers to curate similar datasets that span more diverse subgroups within the early childhood population.

6.3 Data Availability

The original audio files used to create Playlogue can be downloaded from CHILDES [64]. The audio processing metadata and scripts, generated transcripts with token-, word-, and sentence-level alignments, speaker diarization data, DPICS labels, and train/validation/test participant splits are all made available at https://huggingface.co/datasets/playlogue/playlogue-v1. Researchers who wish to use Playlogue will be asked to abide by the CHILDES [64] and TalkBank [58] ground rules for data use and sharing.

7 Conclusion

This paper presents Playlogue, a dataset of naturalistic adult-child interactions during play. Playlogue contains over 33 hours of play-based audio and 5 hours of annotated conversations with 4773 parent and 3895 child DPICS labels. We use Playlogue to evaluate state-of-the-art speech processing models trained on adult data and expose performance gaps when applied to child-centered speech. We demonstrate that fine-tuning with data from Playlogue results in significant performance boosts but that there is still scope for improvement. Further, we investigate whether pretrained large language models and fine-tuned sentence transformer models can be leveraged to derive higher-level insights from adult-child conversations. We run an in-home pilot study to validate the generalizability of models trained on Playlogue to real-world UbiComp contexts. We describe opportunities for future research using Playlogue and discuss open challenges in developing ubiquitous technologies to support adult-child interactions.

Acknowledgments

This work was supported by the National Institute of Mental Health grant R21 MH126326.

References

[1] SentenceTransformers (SBERT). *Pretrained Models*. URL: https://web.archive.org/web/20240730194049/https://www.sbert.net/docs/sentence_transformer/pretrained_models.html#original-models.

[2] Josh Achiam et al. “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (2023).

[3] AI@Meta. “Llama 3 Model Card”. In: (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[4] Alexei Baevski et al. “Data2vec: A general framework for self-supervised learning in speech, vision and language”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 1298–1312.

[5] Alexei Baevski et al. “wav2vec 2.0: A framework for self-supervised learning of speech representations”. In: *Advances in neural information processing systems* 33 (2020), pp. 12449–12460.

[6] Vivek Bhardwaj et al. “Automatic speech recognition (asr) systems for children: A systematic literature review”. In: *Applied Sciences* 12.9 (2022), p. 4419.

[7] Claire Blewitt et al. “Strengthening the quality of educator-child interactions in early childhood education and care settings: A conceptual model to improve mental health outcomes for preschoolers”. In: *Early Child Development and Care* (2020).

[8] Kathleen Bloom, Ann Russell, and Karen Wassenberg. "Turn taking affects the quality of infant vocalizations". In: *Journal of child language* 14.2 (1987), pp. 211–227.
[9] Daniel Bone et al. "Acoustic-prosodic, turn-taking, and language cues in child-psychologist interactions for varying social demand." In: *INTERSPEECH*. 2013, pp. 2400–2404.
[10] Eric Booth et al. "Evaluating and improving child-directed automatic speech recognition". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020, pp. 6340–6345.
[11] C. E. Cameron et al. "Technical codebook for Project Equity: A study to capture, appreciate, and understand young children's language diversity." In: (2023).
[12] Erica A Cartmill et al. "Quality of early parent input predicts child vocabulary 3 years later". In: *Proceedings of the National Academy of Sciences* 110.28 (2013), pp. 11278–11283.
[13] Meng-Ying Chan et al. "WAKEY: assisting parent-child communication for better morning routines". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2017, pp. 2287–2299.
[14] Sanyuan Chen et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), pp. 1505–1518.
[15] Eunji Chong et al. "Detecting gaze towards eyes in natural social interactions and its use in child assessment". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (2017), pp. 1–20.
[16] Jacob Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.
[17] Allison M Cotter and Elizabeth Brestan-Knight. "Convergence of parent report and child behavior using the Dyadic Parent-Child Interaction Coding System (DPICS)". In: *Journal of Child and Family Studies* 29.11 (2020), pp. 3287–3301.
[18] Alejandrina Cristia et al. "A thorough evaluation of the Language Environment Analysis (LENA) system". In: *Behavior research methods* 53 (2021), pp. 467–486.
[19] Alejandrina Cristia et al. "Talker diarization in the wild: The case of child-centered daylong audiorecordings". In: *Interspeech 2018*. 2018, pp. 2583–2587.
[20] Griffin Dietz et al. "Storycoder: Teaching computational thinking concepts through storytelling in a voice-guided app for children". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–15.
[21] Satwik Dutta et al. "Challenges remain in building ASR for spontaneous preschool children speech in naturalistic educational environments". In: *ISCA INTERSPEECH-2022* (2022).
[22] Batya Elbaum, Lynn K Perry, and Daniel S Messinger. "Investigating children's interactions in preschool classrooms: An overview of research using automated sensing technologies". In: *Early childhood research quarterly* 66 (2024), pp. 147–156.
[23] Maxine Eskenazi, Jack Mostow, and David Graff. *The CMU Kids Corpus*. URL: https://doi.org/10.35111/b4v0-ff65.
[24] Florian Eyben, Martin Wöllmer, and Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor". In: *Proceedings of the 18th ACM international conference on Multimedia*. 2010, pp. 1459–1462.
[25] Sheila M Eyberg et al. "Manual for the dyadic parent-child interaction coding system third edition". In: *Unpublished Rating Manual* (2004).
[26] SM Eyberg and B Funderburk. "Parent-child interaction therapy protocol". In: *Gainesville, FL: PCIT International* (2011).
[27] Siyuan Feng et al. "Towards inclusive automatic speech recognition". In: *Computer Speech & Language* 84 (2024), p. 101567.

