CareCorpus+: Expanding and Augmenting Caregiver Strategy Data to Support Pediatric Rehabilitation

*1,3Shahla Farzana, 2Ivana Lucero, 2Vivian Villegas, *2,4Vera Kaelin,
2Mary Khetani, and 1Natalie Parde
1Department of Computer Science, University of Illinois Chicago
2Department of Occupational Therapy, University of Illinois Chicago
3Institute for Population and Precision Health, University of Chicago
4Department of Computing Science, Umeå University
{sfarza3, ilucer3, vvilleg2, mkhetani, parde}@uic.edu, vera.kaelin@umu.se

Abstract

Caregiver strategy classification in pediatric rehabilitation contexts is strongly motivated by real-world clinical constraints but highly under-resourced and seldom studied in natural language processing settings. We introduce a large dataset of 3,062 caregiver strategies in this setting, a five-fold increase over the nearest contemporary dataset. These strategies are manually categorized into clinically established constructs with high agreement (κ=0.68-0.89). We also propose two techniques to further address identified data constraints. First, we manually supplement target task data with relevant public data from online child health forums. Next, we propose a novel data augmentation technique to generate synthetic caregiver strategies with high downstream task utility. Extensive experiments showcase the quality of our dataset. They also establish evidence that both the publicly available data and the synthetic strategies result in large performance gains, with relative F_1 increases of 22.6% and 50.9%, respectively.

1 Introduction

Globally, over 50 million children aged 0-5 years experience disability (Olusanya et al., 2018). These young children and their families benefit from timely access to quality pediatric rehabilitation services in diverse contexts, ranging from hospital to home and community (Olusanya et al., 2024). When rehabilitation providers create conditions for families to share their expertise about their child's attendance and involvement in valued activities, perceived supports and strategies, and priorities for change, they can engage families in shared decision-making to design and monitor a meaningful service plan (Crawford et al., 2022). Providers benefit from gathering this information from families to direct equitable care (Pinto et al., 2022; Jarvis and Fink, 2021; Magnusson and Khetani, 2022). However, parent-generated data is often collected and documented as free text narratives, necessitating efforts to extract and standardize content for clinical and research applications (Newman-Griffis et al., 2022a; Kaelin et al., 2024, 2022b).

Development of web-based tools (e.g., the Participation and Environment Measure (Coster and Khetani, 2008), also known as PEM) can diversify the capture and use of structured and narrative information from families to drive pediatric rehabilitation service design and improvement. Recent work established benchmarks for the detection and classification of caregiver strategies collected using two available versions of a PEM tool (Kaelin et al., 2023; Valizadeh et al., 2024). However, model performance was constrained by a limited availability of caregiver strategy data involving children, across a subset of relevant age ranges and rehabilitation care contexts. A larger, more diverse data source is needed to strengthen applicability across the broader pediatric rehabilitation care continuum.

We respond to that need in this work, making three primary contributions. First, we establish inclusion and exclusion criteria for data sources fitting one or more classes of caregiver strategies. We define primary class characteristics, identify relevant and irrelevant external sources, and construct or select prototypical samples. Next, we identify and prepare a subset of strategies data from (a) three datasets from prior related research (n=185 families (Jarvis et al., 2019; Khetani et al., 2015, 2023)) matching these criteria, and (b) publicly available data instances that also match established guidelines. Data are preprocessed and stored in a format compatible with existing caregiver strategies data (Valizadeh et al., 2024), and labeled according to strategy class with high reliability (κ=0.68-0.89). Finally, we propose a novel data augmentation technique to generate and filter caregiver strategies. We perform quality checks and performance comparisons to assess augmentation

[*] Work completed at University of Illinois Chicago.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6912–6927
November 12-16, 2024 ©2024 Association for Computational Linguistics

feasibility within this task domain. Ultimately, we find that our manually and synthetically augmented data improves strategy classification performance by a wide margin, establishing a new performance ceiling for this task (F_1=0.80).

2 Related Work

2.1 Caregiver Strategy Data

Minimal existing data is suitable for caregiver strategy classification in pediatric rehabilitation settings. Previously, Newman-Griffis et al. (2021) created a dataset of 289 clinical documents associated with claims for federal disability benefits from the U.S. Social Security Administration, and Chorianopoulou et al. (2017) created a dataset with a pediatric focus, collecting video-recorded sessions of children with autism spectrum disorder and typically developing children. However, these data did not relate to caregiver strategies for improving child participation in daily activities.

The closest relevant dataset is the recently released CareCorpus (Valizadeh et al., 2024), which contains 780 caregiver strategies organized into categories based on known drivers of child and youth participation (Imms et al., 2017). CareCorpus is drawn from a subset of data collected during a single pilot implementation trial of PEM in an early intervention program targeting children 0-3 years old (Kaelin et al., 2022a), which limits the generalizability of information that can be drawn from the corpus, in terms of both child demographics and service context. We (1) add data from more diverse pediatric rehabilitation care contexts (hospital, home, and community) as accessed by children across a broader age range (0-5 years); (2) introduce non-strategies from stylistically relevant sources; and (3) incorporate a synthetic data augmentation approach to further diversify our training strategies. This addresses limitations of CareCorpus, including data scarcity, homogeneity, and class imbalances.

2.2 Strategy Classification

Previously, Kaelin et al. (2023) extracted language features pertaining to speech and dependency tags, word sets, and Unified Medical Language System (Bodenreider, 2004) concepts to classify 1,576 caregiver strategies from families of youth ages 11-17 years with childhood-onset disabilities. They achieved promising performance, with macroaveraged F_1=0.58–0.83 across different classification granularities. Valizadeh et al. (2024) experimented with both feature-based models and popular pretrained language models (PLMs), finding competitive performance using a fine-tuned BERT model.

