Assessing Spoken Language Understanding Pipeline of a Multimodal Dialogue System for Kids Learning Math at Home

Eda Okur [1] **Roddy Fuentes Alba** [2] **Saurav Sahay** [1] **Lama Nachman** [1]

Abstract

Enriching the quality of early childhood education with interactive math learning at home systems, empowered by recent advances in conversational AI technologies, is slowly becoming a reality. With this motivation, we implement a multimodal dialogue system to support play-based learning experiences at home, guiding kids to master basic math concepts. This work explores Spoken Language Understanding (SLU) pipeline within a task-oriented dialogue system developed for Kid Space, with cascading Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) components evaluated on our home deployment data with kids going through gamified math learning activities. We validate the advantages of a multi-task architecture for NLU and experiment with a diverse set of pretrained language representations for Intent Recognition and Entity Extraction in the math learning domain. To recognize kids' speech in realistic home environments, we investigate several ASR systems, including the Google Cloud and the latest open-source Whisper solutions with varying model sizes. We evaluate the SLU pipeline by testing our best-performing NLU models on noisy ASR output to inspect the challenges of understanding children for math learning in authentic homes.

1. Introduction and Background

The ongoing progress in Artificial Intelligence (AI) based advanced technologies can assist humanity in reducing the most critical inequities around the globe. The recent widespread interest in conversational AI applications presents exciting opportunities to showcase the positive societal impact of these technologies. The language-based AI systems have already started to mature to a level where we may soon observe their influences in mitigating the most pressing global challenges. Education is among the top priority improvement areas identified by the United Nations (UN) (i.e., poverty, hunger, healthcare, and education). In particular, increasing the inclusiveness and quality of education is within the UN development goals[1] with utmost urgency. One of the preeminent ways to diminish societal inequity is promoting STEM (i.e., Science, Technology, Engineering, Math) education, specifically ensuring that children succeed in mathematics. It is well-known that acquiring basic math skills at younger ages builds students up for success, regardless of their future career choices (Cesarone, 2008; Torpey, 2012). For math education, interactive learning environments through gamification present substantial leverages over more traditional learning settings for studying elementary math subjects, particularly with younger learners (Skene et al., 2022). With that goal, conversational AI technologies can facilitate this interactive learning environment where students can master fundamental math concepts. Despite these motivations, studying spoken language technologies for younger kids to learn basic math is a vastly uncharted area of AI.

This work[2] discusses a modular goal-oriented Spoken Dialogue System (SDS) specifically targeted for kids to learn and practice basic math concepts at home setup. Initially, a multimodal dialogue system (Sahay et al., 2019) is implemented for Kid Space (Anderson et al., 2018), a gamified math learning application for deployment in authentic classrooms. During this preliminary real-world deployment at an elementary school, the COVID-19 pandemic impacted the globe, and school closures forced students to switch to online learning options at home. To support this sudden paradigm shift to at-home learning, previous school use cases are redesigned for new home usages, and our dialogue system is recreated to deal with interactive math games at home. While the play-based learning activities are adjusted for home usages with a much simpler setup, the multimodal aspects of these games are partially preserved

[1]Intel Labs, USA [2]Intel Labs, Mexico. Correspondence to: Eda Okur <eda.okur@intel.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] https://sdgs.un.org/goals

[2] The previous version of this paper has been accepted to the *18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)* at ACL 2023.

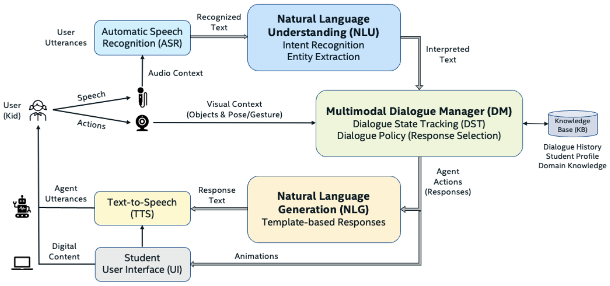

Figure 1. Multimodal Dialogue System Pipeline

along with the fundamental math concepts for early childhood education. These math skills cover using ones and tens to construct numbers and foundational arithmetic concepts and operations such as counting, addition, and subtraction. The multimodal aspects of these learning games include kids' spoken interactions with the system while answering math questions and carrying out game-related conversations, physical interactions with the objects (i.e., placing cubes and sticks as manipulatives) on a visually observed playmat, performing specific pose and gesture actions as part of these interactive games (e.g., jumping, standing, air high-five).

Our domain-specific SDS pipeline (see Figure 1) consists of multiple cascaded components, namely Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), Multimodal Dialogue Manager (DM), Natural Language Generation (NLG), and Text-to-Speech (TTS) synchronizing the agent utterances with virtual character animations on Student User Interface (UI). Here we concentrate on the Spoken Language Understanding (SLU) task on kids' speech at home environments while playing basic math games. Such application-dependent SLU approaches commonly involve two main modules applied sequentially: (i) Speech-to-Text (STT) or ASR module that recognizes speech and transcribes the spoken utterances into text, and (ii) NLU module that interprets the semantics of those utterances by processing the transcribed text. NLU is one of the most integral components of these goal-oriented dialogue systems. It empowers user-agent interactions by understanding the meaning of user utterances via performing domain-specific sub-tasks. Intent Recognition (IR) and Named Entity Recognition (NER) are essential sub-tasks within the NLU module to resolve the complexities of human language and extract meaningful information for the application at hand. Given a user utterance as input, the Intent Classification aims to identify the user's intention (i.e., what the user desires to achieve with that interaction) and categorize the user's objective at that conversational turn. The Entity Extraction targets locating and classifying entities (i.e., specific terms representing existing things such as person names, locations, and organizations) mentioned in user utterances into predefined task-specific categories.

In this study, we present our efforts to convert the task-oriented SDS (Okur et al., 2022c) designed for school use cases (Aslan et al., 2022) to home usages after COVID-19 and inspect the performance of individual SDS modules evaluated on the home deployment data we recently collected from 12 kids individually at their homes. Although the overall dialogue system handles multimodal interactions (e.g., spoken interactions with the agent, physical interactions with the objects, and pose-gesture interactions to perform required game-specific actions), this study narrows down on the kids' verbal interplay with the system. Specifically, the current work focuses on assessing and improving the SLU task performance on kids' utterances at home by utilizing this real-world deployment data. We first investigate the ASR and NLU module evaluations independently. Then, we inspect the overall SLU pipeline (i.e., ASR+NLU) performance on kids' speech by evaluating our NLU tasks on ASR output (i.e., recognized text) at home environments. As the erroneous and noisy speech recognition output would lead to incorrect intent and entity predictions, we aim to understand these error propagation consequences with SLU for children in the math learning domain. We experiment with various recent ASR solutions and diverse model sizes to gain more insights into their capabilities to recognize kids' speech at home. We then analyze the effects of these ASR engines on understanding intents and extracting entities from children's utterances. We discuss our findings and observations for potential enhancements in future deployments of this dialogue system for math learning at home.

2. Related Work

2.1. Conversational AI for Math Learning

With the ultimate goal of improving the quality of education, there has been a growing enthusiasm for exploiting AI-based intelligent systems to boost students' learning experiences (Chassignol et al., 2018; Aslan et al., 2019; Jia et al., 2020; Zhai et al., 2021; Baker, 2021). Among these, interactive frameworks that support guided play-based learning spaces revealed significant advantages for math learning (Pires et al., 2019; Sun et al., 2021; Richey et al., 2021), especially for building foundational math skills in early childhood education (Nrupatunga et al., 2021; Skene et al., 2022). To attain this level of interactivity within smart learning spaces, developing innovative educational applications by utilizing language-based AI technologies is in growing demand (Taghipour & Ng, 2016; Lende & Raghuwanshi, 2016; Raamadhurai et al., 2019; Cahill et al., 2020; Chan et al., 2021; Rathod et al., 2022). In particular, designing conversational agents for intelligent tutoring is a compelling yet challenging area of research, with several attempts presented so far (Winkler & Söllner, 2018; Wambsganss et al., 2020; Winkler et al., 2020; Datta et al., 2020; Okonkwo & Ade-Ibijola, 2021; Wollny et al., 2021), most of them focusing on language learning (Bibauw et al., 2022; Tyen et al., 2022; Zhang et al., 2022).