[28] G David Forney. "The viterbi algorithm". In: *Proceedings of the IEEE* 61.3 (1973), pp. 268–278.
[29] Catherine Garvey and Ginger Berninger. "Timing and turn taking in children's conversations". In: *Discourse processes* 4.1 (1981), pp. 27–57.
[30] Jill Gilkerson et al. "Mapping the early language environment using all-day recordings and automated analysis". In: *American journal of speech-language pathology* 26.2 (2017), pp. 248–265.
[31] Kenneth R Ginsburg, Committee on Psychosocial Aspects of Child, Family Health, et al. "The importance of play in promoting healthy child development and maintaining strong parent-child bonds". In: *Pediatrics* 119.1 (2007), pp. 182–191.
[32] Jen J Gong et al. "Towards an Automated Screening Tool for Developmental Speech and Language Impairments." In: *Interspeech*. 2016, pp. 112–116.
[33] Alex Graves et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 369–376.
[34] Rebecca Grzadzinski et al. "Measuring changes in social communication behaviors: preliminary development of the Brief Observation of Social Communication Change (BOSCC)". In: *Journal of autism and developmental disorders* 46 (2016), pp. 2464–2479.
[35] Joshua Hailpern, Karrie Karahalios, and James Halle. "Creating a spoken impact: encouraging vocalization through audio visual feedback in children with ASD". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2009, pp. 453–462.
[36] John HL Hansen et al. "Speech and language processing for assessing child–adult interaction based on diarization and location". In: *International journal of speech technology* 22 (2019), pp. 697–709.
[37] E Harper et al. *NeMo: A toolkit for conversational AI and large language models*. URL: https://nvidia.github.io/NeMo/.
[38] Betty Hart and Todd R Risley. "Meaningful differences in the everyday experience of young American children". In: *Community Alternatives* 8 (1996), pp. 92–93.
[39] Kunlei He et al. "A Home Study of Parent-Child Co-Reading with a Bilingual Conversational Agent". In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–8.
[40] John Heilmann et al. "Utility of the MacArthur—Bates Communicative Development Inventory in identifying language abilities of late-talking and typically developing toddlers". In: (2005).
[41] Alexander Hermans, Lucas Beyer, and Bastian Leibe. "In defense of the triplet loss for person re-identification". In: *arXiv preprint arXiv:1703.07737* (2017).
[42] Erika Hoff. "The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech". In: *Child development* 74.5 (2003), pp. 1368–1378.
[43] Bernd Huber et al. "SpecialTime: Automatically detecting dialogue acts from speech to support parent-child interaction therapy". In: *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 2019, pp. 139–148.
[44] Inseok Hwang et al. "TalkBetter: family-driven mobile intervention care for children with language delay". In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 2014, pp. 1283–1296.
[45] Rishabh Jain et al. "Adaptation of Whisper models to child speech recognition". In: *arXiv preprint arXiv:2307.13008* (2023).
[46] Eunkyung Jo et al. "GeniAuti: Toward Data-Driven Interventions to Challenging Behaviors of Autistic Children through Caregivers' Tracking". In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW1 (2022), pp. 1–27.