Valizadeh et al. (2024)'s study raised questions about whether general-domain pretraining data is still preferable to more health-focused data for diversified pediatric rehabilitation contexts as accessed by children across a broader age range. Moreover, replicating their study on a larger dataset enables assessment of the reproducibility and generalizability of their findings. Our work creates a sandbox extending from this for the study of data augmentation in specialized healthcare settings.

2.3 Data Augmentation

Data augmentation (DA) tackles data scarcity in low-resource NLP tasks by employing techniques to generate additional similar samples that vary along some dimension from the original source. A popular DA approach is rewriting or paraphrasing, by replacing words with synonyms and varying sentence structure while preserving overall meaning (Wei and Zou, 2019; Kobayashi, 2018; Gupta et al., 2017; Okur et al., 2022). Other common approaches include backtranslation, which involves translating data to and from one or more intermediate languages (Edunov et al., 2018; Yu et al., 2018), and data noising, which involves masking some words with random unigrams or placeholder tokens (Xie et al., 2017). These rule-based DA techniques may struggle with semantic diversity. In contrast, conditional generation involves finetuning a PLM to produce text conditioned on the target label (Bowman et al., 2016; Anaby-Tavor et al., 2020; Yang et al., 2020; Lee et al., 2021). However, it has traditionally required costly human labels (Sahu et al., 2022; Papangelis et al., 2021).

Large language models (LLMs) have emerged as a promising avenue for generating synthetic data, demonstrating remarkable rewriting capabilities (Radford et al., 2021). Their use addresses numerous limitations of prior DA approaches, prioritizing effectiveness and accessibility. However, LLM-generated data may be of dubious quality (Guerreiro et al., 2023). To preserve the quality and diversity of data augmented using LLMs, Ye et al. (2024) proposed *LLM-DA*, evaluating it in a named entity recognition setting and augmenting data at both the context and entity levels to align with characteristics inherent to the task. Ghorbani and Zou (2019) proposed *DATA SHAPLEY* in the biomedical text and image classification domains, generating data and evaluating its training utility for the target task. Likewise, Lin et al. (2023) introduced selective in-context data augmentation, evaluating synthetic examples before training an intent detection model. We adopt the selective data augmentation strategy for our work, leveraging the generative power of LLMs to address data scarcity while evaluating and selecting the most valuable examples to augment the training data to ensure sustained data quality. In contrast to Lin et al. (2023)'s work, we frame data augmentation as a paraphrasing task with various prompts, described in §4.1.

3 Data

3.1 Data Collection

We identified and combined common data elements across three diverse datasets from prior studies. Participant demographics for each data source are provided in Appendix A. We also sourced stylistically-relevant non-strategies to aid in classification. We complied with existing institutional review board (IRB) protocols in accessing all data, and ensured that our acquisition of non-strategies data was consistent with platform-specific terms and conditions.

Strategies Dataset A. These data come from a cross-sectional study establishing the Young Children's PEM (YC-PEM) psychometric properties (Khetani et al., 2015). Eligible caregivers (n=395): 1) could read and write English; 2) resided in the United States or Canada; 3) identified as parents or legal guardians 18+ years old; 4) had a child between 0-5 years old; and 5) had Internet access. A total of 93 caregivers of children with developmental disabilities and delays and accessing rehabilitation services (in the hospital, home, and/or community) are represented in the combined dataset.

Strategies Dataset B. These data come from a trial testing the preliminary effectiveness and implementation of the YC-PEM when paired with a program-specific decision support tool (Kaelin et al., 2022a; Khetani et al., 2023; Rizk et al., 2023). Eligible caregivers (n=76): 1) were at least 18 years of age; 2) identified as the parent or legal guardian of the child already enrolled in early intervention (EI) services at home and in the community; 3) had oral and written proficiency in English; 4) had Internet and phone access; 5) cared for a child 0-3 years old who had received EI for at least 3 months. A total of 39 caregivers assigned to the intervention group are included in the combined dataset.

Setting	% Agreement	κ
Multinomial		
Environment/Context	86.49	0.89
Sense of Self	73.32	0.69
Preferences	76.49	0.77
Activity Competence	69.42	0.68
No Strategy	94.89	0.89
Binary		
Strategy/No Strategy	94.89	0.89
ES/IS	88.40	0.74

Table 1: Inter-rater agreement measured using percent agreement and Cohen's Kappa (κ).

Strategies Dataset C. These data come from a prospective cohort study of children surviving critical illness (Choong et al., 2018; Jarvis et al., 2019; Khetani et al., 2018). Eligible caregivers (n=180) had children that were: 1) between ages 1-17 years old, and 2) had been admitted to the pediatric intensive care unit for at least 48 hours. A total of 53 caregivers with children aged 0-4 years old who thus completed the YC-PEM at study enrollment and 3 and 6 months post-discharge are included in the combined dataset.

Non-Strategies Data. These data come from four public health forums (*Patient.Info*,[1] *Mothering*,[2] *DC Urban Moms and Dads*,[3] and *Netmums*[4]). Eligible caregivers (n=1002): 1) had a child between 0-5 years old, and 2) had a child with a diagnosis or health issue. Data entries included descriptions of the child's behavior (e.g., "eats and plays well like his normal self"), questions (e.g., "Anyone else had this experience with a child?"), and caregiver-reported feelings about the child's health (e.g., "I am concerned").