In the math education context, earlier conversational math tutoring applications exist, such as SKOPE-IT (Nye et al., 2018), which is based on AutoTutor (Graesser et al., 2005) and ALEKS (Falmagne et al., 2013), and MathBot (Grossman et al., 2019). These are often text-based online systems following strict rules in conversational graphs. Later, various studies emerged at the intersection of cutting-edge AI techniques and math learning (Mansouri et al., 2019; Huang et al., 2021; Azerbayev et al., 2022; Uesato et al., 2022; Yang et al., 2022). Among those, employing advanced language understanding methods to assist math learning is relatively new (Peng et al., 2021; Shen et al., 2021; Loginova & Benoit, 2022; Reusch et al., 2022). The majority of those recent work leans on exploring language representations for math-related tasks such as mathematical reasoning, formula understanding, math word problem-solving, knowledge tracing, and auto-grading, to name a few. Recently, TalkMoves dataset (Suresh et al., 2022a) was released with K-12 math lesson transcripts annotated for discursive moves and dialogue acts to classify teacher talk moves in math classrooms (Suresh et al., 2022b).

For the conversational AI tasks, the latest large language models (LLMs) based chatbots, such as BlenderBot (Shuster et al., 2022) and ChatGPT (OpenAI, 2022), gained a lot of traction in the education community (Tack & Piech, 2022; Kasneci et al., 2023), along with some concerns about using generative models in tutoring (Macina et al., 2023; Cotton et al., 2023). ChatGPT is a general-purpose open-ended interaction agent trained on internet-scale data. It is an end-to-end dialogue model without explicit NLU/Intent Recognizer or DM, which currently cannot fully comprehend the multimodal context and proactively generate responses to nudge children in a guided manner without distractions. Using these recent chatbots for math learning is still in the early stages because they are known to miss basic mathematical abilities and carry reasoning flaws (Frieder et al., 2023), revealing a lack of common sense. Moreover, they are known to be susceptible to triggering inappropriate or harmful responses and potentially perpetuate human biases since they are trained on internet-scale data and require carefully-thought guardrails.

On the contrary, our unique application is a task-oriented math learning spoken dialogue system designed to perform learning activities, following structured educational games to assist kids in practicing basic math concepts at home. Our SDS does not require massive amounts of data to understand kids and generate appropriate adaptive responses, and the lightweight models can run locally on client machines. In addition, our solution is multimodal, intermixing the physical and digital hybrid learning experience with audio-visual understanding, object recognition, segmentation, tracking, and pose and gesture recognition.

2.2. Spoken Language Understanding

Conventional pipeline-based dialogue systems with supervised learning are broadly favored when initial domain-specific training data is scarce to bootstrap the task-oriented SDS for future data collection (Serban et al., 2018; Budzianowski et al., 2018; Mehri et al., 2020). Deep learning-based modular dialogue frameworks and practical toolkits are prominent in academic and industrial settings (Bocklisch et al., 2017; Burtsev et al., 2018; Reyes et al., 2019). For task-specific applications with limited in-domain data, current SLU systems often use a cascade of two neural modules: (i) ASR maps the input audio to text (i.e., transcript), and (ii) NLU predicts intent and slots/entities from this transcript. Since our main focus in this work is investigating the SLU pipeline, we briefly summarize the existing NLU and ASR solutions.

2.2.1. Language Representations for NLU

The NLU component processes input text, often detects intents, and extracts referred entities from user utterances. For the mainstream NLU tasks of Intent Classification and Entity Recognition, jointly trained multi-task models are proposed (Liu & Lane, 2016; Zhang & Wang, 2016; Goo et al., 2018) with hierarchical learning approaches (Wen et al., 2018; Okur et al., 2019; Vanzo et al., 2019). Transformer architecture (Vaswani et al., 2017) is a game-changer for several downstream language tasks. With Transformers, BERT (Devlin et al., 2019) is presented, which became one of the most pivotal breakthroughs in language representations, achieving high performance in various tasks, including NLU. Later, Dual Intent and Entity Transformer (DIET) architecture (Bunk et al., 2020) is invented as a lightweight multi-task NLU model. On multi-domain NLU-Benchmark data (Liu et al., 2021a), the DIET model outperformed fine-tuning BERT for joint Intent and Entity Recognition.

For BERT-based autoencoding approaches, RoBERTa (Liu et al., 2019) is presented as a robustly optimized BERT model for sequence and token classification. The Hugging Face introduced a smaller, lighter general-purpose language representation model called DistilBERT (Sanh et al., 2019) as the knowledge-distilled version of BERT. ConveRT (Henderson et al., 2020) is proposed as an efficiently compact model to obtain pretrained sentence embeddings as conversational representations for dialogue-specific tasks. LaBSE (Feng et al., 2022) is a pretrained multilingual model producing language-agnostic BERT sentence embeddings that achieve promising results in text classification.

The GPT family of autoregressive LLMs, such as GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020), perform well at what they are pretrained for, i.e., text generation. GPT models can also be adapted for NLU, supporting few-shot learning capabilities, and NLG in task-oriented dialogue systems (Madotto et al., 2020; Liu et al., 2021b). XLNet (Yang et al., 2019) applies autoregressive pretraining for representation learning that adopts Transformer-XL (Dai et al., 2019) as a backbone model and works well for language tasks with lengthy contexts. DialoGPT (Zhang et al., 2020) extends GPT-2 as a large response generation model for multi-turn conversations trained on Reddit discussions, whose representations can be exploited in dialogue tasks.

For language representations to be utilized in math-related tasks, MathBERT (Shen et al., 2021) is introduced as a math-specific BERT model pretrained on large math corpora. Later, Math-aware-BERT and Math-aware-RoBERTa models (Reusch et al., 2022) are proposed based on BERT and RoBERTa, pretrained on Math Stack Exchange[3].

2.2.2. Speech Recognition with Kids

Speech recognition technology has been around for some time, and numerous ASR solutions are available today, both commercial and open-source. Rockhopper ASR (Stemmer et al., 2017) is an earlier low-power speech recognition engine with LSTM-based language models, where its acoustic models are trained using an open-source Kaldi speech recognition toolkit (Povey et al., 2011). Google Cloud Speech-to-Text[4] is a prominent commercial ASR service powered by advanced neural models and designed for speech-dependant applications. Until recently, Google STT API was arguably the leader in ASR services for recognition performance and language coverage. Franck Dernoncourt (2018) reported that Google ASR could reach a word error rate (WER) of 12.1% on LibriSpeech clean dataset (28.8% on LibriSpeech other) (Panayotov et al., 2015) at that time, which is improved drastically over time. Recently, Open AI released Whisper ASR (Radford et al., 2022) as a game-changer speech recognizer. Whisper models are pretrained on a vast amount of labeled audio-transcription data (i.e., 680k hours), unlike its predecessors (e.g., Wav2Vec 2.0 (Baevski et al., 2020) is trained on 60k hours of unlabeled audio). 117k hours of this data are multilingual, which makes Whisper applicable to over 96 languages, including low-resourced ones. Whisper architecture follows a standard Transformer-based encoder-decoder as many speech-related models (Latif et al., 2023). The Whisper-base model is reported to achieve 5.0% & 12.4% WER on LibriSpeech clean & other datasets.

Although speech recognition systems are substantially improving to achieve human recognition levels, problems still occur, especially in noisy environments, with users having accents and dialects or underrepresented groups like kids. Child speech brings distinct challenges to ASR (Stemmer et al., 2003; Gerosa et al., 2007; Yeung & Alwan, 2018), such as data scarcity and highly varied acoustic, linguistic, physiological, developmental, and articulatory characteristics compared to adult speech (Claus et al., 2013; Shivakumar & Georgiou, 2020; Bhardwaj et al., 2022). Thus, WER for children's voices is reported two-to-five times worse than for adults (Wu et al., 2019), as the younger the child, the poorer ASR performs. There exist efforts to mitigate these difficulties of speech recognition with kids (Shivakumar et al., 2014; Duan & Chen, 2020; Booth et al., 2020; Kelly et al., 2020; Rumberg et al., 2021; Yeung et al., 2021). Few studies also focus on speech technologies in educational settings (Reeder et al., 2015; Blanchard et al., 2015; Bai et al., 2021; 2022; Dutta et al., 2022), often for language acquisition, reading comprehension, and story-telling activities.