[47] Eunkyung Jo et al. "MAMAS: supporting parent–child mealtime interactions using automated tracking and speech recognition". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW1 (2020), pp. 1–32.
[48] Ioannis Katsantonis and Ros McLellan. "The role of parent–child interactions in the association between mental health and prosocial behavior: Evidence from early childhood to late adolescence". In: *International Journal of Behavioral Development* 48.1 (2024), pp. 59–70.
[49] James Kennedy et al. "Child speech recognition in human-robot interaction: evaluations and recommendations". In: *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. 2017, pp. 82–90.
[50] Prasanna V Kothalkar et al. "Tagging child-adult interactions in naturalistic, noisy, daylong school environments using i-vector based diarization system". In: *ISCA SLaTE-2019 Workshop*. Vol. 1. 1. 2020.
[51] Suchitra Krishnamachari et al. "Developing neural representations for robust child-adult diarization". In: *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2021, pp. 590–597.
[52] Manoj Kumar et al. "Improving speaker diarization for naturalistic child-adult conversational interactions using contextual information". In: *The Journal of the Acoustical Society of America* 147.2 (2020), EL196–EL200.
[53] Taeahn Kwon et al. "Captivate! contextual language guidance for parent–child interaction". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–17.
[54] Golan Levy et al. "GECKO - A Tool for Effective Annotation of Human Conversations". In: *20th Annual Conference of the International Speech Communication Association, Interspeech 2019*. Herzliya, Israel, Sept. 2019. URL: https://github.com/gong-io/gecko/blob/master/docs/gecko_interspeech_2019_paper.pdf.
[55] Chaohao Lin et al. "Assessment of Parent–Child Interaction Quality from Dyadic Dialogue". In: *Applied Sciences* 13.20 (2023), p. 11129.
[56] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.
[57] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).
[58] Brian MacWhinney. "The talkbank project". In: *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*. Springer, 2007, pp. 163–180.
[59] Virginia A Marchman et al. "Accuracy of the Language Environment Analyses (LENATM) system for estimating child and adult speech in laboratory settings". In: *Journal of child language* 48.3 (2021), pp. 605–620.
[60] Lillian R Masek et al. "Beyond counting words: A paradigm shift for the study of language acquisition". In: *Child Development Perspectives* 15.4 (2021), pp. 274–280.
[61] Elise F Masur and Jean B Gleason. "Parent–child interaction and the acquisition of lexical information during play." In: *Developmental Psychology* 16.5 (1980), p. 404.
[62] Mary L McHugh. "Interrater reliability: the kappa statistic". In: *Biochemia medica* 22.3 (2012), pp. 276–282.
[63] Cheryl Bodiford McNeil, Toni L Hembree-Kigin, and Karla Anhalt. "Parent-child interaction therapy". In: (2010).
[64] Brian McWhinney. "The CHILDES project: Tools for analyzing talk". In: *Mahwah, NJ* (2000).
[65] Ivonne Monarca et al. "Why doesn't the conversational agent understand me? a language analysis of children speech". In: *Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers*. 2020, pp. 90–93.
[66] Vivian Nguyen et al. "A systematic review and Bayesian meta-analysis of the development of turn taking in adult–child vocal interactions". In: *Child Development* 93.4 (2022), pp. 1181–1200.