3.2 Data Annotation

The diversified dataset includes 3,062 caregiver strategies. Two research team members (one male and one female undergraduate at a large, diverse university in an urban environment) independently annotated 50-250 strategies per week, using the same pediatric rehabilitation categories adopted

[1] https://patient.info/
[2] https://www.mothering.com/
[3] https://www.dcurbanmom.com/
[4] https://www.netmums.com/

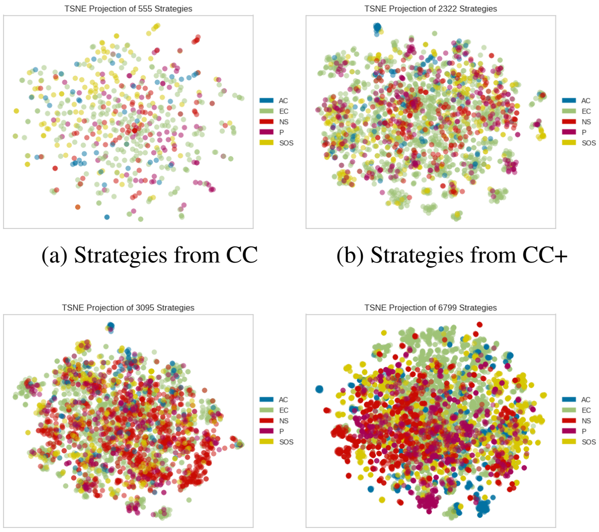
by Valizadeh et al. (2024), from March-August 2023. Annotators were paid $15.20 per hour. They met with a third annotator and research team member (a PhD-holding occupational therapist and re/habilitation scientist) to settle discrepancies by majority rule, seeking feedback from other key informants (project leads who direct rehabilitation science and natural language processing research labs) as needed. Data instances not meeting label criteria were categorized as *non-strategies*.

We calculated inter-annotator agreement using percent agreement and Cohen's kappa (Cohen, 1960), and report per-class agreement statistics as well as inter-annotator agreement across broader categorizations of the data in Table 1. In Appendix B, we present examples of strategies from each class, with the *Pediatric Habilitation Context* corresponding to samples from the data sources A and B, and the *Pediatric Rehabilitation Context* corresponding to examples from data source C.

3.3 Unique Qualities of the Dataset

The combined dataset[5] (n=3,062) is much larger than the original CareCorpus (n=780), with strategies spanning: 1) more diverse rehabilitation service contexts; and 2) the full early childhood period. We sourced our strategies from caregivers whose children were accessing pediatric rehabilitation services across the "clinic to community" care continuum, whereas prior work targeted children solely accessing community-based early intervention services. This more diverse range of service contexts affords for assessing performance in ways that can guide more varied downstream applications.

Similarly, since young children benefit from personalized rehabilitation service design through their preschool and kindergarten years, the expanded age range (0-5 years versus the previous 0-3 years) lends itself to application across a more comprehensive early childhood period. Finally, the non-strategies data responds to limitations in CareCorpus's class balance (Valizadeh et al., 2024). We visualize the semantic diversity of the extended dataset in Figure 1, comparing vector space representations of strategies from the original CareCorpus (CC), our new extended CareCorpus (CC+), CareCorpus+ with additional non-strategies (CC+NS), and CareCorpus+ with synthetically augmented data (described further in §4). In all cases, representations are TF-IDF vectors and visualizations are generated using t-distributed stochastic neighbor embedding (Hinton and Roweis, 2002, t-SNE). This also highlights the rich, complex nature of caregiver strategy classification. We report additional data statistics in Appendix C.



(a) Strategies from CC

(b) Strategies from CC+

(c) Strategies from CC+NS

(d) Strategies from CC+Aug

Figure 1: t-SNE visualizations of strategies from different strategy classes in four datasets: CC, CC+, CC+NS, and CC+Aug. *EC* =Environment/Context; *SOS* =Sense of Self; *AC* =Activity Competence; *P* =Preferences; *NS* =Non-Strategy.

4 Synthetic Data Augmentation

4.1 Prompt-Based Strategy Generation

To investigate the feasibility of synthetic dataset expansion for this task domain, we leveraged the open-source Flan-t5-xl (Chung et al., 2022), a $3B$-parameter autoregressive encoder–decoder LLM. This model is lightweight and has previously proven reliable for zero- or few-shot text generation tasks (Chung et al., 2022; Sterner et al., 2024), as well as for query reformulation tasks (Mo et al., 2023).[6] Its lightweight nature makes it well-suited for environments that are not expected to have substantial compute resources, such as occupational therapy settings. We did not fine-tune the model for the rephrasing task, but rather focused on prompting methods that rely on the model's ex-

[5] Data and code: https://github.com/treena908/CareCorpus-Plus

[6] We use the encoder-decoder based model only to generate synthetic data, and then use the augmented dataset is used to train downstream classification models, allowing us to leverage the power of PLMs while preserving the independence of the strategy classification model design.

ID	Prompt Template
a	Here is an example of **Environment/Context** strategy: Finding restaurants that are kid friendly. Please generate rewrite of the above strategy keeping the style similar. Find restaurants that are family friendly.
b	Here is an example of **Environment/Context** strategy in context of **outing**: Finding restaurants that are kid friendly. Please generate rewrite of the above strategy keeping the style similar. Whether its a cafeteria for school lunch or a fancy restaurant for a date night; you want it to be kid friendly.
c	Here is an example of **Environment/Context** strategy in context of **outing** in **community** setting: Finding restaurants that are kid friendly. Please generate rewrite of the above strategy keeping the style similar. Find out what's going on when it comes to family activities and restaurants that are kid friendly.

Table 2: Examples of the prompts used to generate synthetic examples. The strategy class is in black **bold** text, whereas the broader activity type and environment settings related to the strategy are in **violet** and **orange** bold text in prompt templates (b) and (c), respectively. Completions by the language model are in green.

isting knowledge and understanding to produce desired outputs. Likewise, rather than fine-tuning for noise reduction, we leveraged a filtering technique (described further in §4.2) to mitigate the impact of nonsense strategies generated by the model. These choices also felt more computationally viable for an occupational therapy setting.

