Spoken Language Understanding Evaluations for Home-Based Basic Math Learning

Eda Okur
Intel Labs, USA
eda.okur@intel.com

Saurav Sahay
Intel Labs, USA
saurav.sahay@intel.com

Lama Nachman
Intel Labs, USA
lama.nachman@intel.com

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 the Spoken Language Understanding (SLU) pipeline within a task-oriented dialogue system, with cascading Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) components evaluated on our Kid Space home deployment data with children 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 recent 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's speech for math learning in authentic homes.

1 Introduction and Background

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[1] discusses a modular goal-oriented Spoken Dialogue System (SDS) specifically targeted for kids to learn and practice basic math concepts at home setup Okur et al. [2023a]. Initially, a multimodal dialogue system Sahay et al. [2019] is implemented for Kid Space Anderson et al. [2018] to be deployed in authentic classrooms. During this real-world deployment at an elementary school Aslan et al. [2022], the COVID-19 pandemic impacted the globe, and school closures forced students to switch to online learning at home. To support this sudden shift, previous school use cases are redesigned for new home usages Aslan et al. [2023], and our dialogue system is recreated to deal with interactive math games at home. While the play-based learning activities are adjusted

[1]The previous version Okur et al. [2023b] of this work has been accepted to the *Neural Conversational AI Workshop - What's left to TEACH (Trustworthy, Enhanced, Adaptable, Capable and Human-centric) chatbots?* at ICML 2023.

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Workshop on MATH-AI.

for home with a much simpler setup, the multimodal aspects of these games are partially preserved 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 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 (e.g., jumping, standing, air high-five).

Our domain-specific SDS pipeline Okur et al. [2022c] 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) 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. 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.

This study leans on assessing and improving the SLU task performance on kids' utterances at home by utilizing 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 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.

2 Home Learning Datasets and Use Cases

For our gamified basic math learning usages, we utilize two datasets. The first set is proof-of-concept (POC) data manually constructed based on UX studies 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. 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 are annotated for intent and entity types we identified for each learning activity. Table 1 compares the NLU statistics for Kid Space Home POC and Deployment datasets. Manually transcribed children's utterances in the deployment set

Table 1: Kid Space Home POC and Deployment Dataset Statistics

NLU Data Statistics	**POC**	**Deployment**
# Intents Types	13	12
Total # Utterances	6,245	733
# Entity Types	3	3
Total # Entities	5,023	497
Min # Utterances per Intent	105	1
Max # Utterances per Intent	1,051	270
Avg # Utterances per Intent	480.4	61.1
Min # Tokens per Utterance	1	1
Max # Tokens per Utterance	40	33
Avg # Tokens per Utterance	4.96	2.88
# Unique Tokens (Vocab Size)	1,992	362
Total # Tokens	30,976	2,113

are used to test our best NLU models trained on POC data. We run multiple ASR engines on audio recordings, where automatic transcripts (i.e., ASR output) are utilized to compute the word error rates (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.

Our home deployment setup includes a playmat with physical manipulatives, a laptop with a built-in camera, a lapel mic, and a depth camera on a tripod Aslan et al. [2023]. Home use cases follow a particular flow of activities such as Introduction (Meet & Greet), Warm-up Game (Red Light Green Light), Training Game, Learning Game, and Closure (Dance Party). 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 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). The agent outputs scaffolding utterances and performs animations to show and tell how to solve basic math questions, and the interaction ends with a dance party. 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 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. Our baseline approach is inspired by the StarSpace Wu et al. [2018] model. We enrich this text classifier by incorporating SpaCy Honnibal et al. [2020] pretrained word embeddings in the NLU pipeline. CRF Entity Extractor Lafferty et al. [2001] is also part of this baseline NLU. We explore the advantages of switching to a more recent DIET model Bunk et al. [2020] for joint Intent and Entity Recognition, a multi-task architecture with two-layer Transformers shared for NLU tasks. DIET combines dense features (e.g., pretrained embeddings) with sparse features (e.g., token-level encodings of n-grams). We first feed the SpaCy embeddings used in our baseline (StarSpace) as dense features to DIET. Then, we adopt DIET with pretrained BERT Devlin et al. [2019], RoBERTa Liu et al. [2019], and DistilBERT Sanh et al. [2019] word embeddings, as well as ConveRT Henderson et al. [2020] and LaBSE Feng et al. [2022] sentence embeddings to inspect the effects of these autoencoding-based language representations. We also evaluate pretrained embeddings from models using autoregressive training such as XLNet Yang et al. [2019], GPT-2 Radford et al. [2019], and DialoGPT Zhang et al. [2020] (excluded GPT-3 and beyond that are not open-source). Next, we explore recent math-language representations trained on math datasets for our basic math learning dialogue system. MathBERT Shen et al. [2021] is pretrained on large math corpora covering pre-k to college-graduate materials. We enhance DIET by incorporating embeddings from MathBERT-base and MathBERT-custom models, pretrained with BERT-base original and math-customized vocabularies, respectively. Math-aware-BERT and Math-aware-RoBERTa models Reusch et al. [2022] are initialized from BERT-base and RoBERTa-base, and further pretrained on Math StackExchange with improved LaTeX tokens for math formulas in ARQMath-3 tasks Mansouri et al. [2022].