[67] Behnam Nikbakhtbideh, Linda Duffett-Leger, and Mohammad Moshirpour. "Behavior analysis of parent-child interactions from text". In: *2023 International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2023, pp. 1175–1180.
[68] OpenAI. *GPT 3.5*. URL: https://platform.openai.com/docs/models/gpt-3-5-turbo.
[69] Tae Jin Park et al. "Multi-scale speaker diarization with dynamic scale weighting". In: *arXiv preprint arXiv:2203.15974* (2022).
[70] Pillar Learning. *Meet Codi- An Interactive, AI-Enabled Smart Toy for Kids!* https://www.pillarlearning.com/. Accessed: 2023-04-28. 2021.
[71] Laura Pina et al. "In situ cues for ADHD parenting strategies using mobile technology". In: *Proceedings of the 8th international conference on pervasive computing technologies for healthcare*. 2014, pp. 17–24.
[72] Alexis Plaquet and Hervé Bredin. "Powerset multi-class cross entropy loss for neural speaker diarization". In: *Proc. INTERSPEECH 2023*. 2023.
[73] Alec Radford et al. "Robust speech recognition via large-scale weak supervision". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 28492–28518.
[74] Juan Ramos et al. "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. 1. Citeseer. 2003, pp. 29–48.
[75] James Rehg et al. "Decoding children's social behavior". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 3414–3421.
[76] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: https://arxiv.org/abs/1908.10084.
[77] Dima Rekesh et al. "Fast conformer with linearly scalable attention for efficient speech recognition". In: *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 2023, pp. 1–8.
[78] Sirada Rochanavibhata and Viorica Marian. "Culture at play: A cross-cultural comparison of mother-child communication during toy play". In: *Language Learning and Development* 18.3 (2022), pp. 294–309.
[79] Clare R Rogers et al. "Causal effects on child language development: A review of studies in communication sciences and disorders". In: *Journal of communication disorders* 57 (2015), pp. 3–15.
[80] Deb Roy et al. "The human speechome project". In: *Symbol Grounding and Beyond: Third International Workshop on the Emergence and Evolution of Linguistic Communication, EELC 2006, Rome, Italy, September 30–October 1, 2006. Proceedings*. Springer. 2006, pp. 192–196.
[81] Martin Russell. "The pf-star british english childrens speech corpus". In: *The Speech Ark Limited* (2006).
[82] Martin Russell et al. "Applications of automatic speech recognition to speech and language development in young children". In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. Vol. 1. IEEE. 1996, pp. 176–179.
[83] Neville Ryant et al. *Second DIHARD Challenge Evaluation - SEEDLingS*. URL: https://doi.org/10.35111/mfam-hf33.
[84] Björn Schuller et al. "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language". In: *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*. Vol. 8. ISCA. 2016, pp. 2001–2005.
[85] Alex Sciuto et al. "" Hey Alexa, What's Up?" A Mixed-Methods Studies of In-Home Conversational Agent Usage". In: *Proceedings of the 2018 designing interactive systems conference*. 2018, pp. 857–868.
[86] Mostafa Shahin et al. "Tabby Talks: An automated tool for the assessment of childhood apraxia of speech". In: *Speech Communication* 70 (2015), pp. 49–64.
[87] Kaitao Song et al. "MPNet: Masked and Permuted Pre-training for Language Understanding". In: *arXiv preprint arXiv:2004.09297* (2020).

[88] Seokwoo Song et al. "TalkLIME: mobile system intervention to improve parent-child interaction for children with language delay". In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2016, pp. 304–315.

[89] Vaibhav Srivastav et al. *Open Automatic Speech Recognition Leaderboard*. https://huggingface.co/spaces/hf-audio/open_asr_leaderboard. 2023.

[90] Mai Stafford et al. "Parent–child relationships and offspring's positive mental wellbeing from adolescence to early older age". In: *The journal of positive psychology* 11.3 (2016), pp. 326–337.

[91] Catherine S Tamis-LeMonda et al. "Language and play in parent–child interactions". In: *Handbook of parenting* (2019), pp. 189–213.

[92] Chek Tien Tan et al. "sPeAK-MAN: towards popular gameplay for speech therapy". In: *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death*. 2013, pp. 1–4.

[93] Lukas Teufl and Lieselotte Ahnert. "Parent–child play and parent–child relationship: Are fathers special?" In: *Journal of Family Psychology* 36.3 (2022), p. 416.