We framed strategy augmentation as a paraphrasing task. For each strategy class, we created three versions of a natural language prompt, with different versions including (a) the class name, (b) the class name and broader activity context, and (c) the class name and setting. The broader activity context and setting were additional metadata values stored in our CC+ dataset. Broader activity contexts were drawn from {*chore*, *socializing*, *outing*, *classes and groups*, *basic care routine*, *recreational*, *educational*, *play*}, and settings were one of three environments: {*community*, *daycare/preschool*, *home*}.

Each of these versions was followed by an example from the training set, and then an instruction to rewrite the given example. We show the three different prompt versions for the strategy class *Environment/Context* in Table 2. For each input strategy, we generated nine synthetic outputs using these prompt templates with varying temperature values of $\{0.8, 0.9, 1.0\}$. We adopted random sampling with the repetition penalty set to 1.1 (Keskar et al., 2019). We also set the maximum output sequence length to 276, which was the maximum length of any strategy in the training set.

4.2 PVI Filtering

Given the diversity of our samples coupled with a diversity-oriented stochastic sampling generation strategy, we expected that some generated strategies would not match the desired strategy class. To filter for synthetic strategies anticipated to have high downstream task utility, we adopted the *In-Context Data Augmentation with PVI Filtering* algorithm (Lin et al., 2023). We retained synthetic strategies deemed relevant based on their Pointwise $\mathcal{V}$-Information (PVI) (Ethayarajh et al., 2022). The PVI of an input x with label y is calculated using predictive $\mathcal{V}$-entropy $g = H_{\mathcal{V}}(Y)$ and conditional $\mathcal{V}$-entropy $g' = H_{\mathcal{V}}(Y|X)$, with X and Y being random variables and $\mathcal{V}$ a predictive model family:

$$\text{PVI}(x \rightarrow y) = -\log_2 g[\varnothing](y) + \log_2 g'[x](y) \tag{1}$$

PVI was originally proposed as a mechanism for understanding dataset difficulty: it measures the amount of information that x provides to the classification model for learning y, compared to the absence of that input. High PVI suggests high information content, whereas low PVI suggests that the information gain is unlikely to be useful for learning the target class (Ethayarajh et al., 2022). Following Lin et al. (2023), we set a PVI threshold ϵ for each strategy class, where ϵ was the average PVI for the given strategy class in the validation set. Using these thresholds, our PVI filtering step discarded $11,397$ of the $15,873$ synthetically generated strategies.

5 Experiments and Results

5.1 Strategy Classification Models

To streamline comparison, we experimented with the same models considered by Valizadeh et al.

(2024): logistic regression (Lee et al., 2006), naïve Bayes (Joyce, 2003), BERT (Devlin et al., 2019), and Bio-ClinicalBERT (Alsentzer et al., 2019).[7] We represented strategies in statistical models using TF-IDF vectors with a vocabulary size of the 5000 most-frequent words in our dataset (Zhang et al., 2011). For BERT and Bio-ClinicalBERT, we used embeddings generated by the model's input layer.

We classified caregiver strategies across our full five-class data distribution, and set a baseline performance floor by predicting the most frequent class from the training set for each instance. Following precedent from earlier work, we also used our best-performing model to assess performance for the pipelined strategy/non-strategy (S/NS) and extrinsic/intrinsic strategy (ES/IS) classification tasks introduced by Valizadeh et al. (2024). These tasks predict broader categorizations of the data, with S/NS classification being a useful filtering step for some downstream applications and ES/IS classification reframing the strategy divisions along more general rehabilitation constructs. In motivating inclusion of these additional dataset divisions, we note that extracting and standardizing content from free text for clinical use is important on a more (i.e., in the finer-grained classes EC, SOS, P, and AC) as well as less (i.e., the more simplified classes ES and IS) level. Automated distinction at the ES/IS level may empower clinicians to start differentiating between more specific strategy types, facilitating decision-making without the need for finer-grained precision Valizadeh et al. (2024).

5.2 Experimental Setup

We split the CC+ corpus into 90:10 train:test sets. For experiments with statistical models, we optimized model parameters via 10-fold cross-validation on the training set (Refaeilzadeh et al., 2009). For experiments with pre-trained language models (BERT and Bio-ClinicalBERT) we further subdivided the training set into a training and validation set, resulting in an 80% training, 10% validation, and 10% test split. In these cases, the validation set was used during fine-tuning for hyperparameter optimization. The held-out test set remained consistent across all conditions.

In the CC+NS and CC+Aug conditions, the training data was augmented with the non-strategy data described earlier (CC+NS), as well as with synthetically generated data (CC+Aug). However, the validation and test data was never augmented, meaning that we used the same CC+ test set (with no augmented data present) for all conditions. To avoid potential training biases, we kept all strategies authored by the same caregiver in the same data split for all models. We trained BERT and Bio-ClinicalBERT using a learning rate of 2e-6 and batch size of 16, for 10 epochs. All models were trained and evaluated using one NVIDIA Tesla V100 GPU with 32 GB of memory.

We compared strategy classification performance when training models on the original CareCorpus (CC), our manually expanded dataset without additional non-strategies (CC+), our manually expanded dataset with non-strategies sourced from online forums (CC+NS), and our manually expanded dataset with synthetically augmented data (CC+Aug). For CC+NS, we under-sampled the non-strategy class (retaining $\sim$30% of non-strategies from online forums) due to its comparatively high frequency. We also used class weights to penalize minority class misclassification.