3. Methods

3.1. Home Learning Data and Use Cases

We utilize two datasets for gamified basic math learning at home usages. The first set is a proof-of-concept (POC) data manually constructed based on User Experience (UX) studies (e.g., detailed scripts for new home use cases) and partially adopted from our previous school data (Okur et al., 2022a). This POC data is used to train and cross-validate various NLU models to develop the best practices in later home deployments. The second set is our recent home deployment data collected from 12 kids (ages 7-8) experiencing our multimodal math learning system at authentic homes. The audio-visual data is transcribed manually, and user utterances in these reference transcripts are annotated for intent and entity types we identified for each learning activity at home. Table 1 compares the NLU statistics for Kid Space Home POC and Deployment datasets. Manually transcribed children's utterances in deployment data are employed to test our best NLU models trained on POC data. We run multiple ASR engines on audio recordings from home deployment data, where automatic transcripts (i.e., ASR output) are utilized to compute WER to assess ASR model performances on kids' speech. We also evaluate the SLU pipeline (ASR+NLU) by testing NLU models on ASR output from deployment data.

The simplified home deployment setup includes a playmat with physical manipulatives, a laptop with a built-in camera, a wireless lavalier mic, and a depth camera on a tripod. Home use cases follow a particular flow of activities designed for play-based learning in early childhood education. These activities are Introduction (Meet & Greet), Warm-up Game (Red Light Green Light), Training Game, Learning

[3] https://math.stackexchange.com/

[4] https://cloud.google.com/speech-to-text/

Table 1. Kid Space Home POC and Deployment Data

NLU Data Statistics	POC	Deployment
# Intents Types	13	12
Total # Utterances	4091	733
# Entity Types	3	3
Total # Entities	2244	497
Min # Utterances per Intent	105	1
Max # Utterances per Intent	830	270
Avg # Utterances per Intent	314.7	61.1
Min # Tokens per Utterance	1	1
Max # Tokens per Utterance	40	33
Avg # Tokens per Utterance	4.49	2.30
# Unique Tokens (Vocab Size)	702	149
Total # Tokens	18364	1689

Game, and Closure (Dance Party). Note that these are not fixed static games but rather dynamic interactions, such that the agent's responses and future actions to proceed with the games depend on the kid's utterances and multimodal inputs. After meeting with the virtual character and playing jumping games, the child starts the training game, where the agent asks for help planting flowers. The agent presents tangible manipulatives, cubes representing ones and sticks representing tens, and instructs the kid to answer basic math questions and construct numbers using these objects, going through multiple rounds of practice questions where flowers in child-selected colors bloom as rewards. In the actual learning game, the agent presents clusters of questions involving ones & tens, and the child provides verbal (e.g., stating the numbers) and visual answers (e.g., placing the cubes and sticks on the playmat, detected by the overhead camera). The agent provides scaffolding utterances and performs animations to show and tell how to solve basic math questions. The interaction ends with a dance party to celebrate achievements and say goodbyes in closure. Some of our intents can be considered generic (e.g., *state-name*, *affirm*, *deny*, *repeat*, *out-of-scope*), but some are highly domain-specific (e.g., *answer-flowers*, *answer-valid*, *answer-others*, *state-color*, *had-fun-a-lot*, *end-game*) or math-related (e.g., *state-number*, *still-counting*). The entities we extract are activity-specific (i.e., *name*, *color*) and math-related (i.e., *number*).

3.2. NLU and ASR Models

Customizing open-source Rasa framework (Bocklisch et al., 2017) as a backbone, we investigate several NLU models for Intent Recognition and Entity Extraction tasks to implement our math learning conversational AI system for home usage. Our baseline approach is inspired by the StarSpace (Wu et al., 2018) method, a supervised embedding-based model maximizing the similarity between utterances and intents in shared vector space. We enrich this simple text classifier by incorporating SpaCy (Honnibal et al., 2020) pretrained language models[5] for word embeddings as additional features in the NLU pipeline. CRF Entity Extractor (Lafferty et al., 2001) with BILOU tagging is also part of this baseline NLU. For home usages, we explore the advantages of switching to a more recent DIET model[6] for joint Intent and Entity Recognition, a multi-task architecture with two-layer Transformers shared for NLU tasks. DIET leverages combining dense features (e.g., any given pretrained embeddings) with sparse features (e.g., token-level encodings of char n-grams). To observe the net benefits of DIET, we first pass the identical SpaCy embeddings used in our baseline (StarSpace) as dense features to DIET. Then, we adopt DIET with pretrained BERT[7], RoBERTa[8], and DistilBERT[9] word embeddings, as well as ConveRT[10] and LaBSE[11] sentence embeddings to inspect the effects of these autoencoding-based language representations on NLU performance (see 2.2.1 for more details). We also evaluate pretrained embeddings from models using autoregressive training such as XLNet[12], GPT-2[13][14], and DialoGPT[15] on top of DIET. Next, we explore recently-proposed math-language representations pretrained on math data for our basic math learning dialogue system. MathBERT (Shen et al., 2021) is pretrained on large math corpora (e.g., curriculum, textbooks, MOOCs, arXiv papers) covering pre-k to college-graduate materials. We enhance DIET by incorporating embeddings from MathBERT-base[16] and MathBERT-custom[17] models, pretrained with BERT-base original and math-customized vocabularies, respectively. Math-aware-BERT[18] and Math-aware-RoBERTa[19] models (Reusch et al., 2022) are initialized from BERT-base and RoBERTa-base, and further pretrained on Math Stack-

[5] https://github.com/explosion/spacy-models/releases/tag/en_core_web_md-3.5.0

[6] Please check Bunk et al. (2020) for hyper-parameter tuning, hardware specs, and computational costs.

[7] https://huggingface.co/bert-base-uncased

[8] https://huggingface.co/roberta-base

[9] https://huggingface.co/distilbert-base-uncased

[10] https://github.com/connorbrinton/polyai-models/releases

[11] https://huggingface.co/rasa/LaBSE

[12] https://huggingface.co/xlnet-base-cased

[13] https://huggingface.co/gpt2

[14] Excluded GPT-3 and beyond that are not open-source.

[15] https://huggingface.co/microsoft/DialoGPT-medium

[16] https://huggingface.co/tbs17/MathBERT

[17] https://huggingface.co/tbs17/MathBERT-custom

[18] https://huggingface.co/AnReu/math_pretrained_bert

[19] https://huggingface.co/AnReu/math_pretrained_roberta

Table 2. NLU Model Selection Results in F1-scores (%) Evaluated on Kid Space Home POC Data (10-fold CV)

NLU Model	Intent Detection	Entity Extraction
StarSpace+SpaCy	92.71±0.25	97.08±0.21
DIET+SpaCy	94.29±0.05	98.38±0.12
DIET+BERT	97.25±0.23	99.23±0.02
DIET+RoBERTa	95.50±0.18	99.11±0.12
DIET+DistilBERT	97.41±0.20	99.49±0.12
DIET+ConveRT	**98.80±0.25**	99.61±0.03
DIET+LaBSE	98.19±0.18	**99.72±0.04**
DIET+XLNet	94.99±0.19	98.38±0.14
DIET+GPT-2	95.35±0.27	99.01±0.27
DIET+DialoGPT	96.00±0.49	98.94±0.12
DIET+MathBERT-base	94.55±0.22	98.10±0.21
DIET+MathBERT-custom	94.61±0.34	97.48±0.29
DIET+Math-aware-BERT	95.95±0.15	98.94±0.19
DIET+Math-aware-RoBERTa	94.20±0.16	98.75±0.21

Exchange[20] with extra LaTeX tokens to better tokenize math formulas for ARQMath-3 tasks (Mansouri et al., 2022). We exploit these representations with DIET to investigate their effects on our NLU tasks in the basic math domain.

For the ASR module, we explore three main speech recognizers for our math learning application at home, which are explained further in 2.2.2. Rockhopper ASR[21] is the baseline local approach previously inspected, which can be adjusted slightly for kids. Its acoustic models rely on Kaldi[22] generated resources and are trained on default adult speech data. In the past explorations, when Rockhopper's language models fine-tuned with limited in-domain kids' utterances (Sahay et al., 2021) from previous school usages, WER decreased by 40% for kids but remained 50% higher than adult WER. Although this small-scale baseline solution is unexpected to reach Google Cloud ASR performance, Rockhopper has a few other advantages for our application since it can run offline locally on low-power devices, which could be better for security, privacy, latency, and cost (relative to cloud-based ASR services). Google ASR is a commercial cloud solution providing high-quality speech recognition service but requiring connectivity and payment, which cannot be adapted or fine-tuned as Rockhopper. The third ASR approach we investigate is Whisper[23], which combines the best of both worlds as it is an open-source adjustable solution that can run locally, achieving new state-of-the-art (SOTA) results. We inspect three configurations of varying model sizes (i.e., base, small, and medium) to evaluate the Whisper ASR for our home math learning usage with kids.