[94] Toycloud. *Alpha Egg- An AI learning robot for children, that follows along and reads whatever you point at*. https://www.toycloud.com/channels/198.html. Accessed: 2023-04-28. n.d.

[95] Lori J Van Houten. "The Role of Maternal Input in the Acquisition Process: The Communicative Strategies of Adolescent and Older Mothers with the Language Learning Children." In: (1986).

[96] Mark VanDam et al. "HomeBank: An online repository of daylong child-centered audio recordings". In: *Seminars in speech and language*. Vol. 37. 02. Thieme Medical Publishers. 2016, pp. 128–142.

[97] Yuanyuan Wang et al. "A meta-analysis of the predictability of LENA™ automated measures for child language development". In: *Developmental Review* 57 (2020), p. 100921.

[98] Wayne Ward, Ron Cole, and Sameer Pradhan. "My science tutor and the myst corpus". In: *Boulder Learning Inc* (2019).

[99] Peter Washington et al. "SuperpowerGlass: a wearable aid for the at-home therapy of children with autism". In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1.3 (2017), pp. 1–22.

[100] Jason Wei et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[101] Susan M Williams, Peter G Fairweather, and Don Nix. "Speech recognition to support early literacy". In: *Interactive Literacy Education*. Routledge, 2023, pp. 95–116.

[102] Jiamin Xie et al. "Multi-PLDA Diarization on Children's Speech." In: *Interspeech*. 2019, pp. 376–380.

[103] Dongxin Xu et al. "Signal processing for young child speech language development". In: *First Workshop on Child, Computer and Interaction*. 2008.

[104] Wenjie Xu et al. "MathKingdom: Teaching Children Mathematical Language Through Speaking at Home via a Voice-Guided Game". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–14.

[105] Ying Xu et al. ""Rosita Reads With My Family": Developing A Bilingual Conversational Agent to Support Parent-Child Shared Reading". In: *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 2023, pp. 160–172.

[106] Ying Xu et al. "Same benefits, different communication patterns: Comparing Children's reading with a conversational agent vs. a human partner". In: *Computers & Education* 161 (2021), p. 104059.

[107] Shu-wen Yang et al. "Superb: Speech processing universal performance benchmark". In: *arXiv preprint arXiv:2105.01051* (2021).

[108] Chungkuk Yoo et al. "Mom, I see You Angry at Me! Designing a Mobile Service for Parent-child Conflicts by In-situ Emotional Empathy". In: *Proceedings of the 5th ACM Workshop on Mobile Systems for Computational Social Science*. 2019, pp. 21–26.

[109] Chao Zhang et al. “Mathemyths: Leveraging Large Language Models to Teach Mathematical Language through Child-AI Co-Creative Storytelling”. In: *arXiv preprint arXiv:2402.01927* (2024).
[110] Tianyi Zhang* et al. “BERTScore: Evaluating Text Generation with BERT”. In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.
[111] Zhao Zhao and Rhonda McEwen. “Luka Luka-investigating the interaction of children and their home Reading companion robot: A longitudinal remote study”. In: *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 2021, pp. 141–143.
[112] Shuhan Zhong et al. “DYPA: A Machine Learning Dyslexia Prescreening Mobile Application for Chinese Children”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7.3 (2023), pp. 1–21.

A Methodological Transparency & Reproducibility

The “fixed” prompt used in our evaluation of large language models on parent dialogues is reproduced below:

(PROMPT START)
Introduction: Parent-child interaction therapy (PCIT) helps parents improve the quality of interaction with children who have behavior problems. The therapy trains parents to use effective dialogue acts when interacting with their children. Besides weekly coaching by therapists, the therapy relies on deliberate practice of skills by parents in their homes.
Task: Specifically we want you (GPT4) to label parents’ act dialogues into 8 classes.
Data: samples were collected using a standard set of toys - Fisher Price Farm set and Doll House plus people and furniture - as the props for play-based conversations. For privacy the children’s names are coded as “childsname”.