5.3 Results

We report our results in Table 3. Results using CC are reported directly from Valizadeh et al. (2024)'s paper. Results for BERT and Bio-ClinicalBERT models are averaged across three runs with different random seeds, with standard deviations included in parentheses. We observe a trend of increased performance with added non-strategy data from online forums (CC+NS), as well as with augmented data (CC+Aug). Performance improvements for CC+ over CC are inconsistent, with improvements observed using logistic regression (F_1=0.57 versus F_1=0.46) and Bio-ClinicalBERT (F_1=0.44 versus F_1=0.39) but not for naïve Bayes or BERT. This is unsurprising, given the intentional increased diversity of strategy samples in CC+. We nearly uniformly observe the highest performance using a fine-tuned BERT model, successfully replicating findings from Valizadeh et al. (2024). Our overall highest-performing model is BERT fine-tuned using CC+Aug, achieving F_1=0.80.

We also report five-class strategy classification performance with varying numbers of training instances from the CC+Aug dataset in Figure 2. We observe the best performance in both accuracy and

[7] We note that while it is likely that higher performance may have been achieved with more targeted focus on classification model design and selection, our emphasis in this work was on investigating the impact of manual and synthetic expansion of data in this highly specialized setting; maintaining model consistency with contemporary relevant work allowed us to control more fully for our variables of interest.

Data	Model	Acc.	P	R	F_1
CC	Base	40.78	0.08	0.20	0.11
	LR	57.89	0.69	0.43	0.46
	NB	53.95	0.85	0.38	0.38
	BERT	**64.47**	**0.73**	**0.53**	**0.56**
	Bio	53.94	0.71	0.40	0.39
CC+	Base	59.64	0.12	0.20	0.15
	LR	74.48	0.86	0.49	0.57
	NB	61.72	0.72	0.24	0.23
	BERT	60.78 (0.02)	0.54 (0.01)	0.62 (0.00)	0.53 (0.01)
	Bio	48.74 (0.04)	0.46 (0.01)	0.54 (0.03)	0.44 (0.03)
CC+ NS	Base	59.64	0.11	0.20	0.16
	LR	75.26	0.77	0.56	0.62
	NB	65.36	0.86	0.33	0.34
	BERT	72.77 (0.01)	0.63 (0.01)	0.69 (0.01)	0.65 (0.01)
	Bio	54.46 (0.05)	0.52 (0.03)	0.56 (0.00)	0.48 (0.02)
CC+ Aug	Base	59.64	0.11	0.20	0.16
	LR	82.55	0.82	0.71	0.75
	NB	75.00	0.79	0.54	0.60
	BERT	**83.56** (0.01)	**0.79** (0.01)	**0.82** (0.01)	**0.80** (0.00)
	Bio	80.48 (0.01)	0.75 (0.01)	0.79 (0.01)	0.76 (0.01)

Table 3: Performance in a five-class setting. *Acc.* = accuracy (%), *P* = precision, *R* = recall, *LR* = logistic regression, *NB* = naïve Bayes, and *Bio* = Bio-ClinicalBERT.

F_1 with a training set size of $n = 6799$; as mentioned earlier, the PVI filtering threshold was optimized for each strategy class using the validation set (see Table 5 for example PVI values corresponding to different synthetic samples). Optimized thresholds thus varied with varying training set sizes. When we selected the training instances based on threshold set to the average PVI value (across the strategy class) on the validation set, which increases the training set size to $n = 9773$, the performance dropped. This presents additional evidence supporting the efficacy of per class PVI filtering of synthetically generated instances in the downstream strategy classification task.

In Table 4, we report results under these same conditions for the pipelined S/NS and ES/IS classification settings. We compare these conditions only using BERT, following our findings from Table 3. We do not report ES/IS classification results for CC+NS, since ES/IS classification predicts divisions only between strategy data (leading to unchanged training conditions from CC+). We observe similar trends to those observed in the all-class setting for the S/NS setting, with equivalent performance when training on CC+ versus CC and increased performance when training on both CC+NS (F_1=0.93) and CC+Aug (F_1=0.89). We observe a dramatic performance increase in the ES/IS classification setting when training on CC+ versus CC (F_1=0.83 versus F_1=0.53), and greater performance still when training on CC+Aug (F_1=0.91).

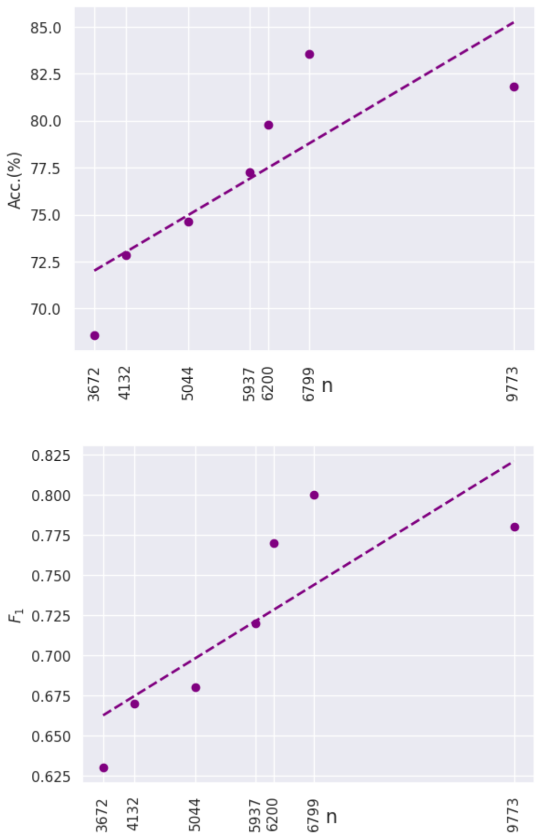


Figure 2: Five-class strategy classification performance with the varying number of training instances (n) from the CC+Aug dataset using the fine-tuned BERT model.