Table 3. NLU Evaluation Results in F1-scores (%) for DIET+ConveRT Models Trained on Kid Space Home POC Data & Tested on Home Deployment Data

	Intent Detection			Entity Extraction		
Activity	POC	Deploy	Δ	POC	Deploy	Δ
Intro (Meet & Greet)	99.9	97.3	-2.6	99.2	97.4	-1.8
Warm-up Game	98.8	93.4	-5.4	-	-	-
Training Game	98.4	94.2	-4.2	99.9	99.8	-0.1
Learning Game	98.9	94.3	-4.6	99.8	99.4	-0.4
Closure (Dance)	98.8	98.7	-0.1	-	-	-
All Activities	**98.8**	**94.2**	**-4.6**	**99.6**	**99.3**	**-0.3**

4. Experimental Results

To build the NLU module of our SLU pipeline, we train Intent and Entity Classification models and cross-validate them over the Kid Space Home POC dataset to decide upon the best-performing NLU architectures moving forward for home. Table 2 summarizes the results of model selection experiments with various NLU models. We report the average of 5 runs, and each run involves a 10-fold cross-validation (CV) on POC data. Compared to the baseline StarSpace algorithm, we gain almost 2% F1 score for intents and more than 1% F1 for entities with multi-task DIET architecture. For language representations, we observe that incorporating DIET with the BERT family of embeddings from autoencoders achieves higher F1 scores relative to the GPT family of embeddings from autoregressive models. We cannot reveal any benefits of employing math-specific representations with DIET, as all such models achieve worse than DIET+BERT results. One reason we identify is the mismatch between our early math domain and advanced math corpora, including college-level math symbols and equations, that these models trained on. Another reason could be that such embeddings are pretrained on smaller math corpora (e.g., 100 million tokens) compared to massive-scale generic corpora (e.g., 3.3 billion words) that BERT models use for training. DIET+ConveRT is the clear winner for intents and achieves second-best but very close results for entities compared to DIET+LaBSE. ConveRT and LaBSE are both sentence-level embeddings, but ConveRT performs well on dialogue tasks as it is pretrained on large conversational corpora, including Reddit discussions. Based on these results, we select DIET+ConveRT as the final multi-task architecture for our NLU tasks at home.

Next, we evaluate our NLU module on Kid Space Home Deployment data collected at authentic homes over 12 sessions with 12 kids. Each child goes through 5 activities within a session, as described in 3.1. In Table 3, we observe overall F1% drops (Δ) of 4.6 for intents and 0.3 for entities when our best-performing DIET+ConveRT models are tested on home deployment data. These findings are expected and relatively lower than the performance drops we

[20] https://archive.org/download/stackexchange

[21] https://docs.openvino.ai/2018_R5/_samples_speech_sample_README.html

[22] https://github.com/kaldi-asr/kaldi

[23] https://github.com/openai/whisper

Table 4. ASR Model Results: Avg Word Error Rates (WER) for Child Speech at Kid Space Home Deployment Data

ASR Model	**Raw Output**	**Lowercase (LC)**	**Remove Punct (RP)**	**Num2Word (NW)**	**LC & RP**	**LC & RP & NW**	**NW & Clean**	**LC & RP & NW & Clean**
Rockhopper	0.939	0.919	0.924	0.937	0.886	0.884	0.937	0.884
Google Cloud	0.829	0.798	0.775	0.763	0.695	0.602	0.763	0.602
Whisper-base	1.042	1.020	0.971	0.985	0.946	0.856	0.622	**0.500**
Whisper-small	0.834	0.804	0.760	0.756	0.720	0.621	0.537	**0.405**
Whisper-medium	0.905	0.870	0.824	0.814	0.785	0.675	0.522	**0.384**

previously observed at school (Okur et al., 2022b). We witness distributional and utterance-length differences between POC/training and deployment/test datasets. Real-world data would always be noisier than anticipated as these utterances come from younger kids playing math games in dynamic conditions.

To further improve the performance of our Kid Space Home NLU models (trained on POC data) by leveraging this recent deployment data, we experiment with merging the two datasets for training and evaluating the performance on individual deployment sessions via leave-one-out (LOO) CV. At each of the 12 runs (for 12 sessions/kids), we merge the POC data with 11 sessions of deployment data for model training and use the remaining session as a test set, then take the average performance of these runs. That would simulate how combining POC with real-world deployment data would help us train more robust NLU models that perform better on unseen data in future deployment sessions. The overall F1-scores reach 96.5% for intents (2.3% gain from 94.2%) and 99.4% for entities (0.1% gain) with LOOCV, which are promising for our future deployments.

To inspect the ASR module of our SLU pipeline, we experiment with Rockhopper, Google, and Whisper-base/small/medium ASR models evaluated on the same audio data collected during home deployments. Using the manual session transcripts as a reference, we compute the average WER for kids with each ASR engine to investigate the most feasible solution. Table 4 summarizes WER results before and after standard pre-processing steps (e.g., lower casing and punctuation removal) as well as application-specific filters (e.g., num2word and cleaning). The numbers are transcribed inconsistently within reference transcripts plus ASR output (e.g., 35 vs. thirty-five), and we need to standardize them all in word forms. The cleaning step is applied to Whisper ASR output only due to known issues such as getting stuck in repeat loops and hallucinations (Radford et al., 2022). We seldom observe trash output from Whisper (4-to-7%) having very long transcriptions with non-sense repetitions/symbols, which hugely affect WER due to their length, yet these samples can be easily auto-filtered. Even after these steps, the relatively high error rates can be attributed to many factors related to the characteristics of these recordings (e.g., incidental voice and phrases), very short utterances to be recognized (e.g., binary yes/no answers or stating numbers with one-or-two words), and recognizing kids' speech in ordinary home environments. Still, the comparative results indicate that Whisper ASR solutions perform better on kids' utterances, and the WER can benefit from increasing the model size from base to small (while the error rates with small vs. medium-sized models are close).

Table 5. SLU Pipeline Evaluation Results in F1-scores (%) for ASR+NLU and VAD-Adjusted ASR+NLU on Kid Space Home Deployment Data

	Intent Detection		**Entity Extraction**	
ASR Model	F1	Adjusted-F1	F1	Adjusted-F1
Rockhopper	36.7	15.5	82.9	35.0
Google Cloud	**78.0**	39.7	96.2	49.0
Whisper-base	64.7	60.0	95.4	88.5
Whisper-small	72.2	68.1	96.6	91.1
Whisper-medium	**76.5**	**73.1**	**98.5**	**94.1**

For SLU pipeline evaluation, we test our highest-performing NLU models on noisy ASR output. Table 5 presents the Intent and Entity Classification results achieved on home deployment data where the DIET+ConveRT models run on varying ASR models output. Note that Voice Activity Detection (VAD) is an integral part of ASR that decides the presence/absence of human speech. We realize that the VAD stage is filtering out a lot of audio chunks with actual kid speech with Rockhopper and Google. Thus, our VAD-ASR nodes can ignore a lot of audio segments with reference transcripts (57.9% for Rokchopper, 49.1% for Google). That is less of an issue with Whisper-base/small/medium, missing 7.1%/5.7%/4.4% of transcribed utterances (often due to filtering very long and repetitive trash Whisper output). When we treat these entirely missed utterances with no ASR output as classification errors for NLU tasks (i.e., missing to predict intent/entities when no speech is detected), we can adjust the F1-scores accordingly to evaluate the VAD-ASR+NLU pipeline. These VAD-adjusted F1-scores are compared in Table 5, aligned with the WER results, where NLU on Whisper ASR performs relatively higher than Google and Rockhopper. For enhanced Intent Recognition in real-world deployments with kids, increasing the

Table 6. NLU Error Analysis: Intent Recognition Error Samples from Kid Space Home Deployment Data

Sample Kid Utterance	Intent	Prediction
Pepper.	*state-name*	*answer-valid*
Wow, that's a lot of red flowers.	*out-of-scope*	*answer-flowers*
None.	*state-number*	*deny*
Nothing.	*state-number*	*deny*
Yeah. Can we have some carrots?	*affirm*	*out-of-scope*
Okay, Do your magic.	*affirm*	*out-of-scope*
Maybe tomorrow.	*affirm*	*out-of-scope*
He's a bear.	*out-of-scope*	*answer-valid*
I like the idea of a bear	*out-of-scope*	*answer-valid*
Oh, 46? Okay.	*still-counting*	*state-number*
94. Okay.	*still-counting*	*state-number*
Now we have mountains.	*out-of-scope*	*answer-valid*
A pond?	*out-of-scope*	*answer-valid*
Sorry, I didn't understand it. Uh, five tens.	*state-number*	*still-counting*
Ah this is 70, 7.	*state-number*	*still-counting*