Labels (numbers behind the classes’ name) with their definition and a few examples:
1: Negative Talk: verbal expression of disapproval of the child or the child’s attributes, activities, products, or choices. Also includes sassy, sarcastic, rude, or impudent speech.
Example 1: You’re being naughty.
Example 2: You put it in the wrong place.
Example 3: That’s not red.
Example 4: Will you stop whining?
Example 5: (sarcastically) That was smart.
Example 6: If you quit making that noise, I’ll play any game you want.
2: Command: statements that contain an order, direction, or suggestion for a vocal or motor behavior to be performed by the child. Can be direct, implied, or stated as a question.
Example 1: Susie, pick the crayon up off the floor.
Example 2: Try and set it down on its edge.
Example 3: Be careful.
Example 4: Why don’t we use the larger wheel in the front?
Example 5: You can color it purple.
Example 6: see.
Example 7: Let’s see.
Example 8: Look.
3: Labeled Praise: provides a positive evaluation of a specific behavior, activity, or product of the child.
Example 1: You did a great job of building the tower.
Example 2: I like the way you drew that circle.

Example 3: You sing so well.
Example 4: Thank you for handing me the box.
4: Unlabeled Praise: provides a positive evaluation of the child, an attribute of the child, or a nonspecific activity, behavior, or product of the child.
Example 1: Nice job.
Example 2: You're a good artist.
Example 3: All right!
Example 4: Thank you.
Example 5: There you go.
5: Question: verbal inquiries, often identifiable from a rising inflection at the end and/or by having the sentence structure of a question. Questions may or may not require an informative response. Questions asking the child to do something should be marked a Command, not Question.
Example 1: What did Santa bring you?
Example 2: Do you want the red or the black pieces?
Example 3: She is the princess?
Example 4: How about if I use the green crayon?
Example 5: Child: Is it high enough? Parent: Is it high enough?
Example 6: Child: There is a monster in the closet. Parent: There is a monster in the closet?
6: Reflective Statement: a statement by the parent that has the same meaning as the immediately preceding child verbalization. The reflection may paraphrase or elaborate upon the child's verbalization but may not change the meaning of the child's statement or interpret unstated ideas.
Example 1: Child: The toy box is full. Parent: The toy box is full.
Example 2: Child: My teacher is taking us to the zoo. Parent: Oh, you're going to the zoo.
Example 3: Child: That's a funny clown. Parent: You think he's funny.
Example 4: Child: It's a horsey. Parent: It is a horse. It's a brown horse.
7: Behavior Description: declarative sentences or phrases in which (i) the subject is the child and (ii) the verb describes the child's ongoing or immediately completed verbal or nonverbal behavior. This applies only to the first sentence after the child speaks.
Example 1: You're building a truck.
Example 2: You drew a rabbit and you gave it long ears.
Example 3: You just finished the red one.
8: Neutral Talk: statements that introduce information about people, objects, events, or activities, or indicate attention to the child, but do not clearly describe or evaluate the child's current or immediately completed behavior.
Example 1: I'm making my rainbow just like yours.
Example 2: I wonder if I left the iron on.
Example 3: Excuse me.
Example 4: Maybe.

Single Parent Sentence Case: Unlike some of the provided examples, I am going to only provide a single dialogue from the parent. I want you to first analyze it comprehensively and finally label it into one of those 8 classes. If the dialogue is only "." or something like that, then you can classify it as Neutral Talk and return 8. Return the label in a separate line (last line) and only return the number without any other character.

"Variable Window" and "Fixed Window" Cases: I will provide a conversation between Child (CHI) and Parents (MOT or FAT) and you should classify each parents' dialogue. You can consider child dialogues for understanding the

context better which is helpful for labeling, but your task is only labeling parents' dialogues. Sometimes a dialogue may only consist of parents' dialogues as well. I want you to first analyze the conversation comprehensively and finally label each of parents' dialogues into one of those 8 classes. If the dialogue is only "." or something like that, then you can classify it as Neutral Talk and return 8. Return the labels in a separate line (last line) and only return a list of labels with the same number of entries as the parent dialogues. So if a conversation with "m" dialogues consists of "n" parents' dialogues, your last line should be a list with n values each corresponds to the predicted label of each parents' dialogue (with the same order).
(PROMPT END)

The "fixed" prompt used in our evaluation of large language models on child dialogues is reproduced below:

(PROMPT START)
Introduction: The Dyadic Parent-Child Interaction Coding System (DPICS) is an observational measure for the quality of parent-child interactions as well as child prosocial and disruptive behaviors. DPICS child codes have been shown to be associated with parent-reported problem behavior.
Task: Specifically we want you to label children's act dialogues into 4 classes.
Data: samples were collected using a standard set of toys - Fisher Price Farm set and Doll House plus people and furniture - as the props for play-based conversations. For privacy the children's names are coded as "childsname".