5.4 Error Analyses

We systematically analyzed errors to identify areas for improvement. We (a) studied synthetic examples in the context of their PVI, calculated as described in §4.2; and (b) examined misclassifications from our best-performing model in Table 3. We provide prototypical samples from each analysis in Tables 5 and 6.

Table 5 shows manually authored strategies and corresponding strategies that were generated when these manual strategies were used as demonstra-

Data	Task	Acc.	P	R	F_1
CC	S/NS	90.60 (0.00)	0.90 (0.00)	0.86 (0.01)	0.87 (0.00)
	ES/IS	58.06	0.64	0.58	0.53
CC+	S/NS	90.60 (0.02)	0.90 (0.00)	0.86 (0.01)	0.87 (0.00)
	ES/IS	84.97 (0.02)	0.82 (0.02)	0.85 (0.00)	0.83 (0.01)
CC+ NS	S/NS	95.02 (0.00)	0.95 (0.00)	0.92 (0.01)	0.93 (0.00)
	ES/IS	-	-	-	-
CC+ Aug	S/NS	91.78 (0.00)	0.92 (0.00)	0.86 (0.00)	0.89 (0.00)
	ES/IS	92.18 (0.00)	0.90 (0.00)	0.92 (0.00)	0.91 (0.00)

Table 4: Model comparison for pipelined classification tasks, using the same metrics as in Table 3. All conditions use a fine-tuned BERT model.

tions during data augmentation, paired with the calculated PVI for the generated strategy. We showcase both high-PVI and low-PVI examples, with low PVIs emphasized in red.[8] Broadly speaking, we observed that high-PVI samples tended to vary the writing style while adhering closely to the content conveyed in the demonstration; often this was because the demonstration was straightforward to process (e.g., *Encourage to help tidy and put away prior to moving to another activity* → *Encourage them to help with the chores ahead of time*). Low-PVI examples typically demonstrated a lack of understanding of the source content, either for unknown reasons (e.g., *Save money to hire a babysitter for parent night out* → *Kidnappers are better at staying up late*) or due to noise or other complexities in the demonstration (e.g., *It takes 2 to talk-??? Program for speech therapy* → *Talking is a very relaxing way to relax*).

Table 6 shows mispredictions with their actual and predicted labels. We find that non-strategies using caregiving language are often misclassified as belonging to strategy classes. Although these cases (e.g., *Teachers are knowledgeable about my child's needs + abilities*) do not include specific strategy content, their style is close to that observed in actual strategies. Non-strategies using caregiving language are a minority; more common non-strategies in CC and CC+ are *N/A* and *None*. Future classification approaches that more closely target underlying intent or actionable language may be able to address challenging non-strategies. This could also allow for better capture of atypically worded actual strategies, which are sometimes missed with our current models (e.g., *If it is a sequence of events we will try and go back to where a step was missed*).

6 Discussion

Our findings broadly support the premise that diverse data can be leveraged for specialized pediatric rehabilitation contexts, justifying our manual data curation and our investigation of synthetic data expansion within this domain. It also replicates findings from Valizadeh et al. (2024) across a broader age range and rehabilitation care continuum; given the heterogeneity of pediatric rehabilitation, these findings encourage the application of automated strategy classification for this use case. Performance continues to improve with the addition of non-strategy data harvested from publicly available online forums pertaining to child health, suggesting an accessible avenue for increasing performance. Useful applications of this may include initiating caregiver education when detecting non-strategy responses during PEM completion (Villegas et al., 2023), enabling skilled engagement for families on waitlists to access rehabilitation services (McManus et al., 2019).

Importantly, we also find that prompt-based synthetic data expansion improves model performance. As LLMs are nascent in rehabilitation (Bonnechère, 2024), this establishes evidence that LLMs are capable of adequately replicating user-generated content in this domain, to an extent that it improves model performance when used as a training source. It also raises intriguing questions for future work, such as whether LLMs could be used to consolidate strategies of the same type. This could address known barriers to family-centered goal-setting practice (Crawford et al., 2022). It could also decrease caregiver burden when searching through a large bank of user-generated strategies, such as that in the PEM-Plus goal setting application, sustaining longer-term feasibility of user-driven strategy exchange. Despite promising technical merit and potential clinical utility, there are ethical considerations to synthetic data expansion, such as its potential to compromise family voice (Newman-Griffis et al., 2022b). This presents a ripe opportunity for further study.

[8] Maximum PVI among our synthetic examples is 1.701, and minimum PVI is -2.417.

Class	Demonstration Strategy	Generated Strategy	PVI
EC	*Playing together as a family*	Playing with friends together	0.674
$\epsilon = 0.450$	*read books about outings*	Read books about the outdoors.	0.581
	Save money to hire a babysitter for parent night out.	Kidnappers are better at staying up late	−0.700
SOS	*Encourage to help tidy and put away prior to moving to another activity*	Encourage them to help with the chores ahead of time.	1.167
$\epsilon = 0.549$	*Positive encouragement*	To encourage a feeling of self you will receive positive reinforcement from others.	0.715
	House guests - I like to have people over that include our son and get along well with him.	A couple who are very close are going to do a good job at this house.	−0.417
P	*Have options to write or color - he focuses well on writing or drawing*	He can choose whether he should write his way or to draw.	1.188
$\epsilon = 0.928$	*finding fun puzzels on tablet*	Find fun puzzles on a computer	0.643
	Finding restaurants that serve food that our son likes.	Find restaurants that serve food with prices that our son considers affordable	−0.144
AC	*Learn sign language or her gestures*	Have kids learn to communicate by pointing fingers and hand gesture	1.071
$\epsilon = 0.740$	*help him interact with others*	Observe how he interacts with others and help him interact with others	0.505
	Development (fine + gross motor)	Using a computer is difficult.	−0.820
NS	*Wish we had more free time for fun outings, but can't with all the therapy.*	We wish we had more free time	1.450
$\epsilon = 0.480$	*He needs help staying occupied.*	He needs help staying occupied in front of other people	1.322
	It takes 2 to talk- ??? Program for speech therapy	Talking is a very relaxing way to relax	−0.456