Table 7. SLU Pipeline (ASR+NLU): Intent Recognition Error Samples from Kid Space Home Deployment Data

Human Transcript	ASR Output	ASR Model	Intent	Prediction
Six.	thanks	Rockhopper	*state-number*	*thank*
fifteen	if he	Rockhopper	*state-number*	*out-of-scope*
fifteen	Mickey	Google Cloud	*state-number*	*state-name*
Five.	bye	Google Cloud	*state-number*	*goodbye*
Blue.	Blair.	Whisper-base	*state-color*	*state-name*
twenty	Plenty.	Whisper-base	*state-number*	*had-fun-a-lot*
A lot.	Oh, la.	Whisper-base	*had-fun-a-lot*	*out-of-scope*
A lot.	Oh, wow.	Whisper-small	*had-fun-a-lot*	*out-of-scope*
Two.	you	Whisper-small	*state-number*	*out-of-scope*
Four.	I'm going to see this floor.	Whisper-small	*state-number*	*out-of-scope*
twenty	Swamy?	Whisper-medium	*state-number*	*state-name*
Eight.	E.	Whisper-medium	*state-number*	*out-of-scope*

ASR model size from small to medium could be worth the trouble for Whisper. Yet, the F1 drop is still huge, from 94.2% with NLU to 73.1% with VAD-ASR+NLU, when VAD-ASR errors propagate into the SLU pipeline.

5. Error Analysis

For NLU error analysis, Table 6 reveals utterance samples from our Kid Space Home Deployment data with misclassified intents obtained by the DIET+ConveRT models on manual/human transcripts. These language understanding errors illustrate the potential pain points solely related to the NLU model performances, as we are assuming perfect or human-level ASR here by feeding the manually transcribed utterances into the NLU. Such intent prediction errors occur in real-world deployments for many reasons. For example, authentic user utterances can have multiple intents (e.g., "*Yeah. Can we have some carrots?*" starts with *affirm* and continues with *out-of-scope*). Some utterances can be challenging due to subtle differences between intent classes (e.g., "*Ah this is 70, 7.*" is submitting a verbal answer with *state-number* but can easily be mixed with *still-counting* too). Moreover, we observe utterances having *colors* and "*flowers*" within *out-of-scope* (e.g., "*Wow, that's a lot of red flowers.*"), which can be confusing for the NLU models trained on relatively cleaner POC datasets.

For further error analysis on the SLU pipeline (ASR+NLU), Table 7 demonstrates Intent Recognition error samples from Kid Space Home Deployment data obtained on ASR output with several speech recognition models we explored. These samples depict anticipated error propagation from speech recognition to language understanding modules in the cascaded SLU approach. Please check Appendix A for a more detailed ASR error analysis.

6. Conclusion

To increase the quality of math learning experiences at home for early childhood education, we develop a multimodal dialogue system with play-based learning activities, helping the kids gain basic math skills. This study investigates a modular SLU pipeline for kids with cascading ASR and NLU modules, evaluated on our first home deployment data with 12 kids at individual homes. For NLU, we examine the advantages of a multi-task architecture and experiment with numerous pretrained language representations for Intent Recognition and Entity Extraction tasks in our application domain. For ASR, we inspect the WER with several solutions that are either low-power and local (e.g., Rockhopper), commercial (e.g., Google Cloud), or open-source (e.g., Whisper) with varying model sizes and conclude that Whisper-medium outperforms the rest on kids' speech at authentic home environments. Finally, we evaluate the SLU pipeline by running our best-performing NLU models, DIET+ConveRT, on VAD-ASR output to observe the significant effects of cascaded errors due to noisy voice detection and speech recognition performance with kids in realistic home deployment settings. In the future, we aim to fine-tune the Whisper ASR acoustic models on kids' speech and language models on domain-specific math content. Moreover, we consider exploring N-Best-ASR-Transformers (Ganesan et al., 2021) to leverage multiple Whisper ASR hypotheses and mitigate errors propagated into cascading SLU.

Limitations

By building this task-specific dialogue system for kids, we aim to increase the overall quality of basic math education and learning at-home experiences for younger children. In our previous school deployments, the overall cost of the whole school/classroom setup, including the wall/ceiling-mounted projector, 3D/RGB-D cameras, LiDAR sensor, wireless lavalier microphones, servers, etc., can be considered as a limitation for public schools and disadvantaged populations. When we shifted our focus to home learning usages after the COVID-19 pandemic, we simplified the overall setup for 1:1 learning with a PC laptop with a built-in camera, a depth camera on a tripod, a lapel mic, and a playmat with cubes and sticks. However, even this minimal instrumentation suitable for home setup can be a limitation for kids with lower socioeconomic status. Moreover, the dataset size of our initial home deployment data collected from 12 kids in 12 sessions is relatively small, with around 12 hours of audio data manually transcribed and annotated. Collecting multimodal data at authentic homes of individual kids within our target age group (e.g., 5-to-8 years old) and labor-intensive labeling process is challenging and costly. To overcome these data scarcity limitations and develop dialogue systems for kids with such small-data regimes, we had to rely on transfer learning approaches as much as possible. However, the dataset sizes affect the generalizability of our explorations, the reliability of some results, and ultimately the robustness of our multimodal dialogue system for deployments with kids in the real world. We aim to collect more deployment data (both at school and home) to try to mitigate the known data scarcity issues and strengthen our investigation results to build a more robust system. Please note that although our dialogue system and data are constructed for English-language, it can be adapted easily to other languages by exploiting the available multilingual resources for NLU (e.g., pretrained non-English language representations) and ASR (e.g., Whisper supports both English-only and multilingual ASR).

Ethics Statement

Prior to our initial research deployments at home, a meticulous process of Privacy Impact Assessment is pursued. The legal approval processes are completed to operate our research with educators, parents, and the kids. Individual participants and parties involved have signed the relevant consent forms in advance, which inform essential details about our research studies. The intentions and procedures and how the participant data will be collected and utilized to facilitate our research are explained in writing in these required consent forms. Our collaborators comply with stricter data privacy policies as well. For further discussion on ethical implications of this work, please check Appendix B.

Acknowledgements

We aspire to share our gratitude and acknowledge our former and current colleagues in the Kid Space team at Intel Labs. Particularly: (i) Hector Cordourier Maruri, Juan Del Hoyo Ontiveros, and Georg Stemmer for developing the VAD-ASR node to obtain the ASR output that we use in our SLU pipeline; (ii) Benjamin Bair, Lenitra Durham, Sai Prasad, Giuseppe Raffa, Celal Savur, and Sangita Sharma for designing the HW/SW architectural setup, developing the Wizard UI and game logic nodes, and supporting data collection; (iii) Ankur Agrawal, Arturo Bringas Garcia, Vishwajeet Narwal, and Guillermo Rivas Aguilar for developing the Student UI via the Unity game engine; (iv) Glen Anderson, Sinem Aslan, Rebecca Chierichetti, Pete Denman, John Sherry, and Meng Shi for conducting UX studies and performing interaction design to conceptualize school and home usages; (v) David Gonzalez Aguirre, Gesem Gudino Mejia, and Julio Zamora Esquivel for developing the visual understanding nodes to support this research. We would also want to gratefully acknowledge our field team members from Summa Linguae Technologies, especially Rick Lin and Brenda Tumbalobos Cubas, for their exceptional support in executing the data collection and transcription/annotation tasks in collaboration with our Intel Labs Kid Space team. Finally, we should thank the Rasa team and community developers for their open-source framework and contributions that empowered us to conduct our research.

References

Anderson, G. J., Panneer, S., Shi, M., Marshall, C. S., Agrawal, A., Chierichetti, R., Raffa, G., Sherry, J., Loi, D., and Durham, L. M. Kid space: Interactive learning in a smart environment. In *Proceedings of the Group Interaction Frontiers in Technology*, GIFT'18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360777. doi: 10.1145/3279981.3279986. URL https://doi.org/10.1145/3279981.3279986.

Aslan, S., Alyuz, N., Tanriover, C., Mete, S. E., Okur, E., D'Mello, S. K., and Arslan Esme, A. Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300534. URL https://doi.org/10.1145/3290605.3300534.