For the ASR module, we explore three main speech recognizers. Rockhopper ASR Stemmer et al. [2017] is the baseline local approach, whose acoustic models rely on Kaldi Povey et al. [2011] resources trained on adult speech data. In the past, when Rockhopper's language models fine-tuned with limited in-domain kids' utterances from previous school usages, WER decreased by 40% for kids but remained 50% higher than adult WER. Rockhopper can run offline locally on low-power devices, which is better for security, privacy, latency, and cost. Google ASR is a commercial cloud solution providing high-quality speech recognition service but requiring connectivity and payment, which cannot be fine-tuned as Rockhopper. Whisper ASR Radford et al. [2022] is an open-source adjustable solution that can run locally, achieving new state-of-the-art (SOTA) results. We inspect four model sizes (i.e., tiny, base, small, and medium) to evaluate the Whisper ASR for our math learning usages with kids.

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

NLU Model	Intent Detection	Entity Extraction
StarSpace+SpaCy	92.83±0.28	97.14±0.21
DIET+SpaCy	94.40±0.08	98.45±0.11
DIET+BERT	97.37±0.26	99.29±0.01
DIET+RoBERTa	95.62±0.21	99.17±0.11
DIET+DistilBERT	97.52±0.23	99.54±0.11
DIET+ConveRT	**98.92±0.28**	99.66±0.02
DIET+LaBSE	98.31±0.21	**99.78±0.03**
DIET+XLNet	95.11±0.22	98.44±0.13
DIET+GPT-2	95.46±0.30	99.07±0.27
DIET+DialoGPT	96.12±0.52	99.00±0.11
DIET+MathBERT-base	94.67±0.25	98.15±0.20
DIET+MathBERT-custom	94.73±0.37	97.54±0.28
DIET+Math-aware-BERT	96.07±0.18	99.00±0.18
DIET+Math-aware-RoBERTa	94.31±0.19	98.81±0.20

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

	Intent Detection			Entity Extraction		
Activity	POC	Deploy	Δ	POC	Deploy	Δ
Intro (Meet & Greet)	99.92	97.46	-2.46	99.32	97.55	-1.77
Warm-up Game	98.91	93.54	-5.37	-	-	-
Training Game	98.48	94.27	-4.21	99.92	99.91	-0.01
Learning Game	99.02	94.37	-4.65	99.95	99.50	-0.45
Closure (Dance)	98.91	98.82	-0.09	-	-	-
All Activities	**98.92**	**94.36**	**-4.56**	**99.66**	**99.42**	**-0.24**

4 Experimental Results and Discussions

We train Intent and Entity Recognition models and cross-validate them on the Home POC data to pick the best-performing NLU architecture. Table 2 summarizes the results of NLU model selection experiments. Compared to StarSpace (baseline), we gain 2% F1 score for intents and 1% F1 for entities with multi-task DIET. We observe that incorporating DIET with BERT-family of embeddings achieves higher F1 relative to GPT-family. We cannot reveal any benefits of math-specific representations with DIET. Based on these results, we select DIET+ConveRT as our final NLU architecture.