Table 5: Synthetic examples indicate paired with their corresponding demonstrations. Red texts with low PVI values highlight synthetic strategies that were hard to classify according to the PVI threshold (ϵ) for each class. *EC*=Environment/Context; *SOS*=Sense of Self; *AC*=Activity Competence; *P*=Preferences; *NS*=Non-Strategy.

Strategy	Actual	Pred.
then a hug	NS	SOS
and miss out on treats through her own choice.	NS	P
write and draw absolutely fine.	NS	EC
If it is a sequence of events we will try and go back to where a step was missed	EC	NS
Teachers are knowledgeable about my child's needs + abilities	NS	AC

Table 6: Misclassified examples (BERT model). Abbreviations are similar to Table 5.

7 Conclusion

In this work, we introduced an expanded dataset (n=3,062, an approximately five-fold increase over prior work) of caregiver strategies in diverse pediatric rehabilitation contexts. The strategies were identified from prior pediatric rehabilitation studies and manually assigned to clinically-established constructs by trained researchers spanning a rigorous five-month annotation process, resulting in strong agreement (κ=0.68–0.89). We also identified non-strategies from publicly available online child health forums to supplement the data and address previously-identified class balance issues (Valizadeh et al., 2024).

Additionally, we proposed a new technique for synthetic data augmentation in this domain, guided by three diverse prompts leveraging task-relevant contextual information while filtering for synthetic strategies with high anticipated task utility. We demonstrated the value of the additional manually curated strategies, publicly available task-relevant non-strategies, and our novel data augmentation approach for the downstream task of caregiver strategy classification. Our results establish evidence that both publicly available non-strategies (F_1=0.65, a 22.6% relative increase over the use of CC+ alone) and prompting-based synthetic strategies (F_1=0.80, a 50.9% relative increase) can support impressive performance gains in this highly specialized and under-resourced domain.

Limitations

This work is limited by two factors. First, we curated a larger, more diverse, and more balanced dataset relative to prior work, but implementation of PEM as a candidate common data element into data capture systems across diverse pediatric rehabilitation contexts is both possible and needed (Schiariti et al., 2018; Pinto et al., 2022) to further overcome data scarcity and homogeneity when examining generalizability in the longer-term. Second, our dataset is limited to English-language strategies. It remains unclear whether the results can be reproduced in less-resourced languages. Culturally adapted versions of PEM do exist (Krieger et al., 2020; Tomas et al., 2022), although they have been the subject of less research to date. We are committed to extending this work to additional data sources derived from use of culturally adapted PEM version(s) as permitted.

Ethical Considerations

A guiding motivation of our work is to enable more equitable support for caregivers of children with diverse rehabilitation needs. In pursuing this goal, we have been cognizant of the intersectional biases present in the contemporary pediatric rehabilitation community, along dimensions including race, ethnicity, and socioeconomic status. We report demographic items pertaining to these factors for our data sources in Table 7, to the extent that they are available and recognizing that education level is an imperfect proxy for socioeconomic status.

We intend for our caregiver strategy classification and strategy augmentation approaches to be deployed for the specific uses outlined in §6. Specifically, classification of caregiver strategies can reduce burden for caregivers searching through public strategy banks and enable educational opportunities regarding strategy development. Strategy augmentation can foster improved strategy classification, as established in Tables 3 and 4. A risk of strategy augmentation is that it holds potential to diminish family voice, although we note that in our work generated strategies are not provided to users nor used in any way to recommend strategy quality. We urge future research studying strategy augmentation in other pediatric rehabilitation contexts to examine this ethical consideration further.

Our dataset will be distributed under the licensing terms of the source datasets and in ways that are consistent with our prior work. CareCorpus (Valizadeh et al., 2024) was made available in the Inter-university Consortium for Political and Social Research (ICPSR) portal (Kaelin et al., 2023b). ICPSR requests IRB approval for researchers to access this additional linked dataset to ensure research done with this dataset aligns with ethical regulations and principles.

Similarly, the CareCorpus+ dataset is comprised of deidentified strategies data from sources A (Khetani et al., 2015), B (Kaelin et al., 2022a; Khetani et al., 2023; Rizk et al., 2023), and C (Choong et al., 2018; Jarvis et al., 2019; Khetani et al., 2018), and ethics approval was obtained prior to participant recruitment, each participant provided consent for study participation and was informed about their rights to withdraw their participation at any time, and in most cases were compensated with a gift card. Our use of these existing artifacts was consistent with their intended use, as specified in those source publications. Data are anonymized and available upon author request, provided that existing institutional review board approval is provided and protocol is followed. Manually-curated non-strategy data is publicly available following the terms and conditions of the web sources from which it was downloaded. We provide a link to a publicly available repository to facilitate straightforward acquisition of data, as well as source code to replicate our experiments, under a Creative Commons Attribution-NonCommercial 4.0 International license. Derivatives of our work accessed for research purposes should not be deployed for purposes other than as a research prototype. To foster reproducibility, we report our experimental setup, relevant statistics for running, and hyperparameters in §5. We report means and standard deviations in Tables 3 and 4 for BERT and Bio-ClinicalBERT, averaging results across three runs with different random seeds.