Aslan, S., Agrawal, A., Alyuz, N., Chierichetti, R., Durham, L. M., Manuvinakurike, R., Okur, E., Sahay, S., Sharma, S., Sherry, J., Raffa, G., and Nachman, L. Exploring kid space in the wild: a preliminary study of multimodal and immersive collaborative play-based learning experiences. *Educational Technology Research and Development*, 70:205–230, 2022. doi: 10.1007/s11423-021-10072-x. URL https://doi.org/10.1007/s11423-021-10072-x.

Azerbayev, Z., Piotrowski, B., and Avigad, J. Proofnet: A benchmark for autoformalizing and formally proving undergraduate-level mathematics problems. In *Workshop MATH-AI: Toward Human-Level Mathematical Reasoning, 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, Louisiana, USA*, 2022. URL https://mathai2022.github.io/papers/20.pdf.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Bai, Y., Hubers, F., Cucchiarini, C., and Strik, H. An ASR-based Reading Tutor for Practicing Reading Skills in the First Grade: Improving Performance through Threshold Adjustment. In *Proc. IberSPEECH 2021*, pp. 11–15, 2021. doi: 10.21437/IberSPEECH.2021-3. URL http://dx.doi.org/10.21437/IberSPEECH.2021-3.

Bai, Y., Hubers, F., Cucchiarini, C., van Hout, R., and Strik, H. The Effects of Implicit and Explicit Feedback in an ASR-based Reading Tutor for Dutch First-graders. In *Proc. Interspeech 2022*, pp. 4476–4480, 2022. doi: 10.21437/Interspeech.2022-10810.

Baker, R. S. Artificial intelligence in education: Bringing it all together. *OECD Digital Education Outlook 2021: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots*, pp. 43–51, 2021.

Bhardwaj, V., Ben Othman, M. T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., Rehman, A. U., Shafiq, M., and Hamam, H. Automatic speech recognition (asr) systems for children: A systematic literature review. *Applied Sciences*, 12(9):4419, 2022.

Bibauw, S., Van den Noortgate, W., François, T., and Desmet, P. Dialogue systems for language learning: a meta-analysis. *Language Learning & Technology*, 26(1), 2022.

Blanchard, N., Brady, M., Olney, A. M., Glaus, M., Sun, X., Nystrand, M., Samei, B., Kelly, S., and D'Mello, S. A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In Conati, C., Heffernan, N., Mitrovic, A., and Verdejo, M. F. (eds.), *Artificial Intelligence in Education*, pp. 23–33, Cham, 2015. Springer International Publishing. ISBN 978-3-319-19773-9.

Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. Rasa: Open source language understanding and dialogue management. In *Conversational AI Workshop, NIPS 2017*, 2017. URL http://arxiv.org/abs/1712.05181.

Booth, E., Carns, J., Kennington, C., and Rafla, N. Evaluating and improving child-directed automatic speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6340–6345, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.778.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547.

Bunk, T., Varshneya, D., Vlasov, V., and Nichol, A. DIET: lightweight language understanding for dialogue systems. *CoRR*, abs/2004.09936, 2020. URL https://arxiv.org/abs/2004.09936.

Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Bushkov, N., Gureenkova, O., Khakhulin, T., Kuratov, Y., Kuznetsov, D., Litinsky, A., Logacheva, V., Lymar, A., Malykh, V., Petrov, M., Polulyakh, V., Pugachev, L., Sorokin, A., Vikhreva, M., and Zaynutdinov, M. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pp. 122–127, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-4021. URL https://aclanthology.org/P18-4021.

Cahill, A., Fife, J. H., Riordan, B., Vajpayee, A., and Galochkin, D. Context-based automated scoring of complex mathematical responses. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 186–192, Seattle, WA, USA → Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.bea-1.19. URL https://aclanthology.org/2020.bea-1.19.

Cesarone, B. Early childhood mathematics: Promoting good beginnings. *Childhood Education*, 84(3):189, 2008.

Chan, Y., Chung, H., and Fan, Y. Improving controllability of educational question generation by keyword provision. *CoRR*, abs/2112.01012, 2021. URL https://arxiv.org/abs/2112.01012.

Chassignol, M., Khoroshavin, A., Klimova, A., and Bilyatdinova, A. Artificial intelligence trends in education: a narrative overview. *Procedia Computer Science*, 136:16–24, 2018. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2018.08.233. URL https://www.sciencedirect.com/science/article/pii/S1877050918315382. 7th International Young Scientists Conference on Computational Science, YSC2018, 02-06 July2018, Heraklion, Greece.

Claus, F., Rosales, H. G., Petrick, R., Hain, H.-U., and Hoffmann, R. A survey about databases of children's speech. In *INTERSPEECH*, pp. 2410–2414, 2013.

Cotton, D. R., Cotton, P. A., and Shipway, J. R. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in Education and Teaching International*, pp. 1–12, 2023.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285.

Datta, D., Phillips, M., Chiu, J. L., Watson, G. S., Bywater, J. P., Barnes, L. E., and Brown, D. E. Improving classification through weak supervision in context-specific conversational agent development for teacher education. *CoRR*, abs/2010.12710, 2020. URL https://arxiv.org/abs/2010.12710.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Duan, R. and Chen, N. F. Unsupervised feature adaptation using adversarial multi-task training for automatic evaluation of children's speech. In *INTERSPEECH*, pp. 3037–3041, 2020.

Dutta, S., Irvin, D., Buzhardt, J., and Hansen, J. H. Activity focused speech recognition of preschool children in early childhood classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pp. 92–100, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bea-1.13. URL https://aclanthology.org/2022.bea-1.13.

Falmagne, J.-C., Albert, D., Doble, C., Eppstein, D., and Hu, X. *Knowledge spaces: Applications in education*. Springer Science & Business Media, 2013.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62.

Franck Dernoncourt, Trung Bui, W. C. A framework for speech recognition benchmarking. In *Interspeech*, 2018.

Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., and Berner, J. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023.

Ganesan, K., Bamdev, P., B, J., Venugopal, A., and Tushar, A. N-best ASR transformer: Enhancing SLU performance using multiple ASR hypotheses. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 93–98, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.14. URL https://aclanthology.org/2021.acl-short.14.

Gerosa, M., Giuliani, D., and Brugnara, F. Acoustic variability and automatic recognition of children's speech. *Speech Communication*, 49(10-11):847–860, 2007.

Goo, C.-W., Gao, G., Hsu, Y.-K., Huo, C.-L., Chen, T.-C., Hsu, K.-W., and Chen, Y.-N. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 753–757, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2118. URL https://aclanthology.org/N18-2118.

Graesser, A., Chipman, P., Haynes, B., and Olney, A. Autotutor: an intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612–618, 2005. doi: 10.1109/TE.2005.856149.

Grossman, J., Lin, Z., Sheng, H., Wei, J. T.-Z., Williams, J. J., and Goel, S. Mathbot: Transforming online resources for learning math into conversational interactions. *AAAI 2019 Story-Enabled Intelligence*, 2019.

Henderson, M., Casanueva, I., Mrkšić, N., Su, P.-H., Wen, T.-H., and Vulić, I. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2161–2174, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.196. URL https://aclanthology.org/2020.findings-emnlp.196.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. spaCy: Industrial-strength natural language processing in python, 2020. URL https://github.com/explosion/spaCy.

Huang, S., Wang, J., Xu, J., Cao, D., and Yang, M. Real2: An end-to-end memory-augmented solver for math word problems. In *Workshop on Math AI for Education (MATHAI4ED), 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021. URL https://mathai4ed.github.io/papers/papers/paper_7.pdf.

Jia, J., He, Y., and Le, H. A multimodal human-computer interaction system and its application in smart learning environments. In Cheung, S. K. S., Li, R., Phusavat, K., Paoprasert, N., and Kwok, L. (eds.), *Blended Learning. Education in a Smart Learning Environment*, pp. 3–14, Cham, 2020. Springer International Publishing.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

Kelly, A. C., Karamichali, E., Saeb, A., Vesely̌, K., Parslow, N., Deng, A., Letondor, A., O'Regan, R., and Zhou, Q. Soapbox labs verification platform for child speech. In *INTERSPEECH*, pp. 486–487, 2020.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, ICML, pp. 282–289, 2001.

Latif, S., Zaidi, A., Cuayahuitl, H., Shamshad, F., Shoukat, M., and Qadir, J. Transformers in speech processing: A survey. *arXiv preprint arXiv:2303.11607*, 2023.

Lende, S. P. and Raghuwanshi, M. Question answering system on education acts using nlp techniques. In *2016 world conference on futuristic trends in research and innovation for social welfare (Startup Conclave)*, pp. 1–6. IEEE, 2016.