Next, we evaluate our NLU module on Home Deployment data collected at authentic homes over 12 sessions/kids. In Table 3, we observe overall F1% drops (Δ) of 4.56 for intents and 0.24 for entities when our best DIET+ConveRT models are tested on Home Deployment data. These drops are relatively lower than what we observed at school Okur et al. [2022b]. We witness distributional and utterance-length differences between POC and Deployment 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 NLU models (trained on POC data) by leveraging 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 of these runs. That would simulate how combining POC with real-world deployment data can help us train more robust NLU models on unseen data in future deployments. The overall F1-scores reach 96.6% for intents (2.2% gain from 94.4%) and 99.6% for entities (0.2% gain) with LOOCV, which are promising for our future deployments.

Table 4: ASR Model Results: Average 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-tiny	1.055	1.027	1.002	1.027	0.964	0.919	0.983	0.880
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**

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

	Intent Detection		Entity Extraction	
ASR Model	F1	Adjusted-F1	F1	Adjusted-F1
Rockhopper	37.3	15.7	84.4	35.5
Google Cloud	**79.1**	40.3	97.0	49.4
Whisper-tiny	58.1	56.6	94.0	89.1
Whisper-base	64.9	60.3	95.9	91.0
Whisper-small	72.1	68.0	96.5	91.6
Whisper-medium	**76.7**	**73.3**	**98.2**	**93.8**

To inspect the ASR module, we evaluated various ASR engines on the same audio from Home Deployment data. We compute the average WER for kids with each ASR engine to investigate the most feasible solution. Table 4 summarizes WER after pre-processing steps (e.g., lower casing and punctuation removal) and domain-specific filters (e.g., num2word and cleaning). The cleaning applied to Whisper output due to known issues like repeat loops and hallucinations Radford et al. [2022]. We observe 4-to-7% Whisper output are trash (e.g., long transcriptions with repetitions), which hugely affect WER, yet these can be auto-filtered. Relatively high error rates can be attributed to the characteristics of recordings (e.g., incidental voice and phrases), very short utterances to be recognized (e.g., binary yes/no answers or stating numbers), and recognizing kids' speech in ordinary home environments. The results indicate that Whisper ASR performs better on kids, and we can benefit from increasing the model size.

For SLU pipeline evaluation, we test our best-performing NLU models on noisy ASR output. Table 5 presents the results achieved on Home Deployment data where the DIET+ConveRT models run on varying ASR models output. Voice Activity Detection (VAD) is an integral part of ASR that detects the presence of human speech. We realize that the VAD stage is filtering out many audio chunks with actual kid speech. Thus, VAD-adjusted F1s are compared in Table 5, aligned with the WER results, where NLU on Whisper performs relatively higher than Google and Rockhopper. Increasing the model size from tiny to medium worth the trouble for Whisper. Yet, F1 drop is still huge when VAD-ASR errors propagate into the SLU pipeline.

5 Conclusion

To increase the quality of early math education, we develop a multimodal dialogue system for play-based learning, helping the kids gain basic math skills. We investigate a modular SLU pipeline with cascading ASR and NLU modules, evaluated on our home deployment data with 12 kids. For NLU, we experiment with numerous pretrained language representations on top of a multi-task architecture for Intent and Entity Recognition. For ASR, we inspect the WER with several low-power, commercial, or open-source solutions with varying model sizes to conclude that Whisper-medium outperforms the rest on kids' speech at authentic homes. Finally, we evaluate the SLU pipeline by testing our best-performing NLU models on VAD-ASR output to observe the effects of cascaded errors due to noisy speech recognition with kids in realistic settings. In the future, we consider fine-tuning Whisper ASR acoustic models on kids' speech and language models on domain-specific math data, as well as exploring N-Best-ASR-Transformers Ganesan et al. [2021] to mitigate errors propagated in SLU.

Acknowledgments

We strive to share our gratitude and acknowledge our invaluable former and current colleagues in the Kid Space team at Intel Labs. Particularly: (i) Roddy Fuentes Alba and Celal Savur for developing the SDS ROS node and performing the unit/integration testing; (ii) Lenitra Durham for designing the HW/SW architectural setup, leading the deployment data collection, developing the game logic node, performing the integration testing, and supporting the Wizard UI development; (iii) Hector Cordourier Maruri, Juan Del Hoyo Ontiveros, and Georg Stemmer for developing the VAD-ASR ROS node and extracting the ASR output that we use in our SLU pipeline; (iv) Sinem Aslan for leading the UX studies to conceptualize school/home usages and pilot studies for data collection; (v) Ankur Agrawal, Arturo Bringas Garcia, Vishwajeet Narwal, and Guillermo Rivas Aguilar for developing the Student UI via the Unity game engine; (vi) Glen Anderson, Rebecca Chierichetti, Pete Denman, John Sherry, and Meng Shi for supporting the UX studies and performing interaction designs; (vii) David Gonzalez Aguirre, Gesem Gudino Mejia, and Julio Zamora Esquivel for developing the visual understanding nodes; (viii) Benjamin Bair, Sai Prasad, Giuseppe Raffa, and Sangita Sharma for their contributions to the HW/SW setup to support this research. In addition, we would like 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 thank the Rasa team and community developers for their open-source framework and contributions that empowered us to conduct our research.