Labels (numbers behind the classes' name) with their definition and a few examples:

1: Negative Talk: Negative talk is a verbal expression of disapproval of the parent or the parent's attributes, activities, products, or choices. Negative talk also includes sassy, sarcastic, rude, or impudent speech.
Example 1: MOT: Put your shoes on. CHI: No.
Example 2: CHI: Your picture is ugly.
Example 3: CHI: Yuck.
Example 4: MOT: The teacher said you started the fight. CHI: She's a liar.
Example 5: MOT: This is Barney. CHI: You're wrong.
Example 6: MOT: The pig's eating. CHI: That's not a pig.
2: Command: Commands are statements in which the child directs the vocal or motor behavior of the parent. Commands may be given directly, as an order, or implied, as a suggestion.
Example 1: Dad.
Example 2: Wait.
Example 3: see.
Example 4: Let's see.
Example 5: Look.
Example 6: Can you think of what should go here?
Example 7: Give me the purple crayon.
Example 8: Let's make the house bigger.
Example 9: We should tell dad what we made.
Example 10: We need to draw the grass here.
3: Question: Questions are verbal inquiries from one person to another that are distinguishable from declarative statements by having a rising inflection at the end and/or by having the sentence structure of a question. Questions request an answer but do not suggest that a behavior is to be performed by the other person.
Example 1: Huh?
Example 2: What do you wanna play?

Example 3: Where is the block?
Example 4: This is the runway?
Example 5: MOT: I'm drawing a dinosaur. CHI: Yeah?
Example 6: MOT: They have new crayons. CHI: Want to color?
4: Prosocial Talk: Prosocial talk incorporates several categories of verbalizations which contribute positively to the parent-child interaction. Prosocial Talk includes all statements that positively evaluate an attribute, product, or behavior of the parent (specifically or generally); describe the parent's behavior; provide neutral information; reflect the parent's verbalizations; or acknowledge the parent.
Example 1: Alright!
Example 2: I need one more pink stick.
Example 3: You are making a hat.
Example 4: MOT: The farmer's feeding his cows. CHI: Feeding his cows.
Example 5: Yea.
Example 6: The cow is eating the hay.
Example 7: Thanks.
Example 8: MOT: Thanks. CHI: You're welcome.
Example 9: MOT: A car. CHI: A car.
Example 10: MOT: Which one is prettier? CHI: I don't know.

Single Child Sentence Case: Unlike some of the provided examples, I am going to only provide a single dialogue from the Child. I want you to first analyze it comprehensively and finally label it into one of those 4 classes. If the dialogue is only "." or something like that, then you can classify it as Prosocial Talk and return 4. Return the label in a separate line (last line) and only return the number without any other character.

"Variable Window" and "Fixed Window" Cases: I will provide a conversation between Child (CHI) and Parents (MOT and FAT) and you should classify each child's dialogue. You can consider the parents dialogues for understanding the context better which is helpful for labeling, but your task is only labeling children's dialogues. Sometimes a dialogue may only consist of child's dialogues as well. I want you to first analyze the conversation comprehensively and finally label each of child's dialogues into one of those 4 classes. If the dialogue is only "." or something like that, then you can classify it as Prosocial Talk and return 4. Return the labels in a separate line (last line) and only return a list of labels with the same number of entries as the Child (CHI) dialogues. So if a conversation with "m" dialogues consists of "n" children's dialogues, your last line should be a list with n values each corresponds to the predicted label of each child's dialogue (with the same order).
(PROMPT END)

Figure 10 demostrates the single sentence, "variable window", and "fixed window" approaches used in evaluating GPT-4 with illustrative examples of parent DPICS classification.