Liu, B. and Lane, I. Attention-based recurrent neural network models for joint intent detection and slot filling. In *Interspeech 2016*, pp. 685–689, 2016. doi: 10.21437/Interspeech.2016-1352. URL http://dx.doi.org/10.21437/Interspeech.2016-1352.

Liu, X., Eshghi, A., Swietojanski, P., and Rieser, V. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pp. 165–183. Springer, 2021a.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., and Tang, J. Gpt understands, too. *arXiv preprint arXiv:2103.10385*, 2021b.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Loginova, E. and Benoit, D. Structural information in mathematical formulas for exercise difficulty prediction: a comparison of nlp representations. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pp. 101–106, 2022.

Macina, J., Daheim, N., Wang, L., Sinha, T., Kapur, M., Gurevych, I., and Sachan, M. Opportunities and challenges in neural dialog tutoring. *arXiv preprint arXiv:2301.09919*, 2023.

Madotto, A., Liu, Z., Lin, Z., and Fung, P. Language models as few-shot learner for task-oriented dialogue systems. *arXiv preprint arXiv:2008.06239*, 2020.

Mansouri, B., Rohatgi, S., Oard, D. W., Wu, J., Giles, C. L., and Zanibbi, R. Tangent-cft: An embedding model for mathematical formulas. In *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*, pp. 11–18, 2019.

Mansouri, B., Novotný, V., Agarwal, A., Oard, D. W., and Zanibbi, R. Overview of arqmath-3 (2022): Third clef lab on answer retrieval for questions on math. In Barrón-Cedeño, A., Da San Martino, G., Degli Esposti, M., Sebastiani, F., Macdonald, C., Pasi, G., Hanbury, A., Potthast, M., Faggioli, G., and Ferro, N. (eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pp. 286–310, Cham, 2022. Springer International Publishing. ISBN 978-3-031-13643-6.

Mehri, S., Eric, M., and Hakkani-Tür, D. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *CoRR*, abs/2009.13570, 2020. URL https://arxiv.org/abs/2009.13570.

Nrupatunga, Kumar, A., and Rajagopal, A. Phygital math learning with handwriting for kids. In *Workshop on Math AI for Education (MATHAI4ED), 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021. URL https://mathai4ed.github.io/papers/papers/paper_5.pdf.

Nye, B. D., Pavlik, P. I., Windsor, A., Olney, A. M., Hajeer, M., and Hu, X. Skope-it (shareable knowledge objects as portable intelligent tutors): overlaying natural language tutoring on an adaptive learning system for mathematics. *International journal of STEM education*, 5:1–20, 2018.

Okonkwo, C. W. and Ade-Ibijola, A. Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 2:100033, 2021.

Okur, E., Kumar, S. H., Sahay, S., Arslan Esme, A., and Nachman, L. Natural language interactions in autonomous vehicles: Intent detection and slot filling from passenger utterances. In Gelbukh, A. (ed.), *Computational Linguistics and Intelligent Text Processing*, pp. 334–350, Cham, 2019. Springer Nature Switzerland. ISBN 978-3-031-24340-0. doi: 10.1007/978-3-031-24340-0_25. URL https://doi.org/10.1007/978-3-031-24340-0_25.

Okur, E., Sahay, S., Fuentes Alba, R., and Nachman, L. End-to-end evaluation of a spoken dialogue system for learning basic mathematics. In *Proceedings of the 1st Workshop on Mathematical Natural Language Processing (MathNLP)*, pp. 51–64, Abu Dhabi, United Arab Emirates (Hybrid), December 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.mathnlp-1.7.

Okur, E., Sahay, S., and Nachman, L. NLU for game-based learning in real: Initial evaluations. In *Proceedings of the 9th Workshop on Games and Natural Language Processing within the 13th Language Resources and Evaluation Conference*, pp. 28–39, Marseille, France, June 2022b. European Language Resources Association. URL https://aclanthology.org/2022.games-1.4.

Okur, E., Sahay, S., and Nachman, L. Data augmentation with paraphrase generation and entity extraction for multimodal dialogue system. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4114–4125, Marseille, France, June 2022c. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.437.

OpenAI. Chatgpt: Optimizing language models for dialogue, 2022.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Peng, S., Yuan, K., Gao, L., and Tang, Z. Mathbert: A pre-trained model for mathematical formula understanding. *CoRR*, abs/2105.00377, 2021. URL https://arxiv.org/abs/2105.00377.

Pires, A. C., González Perilli, F., Bakała, E., Fleisher, B., Sansone, G., and Marichal, S. Building blocks of mathematical learning: Virtual and tangible manipulatives lead to different strategies in number composition. *Frontiers in Education*, 4, 2019. ISSN 2504-284X. doi: 10.3389/feduc.2019.00081. URL https://www.frontiersin.org/article/10.3389/feduc.2019.00081.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

Raamadhurai, S., Baker, R., and Poduval, V. Curio SmartChat : A system for natural language question answering for self-paced k-12 learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 336–342, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4435. URL https://aclanthology.org/W19-4435.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

Rathod, M., Tu, T., and Stasaski, K. Educational multi-question generation for reading comprehension. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pp. 216–223, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bea-1.26. URL https://aclanthology.org/2022.bea-1.26.

Reeder, K., Shapiro, J., Wakefield, J., and D'Silva, R. Speech recognition software contributes to reading development for young learners of english. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, 5(3):60–74, 2015.

Reusch, A., Thiele, M., and Lehner, W. Transformer-encoder and decoder models for questions on math. *Proceedings of the Working Notes of CLEF 2022*, pp. 5–8, 2022.

Reyes, R., Garza, D., Garrido, L., De la Cueva, V., and Ramirez, J. Methodology for the implementation of virtual assistants for education using google dialogflow. In *Advances in Soft Computing: 18th Mexican International Conference on Artificial Intelligence, MICAI 2019, Xalapa, Mexico, October 27–November 2, 2019, Proceedings 18*, pp. 440–451. Springer, 2019.

Richey, J. E., Zhang, J., Das, R., Andres-Bray, J. M., Scruggs, R., Mogessie, M., Baker, R. S., and McLaren, B. M. Gaming and confrustion explain learning advantages for a math digital learning game. In Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., and Dimitrova, V. (eds.), *Artificial Intelligence in Education*, pp. 342–355, Cham, 2021. Springer International Publishing. ISBN 978-3-030-78292-4.

Rumberg, L., Ehlert, H., Lüdtke, U., and Ostermann, J. Age-invariant training for end-to-end child speech recognition using adversarial multi-task learning. In *Interspeech*, pp. 3850–3854, 2021.

Sahay, S., Kumar, S. H., Okur, E., Syed, H., and Nachman, L. Modeling intent, dialog policies and response adaptation for goal-oriented interactions. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, London, United Kingdom, September 2019. SEMDIAL. URL http://semdial.org/anthology/Z19-Sahay_semdial_0019.pdf.

Sahay, S., Okur, E., Hakim, N., and Nachman, L. Semi-supervised interactive intent labeling. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pp. 31–40, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.dash-1.5. URL https://aclanthology.org/2021.dash-1.5.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *5th EMC2 Workshop - Energy Efficient Training and Inference of Transformer Based Models, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019. URL http://arxiv.org/abs/1910.01108.

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9(1):1–49, 2018.

Shen, J. T., Yamashita, M., Prihar, E., Heffernan, N. T., Wu, X., and Lee, D. Mathbert: A pre-trained language model for general NLP tasks in mathematics education. *CoRR*, abs/2106.07340, 2021. URL https://arxiv.org/abs/2106.07340.

Shivakumar, P. G. and Georgiou, P. Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations. *Computer speech & language*, 63:101077, 2020.

Shivakumar, P. G., Potamianos, A., Lee, S., and Narayanan, S. Improving speech recognition for children using acoustic adaptation and pronunciation modeling. In *Fourth Workshop on Child Computer Interaction (WOCCI 2014)*, 2014. URL https://www.isca-speech.org/archive_v0/wocci_2014/papers/wc14_015.pdf.

Shuster, K., Xu, J., Komeili, M., Ju, D., Smith, E. M., Roller, S., Ung, M., Chen, M., Arora, K., Lane, J., et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*, 2022.

Skene, K., O'Farrelly, C. M., Byrne, E. M., Kirby, N., Stevens, E. C., and Ramchandani, P. G. Can guidance during play enhance children's learning and development in educational contexts? a systematic review and meta-analysis. *Child Development*, 2022.