References

G. J. Anderson, S. Panneer, M. Shi, C. S. Marshall, A. Agrawal, R. Chierichetti, G. Raffa, J. Sherry, D. Loi, and L. M. Durham. 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.

S. Aslan, A. Agrawal, N. Alyuz, R. Chierichetti, L. M. Durham, R. Manuvinakurike, E. Okur, S. Sahay, S. Sharma, J. Sherry, G. Raffa, and L. Nachman. 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.

S. Aslan, L. M. Durham, N. Alyuz, R. Chierichetti, P. A. Denman, E. Okur, D. I. G. Aguirre, J. C. Z. Esquivel, H. A. Cordourier Maruri, S. Sharma, G. Raffa, R. E. Mayer, and L. Nachman. What is the impact of a multi-modal pedagogical conversational ai system on parents' concerns about technology use by young children? *British Journal of Educational Technology*, 2023. doi: https://doi.org/10.1111/bjet.13399. URL https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13399.

Z. Azerbayev, B. Piotrowski, and J. Avigad. 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.

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

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

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

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

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

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

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

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

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

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

P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić. 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*, pages 5016–5026, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL `https://aclanthology.org/D18-1547`.

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

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

A. Cahill, J. H. Fife, B. Riordan, A. Vajpayee, and D. Galochkin. Context-based automated scoring of complex mathematical responses. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 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`.

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

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

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

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

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

D. Datta, M. Phillips, J. L. Chiu, G. S. Watson, J. P. Bywater, L. E. Barnes, and D. E. Brown. 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.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

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

S. Dutta, D. Irvin, J. Buzhardt, and J. H. Hansen. 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)*, pages 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.

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

F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 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.

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

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

K. Ganesan, P. Bamdev, J. B, A. Venugopal, and A. Tushar. 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)*, pages 93–98, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.14. URL https://aclanthology.org/2021.acl-short.14.

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

C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen. 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)*, pages 753–757, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2118. URL https://aclanthology.org/N18-2118.

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

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

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

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

S. Huang, J. Wang, J. Xu, D. Cao, and M. Yang. 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.

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

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

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

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

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

S. P. Lende and M. Raghuwanshi. 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)*, pages 1–6. IEEE, 2016.

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

X. Liu, A. Eshghi, P. Swietojanski, and V. Rieser. 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*, pages 165–183. Springer, 2021a.

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

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

E. Loginova and D. Benoit. 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)*, pages 101–106, 2022.

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

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

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

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

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

Nrupatunga, A. Kumar, and A. Rajagopal. 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`.

B. D. Nye, P. I. Pavlik, A. Windsor, A. M. Olney, M. Hajeer, and X. Hu. 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.

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

E. Okur, S. Sahay, R. Fuentes Alba, and L. Nachman. 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)*, pages 51–64, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022a. Association for Computational Linguistics. URL `https://aclanthology.org/2022.mathnlp-1.7`.

E. Okur, S. Sahay, and L. Nachman. 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*, pages 28–39, Marseille, France, June 2022b. European Language Resources Association. URL `https://aclanthology.org/2022.games-1.4`.

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

E. Okur, R. Fuentes Alba, S. Sahay, and L. Nachman. Inspecting spoken language understanding from kids for basic math learning at home. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, and T. Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 692–708, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.bea-1.56. URL `https://aclanthology.org/2023.bea-1.56`.

E. Okur, R. Fuentes Alba, S. Sahay, and L. Nachman. Assessing spoken language understanding pipeline of a multimodal dialogue system for kids learning math at home. In *Neural Conversational AI Workshop - What's left to TEACH (Trustworthy, Enhanced, Adaptable, Capable and Human-centric) chatbots? at ICML 2023*, 2023b. URL `https://drive.google.com/file/d/10gE09u6cKcHr8ZjR95Iqrc_ilrmCSnb8/view`.