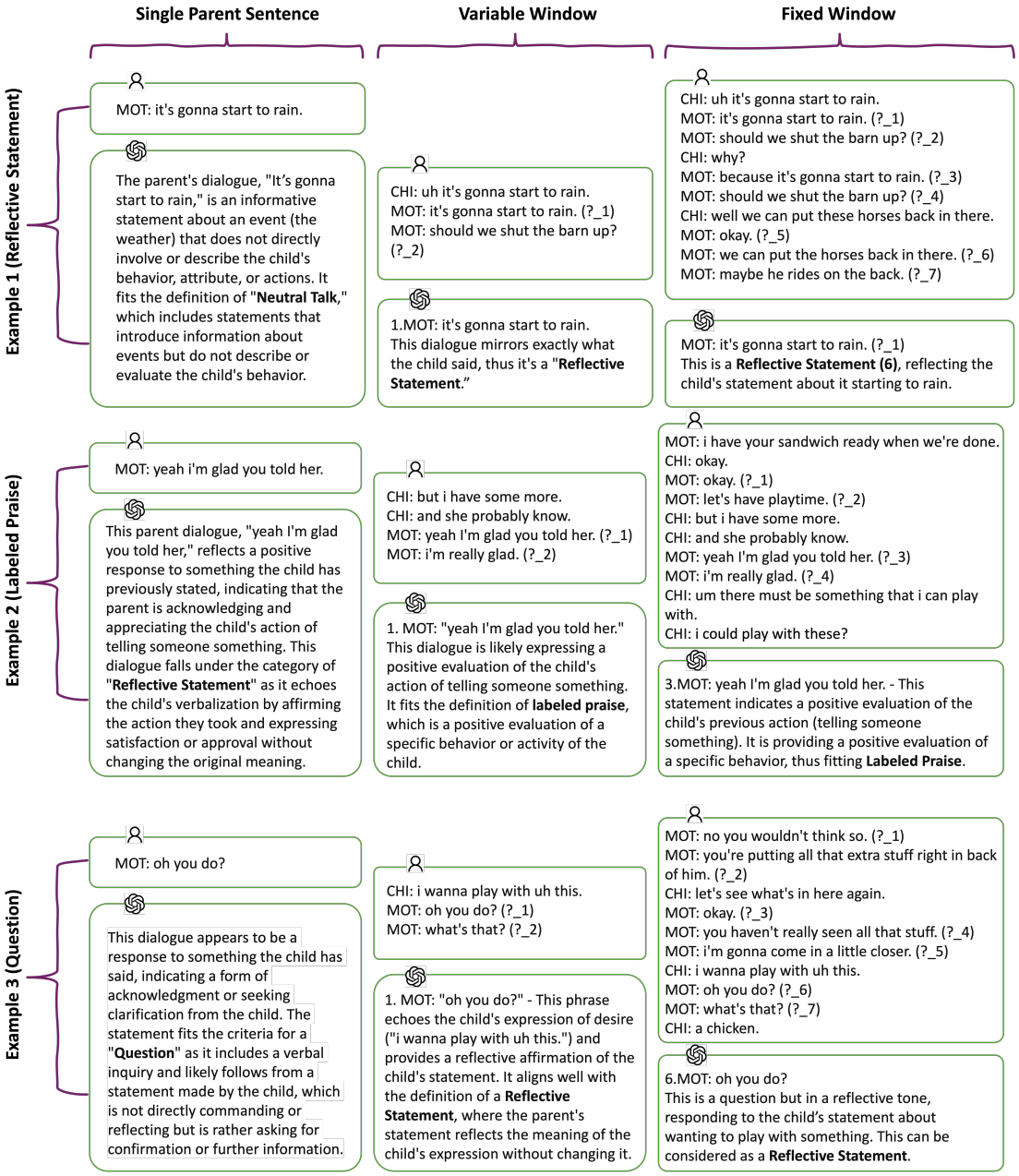

Fig. 10. Examples test prompts and GPT-4 outputs for single parent sentence, "variable window", and "fixed window" modes of evaluation (truncated due to space limitations). Ground-truth labels for examples 1, 2, and 3 are "Reflective Statement", "Labeled Praise", and "Question" respectively.