Stemmer, G., Hacker, C., Steidl, S., and Nöth, E. Acoustic normalization of children's speech. In *Eighth European Conference on Speech Communication and Technology*, 2003.

Stemmer, G., Georges, M., Hofer, J., Rozen, P., Bauer, J. G., Nowicki, J., Bocklet, T., Colett, H. R., Falik, O., Deisher, M., et al. Speech recognition and understanding on hardware-accelerated dsp. In *Interspeech*, pp. 2036–2037, 2017.

Sun, Y., Play, T., Nambiar, R., and Vidyasagaran, V. Gamifying math education using object detection. In *Workshop on Math AI for Education (MATHAI4ED), 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021. URL https://mathai4ed.github.io/papers/papers/paper_11.pdf.

Suresh, A., Jacobs, J., Harty, C., Perkoff, M., Martin, J. H., and Sumner, T. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4654–4662, Marseille, France, June 2022a. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.497.

Suresh, A., Jacobs, J., Perkoff, M., Martin, J. H., and Sumner, T. Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pp. 71–81, Seattle, Washington, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.bea-1.11. URL https://aclanthology.org/2022.bea-1.11.

Tack, A. and Piech, C. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. In *Proceedings of the 15th International Conference on Educational Data Mining*, pp. 522, 2022.

Taghipour, K. and Ng, H. T. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1193. URL https://aclanthology.org/D16-1193.

Torpey, E. Math at work: Using numbers on the job. *Occupational Outlook Quarterly*, 56(3):2–13, 2012.

Tyen, G., Brenchley, M., Caines, A., and Buttery, P. Towards an open-domain chatbot for language practice. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pp. 234–249, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bea-1.28. URL https://aclanthology.org/2022.bea-1.28.

Uesato, J., Kushman, N., Kumar, R., Song, H. F., Siegel, N. Y., Wang, L., Creswell, A., Irving, G., and Higgins, I. Solving math word problems with process-based and outcome-based feedback. In *Workshop MATH-AI: Toward Human-Level Mathematical Reasoning, 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, Louisiana, USA*, 2022. URL https://mathai2022.github.io/papers/26.pdf.

Vanzo, A., Bastianelli, E., and Lemon, O. Hierarchical multi-task natural language understanding for cross-domain conversational AI: HERMIT NLU. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 254–263, Stockholm, Sweden, September 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5931. URL https://aclanthology.org/W19-5931.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Wambsganss, T., Winkler, R., Söllner, M., and Leimeister, J. M. A conversational agent to improve response quality in course evaluations. In *Extended Abstracts of the 2020 CHI conference on human factors in computing systems*, pp. 1–9, 2020.

Wen, L., Wang, X., Dong, Z., and Chen, H. Jointly modeling intent identification and slot filling with contextual and hierarchical information. In Huang, X., Jiang, J., Zhao, D., Feng, Y., and Hong, Y. (eds.), *Natural Language Processing and Chinese Computing*, pp. 3–15, Cham, 2018. Springer International Publishing.

Winkler, R. and Söllner, M. Unleashing the potential of chatbots in education: A state-of-the-art analysis. In *Academy of Management Annual Meeting (AOM)*, 2018. URL http://www.alexandria.unisg.ch/254848/.

Winkler, R., Hobert, S., Salovaara, A., Söllner, M., and Leimeister, J. M. Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–14, 2020.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., and Drachsler, H. Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, 4:654924, 2021.

Wu, F., García-Perera, L. P., Povey, D., and Khudanpur, S. Advances in automatic speech recognition for child speech using factored time delay neural network. In *Interspeech*, pp. 1–5, 2019.

Wu, L., Fisch, A., Chopra, S., Adams, K., and Weston, A. B. J. Starspace: Embed all the things! In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Yang, Z., Qin, J., Chen, J., Lin, L., and Liang, X. Logic-Solver: Towards interpretable math word problem solving with logical prompt-enhanced learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1–13, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.1.

Yeung, G. and Alwan, A. On the difficulties of automatic speech recognition for kindergarten-aged children. *Interspeech 2018*, 2018.

Yeung, G., Fan, R., and Alwan, A. Fundamental frequency feature normalization and data augmentation for child speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6993–6997. IEEE, 2021.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., and Li, Y. A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity*, 2021, 2021.

Zhang, X. and Wang, H. A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pp. 2993–2999. AAAI Press, 2016. ISBN 978-1-57735-770-4. URL http://dl.acm.org/citation.cfm?id=3060832.3061040.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 270–278, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.30. URL https://aclanthology.org/2020.acl-demos.30.

Zhang, Z., Xu, Y., Wang, Y., Yao, B., Ritchie, D., Wu, T., Yu, M., Wang, D., and Li, T. J.-J. Storybuddy: A human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2022.

A. Appendix: Additional Error Analysis

Please refer to Table 8 for additional error analysis on ASR output from our home deployment data. Here, we compare manually transcribed utterances (i.e., human transcripts) with the speech recognition output (i.e., raw ASR transcripts) using five different ASR models that we investigated in this study. These ASR errors demonstrate the challenges faced in the speech recognition model performances on kids' speech, which potentially would be propagated into the remaining modules in the conventional task-oriented dialogue pipeline.

We may attribute various factors to these speech recognition errors, often related to our deployment data characteristics. Incidental voices and phrases constitute a good chunk of the overall home deployment data, along with very short utterances to be recognized (e.g., stating names, colors, types of flowers, numbers, and binary answers with one-or-two words), plus the remaining known challenges present with recognizing kids' speech in noisy real-world environments.

Table 8. ASR Error Samples from Kid Space Home Deployment Data

Human Transcript	**Rockhopper**	**Google Cloud**	**Whisper-base**	**Whisper-small**	**Whisper-medium**
Atticus.	-	-	Yeah, that's cute.	I have a kiss.	Now I have to kiss.
I am Genevieve.	i'm twenty-two	I'm going to be	I'm Kennedy.	I'm Genevieve.	I'm Genevieve.
Red.	rab	-	Ralph.	Red.	Red.
Blue.	lil	blue	Blair.	Blue.	Blue.
Yes,	laughs	yes	Yes?	Yes?	Yes?
Roses.	it is	roses	Okay.	Okay	focus
Zero.	you know	no	No.	No, no.	No.
four.	you swore	-	forward.	Over.	Over.
five.	-	bye	Bye.	Bye.	Bye.
eight	all	-	Thank you.	Bye.	Oh
forty eight	wall e	48	48	48	48
forty nine	already	49	49	49	49
fifty one	if you want	51	51	51	51
seventy four	stopping before	74	74	74	74
Maybe tomorrow.	novarro	tomorrow	I need some water, though.	I'm going to leave it tomorrow.	I'm leaving tomorrow.
Flowers, flowers in the greenhouse?	lean forward phelps hours than we	Greenhouse	In forward, in forward, in the green house.	I think forward, both flowers and the greenhouse.	In the green house.
There are seventeen, and seventeen minus ten equals seven.	seventeen seventeen rooms	17 + 17 - 27	There are 17 and 17 minus 10 equals 7.	There are 17 and 17 minus 10 equals 7.	What is the maximum number of children in the world? Um... There are 17 and 17 minus 10 equals 7.

B. Appendix: Further Ethical Considerations

As we briefly discussed in the Ethics Statement section (see page 9), all the required legal approval procedures are completed, and the signed consent forms are collected from the participants that provided their data for research purposes during our home deployment sessions. The multimodal data we collected in these sessions include the video streams from the built-in laptop and depth cameras, audio streams from built-in and lapel mics, all relevant system and interaction logs with the users, plus the UX research data such as interviews with the parents and children prior/after the sessions. To address privacy concerns due to the sensitive nature of this data involving kids, we comply with rigorous data privacy and security policies to prevent any attacks or information leakage.

Our application area, education, is also highly critical to be preserved from any uncertainties and forms of biases. To increase our control over the generated agent responses to kids, currently, we are exploiting template-based or canned responses at the NLG module of our SDS pipeline. When the multimodal DM module predicts the verbal response types in the form of agent actions, the NLG retrieves these pre-defined agent response templates. Creating variety in the final response text has been ensured by preparing multiple templates for each response type, usually with 3-to-6 variations. Among these variations in response templates, a final response text is picked randomly at run-time. Note that each response text is carefully designed by the UX experts and vetted by educators for age and grade appropriateness in advance. Employing this version of the template-based response approach makes the overall system more reliable and consistent, which is crucial for our application domain. These pre-defined templates would also serve as a guardrail to prevent harmful or inappropriate responses to children and mitigate potential bias issues.