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

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

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

A. C. Pires, F. González Perilli, E. Bakała, B. Fleisher, G. Sansone, and S. Marichal. 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.

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

S. Raamadhurai, R. Baker, and V. Poduval. 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*, pages 336–342, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4435. URL https://aclanthology.org/W19-4435.

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

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

M. Rathod, T. Tu, and K. Stasaski. Educational multi-question generation for reading comprehension. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 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.

K. Reeder, J. Shapiro, J. Wakefield, and R. D'Silva. 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.

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

R. Reyes, D. Garza, L. Garrido, V. De la Cueva, and J. Ramirez. 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*, pages 440–451. Springer, 2019.

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

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

S. Sahay, S. H. Kumar, E. Okur, H. Syed, and L. Nachman. 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, Sept. 2019. SEMDIAL. URL http://semdial.org/anthology/Z19-Sahay_semdial_0019.pdf.

S. Sahay, E. Okur, N. Hakim, and L. Nachman. Semi-supervised interactive intent labeling. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 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.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 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.

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

J. T. Shen, M. Yamashita, E. Prihar, N. T. Heffernan, X. Wu, and D. Lee. 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.

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

P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan. 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.

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

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

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

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

Y. Sun, T. Play, R. Nambiar, and V. Vidyasagaran. 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.

A. Suresh, J. Jacobs, C. Harty, M. Perkoff, J. H. Martin, and T. Sumner. 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*, pages 4654–4662, Marseille, France, June 2022a. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.497.

A. Suresh, J. Jacobs, M. Perkoff, J. H. Martin, and T. Sumner. 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)*, pages 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.

A. Tack and C. Piech. 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*, page 522, 2022.

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

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

G. Tyen, M. Brenchley, A. Caines, and P. Buttery. 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)*, pages 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.

J. Uesato, N. Kushman, R. Kumar, H. F. Song, N. Y. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins. 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.

A. Vanzo, E. Bastianelli, and O. Lemon. 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*, pages 254–263, Stockholm, Sweden, Sept. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5931. URL https://aclanthology.org/W19-5931.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 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*, pages 5998–6008, 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.

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

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

R. Winkler and M. Söllner. 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/.

R. Winkler, S. Hobert, A. Salovaara, M. Söllner, and J. M. Leimeister. 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*, pages 1–14, 2020.

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

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

L. Wu, A. Fisch, S. Chopra, K. Adams, and A. B. J. Weston. 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.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *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.

Z. Yang, J. Qin, J. Chen, L. Lin, and X. Liang. LogicSolver: Towards interpretable math word problem solving with logical prompt-enhanced learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1–13, Abu Dhabi, United Arab Emirates, Dec. 2022. Association

for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.1.

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

G. Yeung, R. Fan, and A. Alwan. 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)*, pages 6993–6997. IEEE, 2021.

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

X. Zhang and H. Wang. 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, pages 2993–2999. AAAI Press, 2016. ISBN 978-1-57735-770-4. URL http://dl.acm.org/citation.cfm?id=3060832.3061040.

Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. 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*, pages 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.

Z. Zhang, Y. Xu, Y. Wang, B. Yao, D. Ritchie, T. Wu, M. Yu, D. Wang, and T. J.-J. Li. 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*, pages 1–21, 2022.

6 Appendices

6.1 Related Work

6.1.1 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 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 and Ng [2016], Lende and 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 and Söllner [2018], Wambsganss et al. [2020], Winkler et al. [2020], Datta et al. [2020], Okonkwo and 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 and 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 and 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.

6.1.2 Language Representations for SLU

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 set-

tings 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.

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 and Lane [2016], Zhang and Wang [2016], Goo et al. [2018] with hierarchical learning approaches Wen et al. [2018], 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[2].

6.1.3 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[3] 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.

[2] `https://math.stackexchange.com/`

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

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 and 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 and 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 and 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.

6.2 Error Analysis

For NLU error analysis, Table 6 reveals utterance samples from our 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 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 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.

Table 6: NLU Error Analysis: Intent Recognition Error Samples from 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 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*

Table 8: ASR Error Samples from 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.

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.

6.3 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 Sahay et al. [2021]. 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).

6.4 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.

The multimodal data we collected for research purposes during our home deployment 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.