8 Acknowledgements

Data for this study were collected with funding from the National Institutes of Health (NIH grant number 1K12 HD055931; PI: M. Khetani), Academic Health Sciences AFP Innovation Fund (PI: K. Choong, Collaborator: M. Khetani), and the American Occupational Therapy Foundation (AOTF grant number AOTFIR20KHETANI; PI: M. Khetani). Analyses of these data were supported by the National Science Foundation (award #2125411; PIs: N. Parde, M. Khetani, J. Dooling-Litfin). The content is the responsibility of the authors and does not necessarily represent the official views of these funding agencies. We thank Shayan Rasheed at University of Illinois Chicago for assisting with data annotation.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7383–7390.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Bruno Bonnechère. 2024. Unlocking the black box? a comprehensive exploration of large language models in rehabilitation. *American Journal of Physical Medicine & Rehabilitation*, pages 10–1097.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Karen Choong, Samah Al-Harbi, Asm Borham, Jill Cameron, Saoirse Cameron, Ji (Emmy) Cheng, Heather Clark, Tim Doherty, Nora Fayed, Jan Willem Gorter, Margaret Herridge, Mary Khetani, Kusum Menon, Jamie Seabrook, Racquel Simpson, and Lehana Thabane. 2018. Functional recovery in critically ill children, the "weecover" multicenter study. *Pediatric Critical Care Medicine*, 19:145–154.

Arodami Chorianopoulou, Efthymios Tzinis, Elias Iosif, Asimenia Papoulidi, Christina Papailiou, and Alexandros Potamianos. 2017. Engagement detection for children with autism spectrum disorder. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5055–5059, New Orleans, USA. IEEE.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Wendy Coster and Mary Alunkal Khetani. 2008. Measuring participation of children with disabilities: Issues and challenges. *Disability and rehabilitation*, 30(8):639–648.

L Crawford, J Maxwell, H Colquhoun, S Kingsnorth, D Fehlings, S Zarshenas, S McFarland, and Nora Fayed. 2022. Facilitators and barriers to patient-centred goal-setting in rehabilitation: a scoping review. *Clinical Rehabilitation*, 36(12):1694–1704.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Amirata Ghorbani and James Y. Zou. 2019. Data shapley: Equitable valuation of data for machine learning. *ArXiv*, abs/1904.02868.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2017. A deep generative framework for paraphrase generation. In *AAAI Conference on Artificial Intelligence*.

Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Christine Imms, Mats Granlund, Peter H. Wilson, Bert Steenbergen, Peter L. Rosenbaum, and Andrew M. Gordon. 2017. Participation, both a means and an end: A conceptual analysis of processes and outcomes in childhood disability. *Developmental Medicine & Child Neurology*, 59(1):16–25.

Jessica M. Jarvis, Karen Choong, and Mary A. Khetani. 2019. Associations of participation-focused strategies and rehabilitation service use with caregiver stress after pediatric critical illness. *Archives of Physical Medicine and Rehabilitation*, 100(4):703–710.

Jessica M Jarvis and Ericka L Fink. 2021. More than a feeling: understanding function and health related quality of life after pediatric neurocritical illness. *Neurocritical care*, 35(2):308–310.

James Joyce. 2003. Bayes' theorem. In *The Stanford Encyclopedia of Philosophy (Fall 2021 Edition)*. The Metaphysics Research Lab, Philosophy Department, Stanford University.

Vera Kaelin, Vivian Villegas, Yi-Fan Chen, Natalie Murphy, Elizabeth Papautsky, Jodi Litfin, Natalie Leland, Varun Maheshwari, Beth McManus, and Mary Khetani. 2022a. Effectiveness and scalability of an electronic patient-reported outcome measure and decision support tool for family-centred and participation-focused early intervention: PROSPECT hybrid type 1 trial protocol. *BMJ open*, 12(1):e051582.

Vera C. Kaelin, Dianna L. Bosak, Shivani Saluja, Denis Newman-Griffis, Andrew D. Boyd, and Mary A. Khetani. 2024. Representation of child and youth participation within the unified medical language system (umls). *Disability and Rehabilitation*, 0(0):1–6. PMID: 38596871.

Vera C. Kaelin, Andrew D. Boyd, Martha M. Werler, Natalie Parde, and Mary A. Khetani. 2023. Natural language processing to classify caregiver strategies supporting participation among children and youth with craniofacial microsomia and other childhood-onset disabilities. *Journal of Healthcare Informatics Research*, 7(4):480–500.

Vera C. Kaelin, Mina Valizadeh, Zurisadai Salgado, Julia G. Sim, Dana Anaby, Andrew D. Boyd, Natalie Parde, and Mary A. Khetani. 2022b. Capturing and operationalizing participation in pediatric re/habilitation research using artificial intelligence: A scoping review. *Frontiers in Rehabilitation Sciences*, 3.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

M.A. Khetani, V. Kaelin, S. Rizk, M. Angulo, Z. Salgado, Y.F. Chen, V. Villegas, J. Dooling-Litfin, N. Leland, E. Lerner Papautsky, N. Murphy, B. McManus, and High Value Early Intervention Research Group. 2023. Preliminary effectiveness of an electronic patient-reported outcome measure and decision support tool on early intervention service quality. *Developmental Medicine & Child Neurology*, 65(S3):5–87.

Mary Khetani, Erin Albrecht, Jessica Jarvis, David Pogorzelski, Ji (Emmy) Cheng, and Karen Choong. 2018. Determinants of change in home participation among critically ill children. *Developmental Medicine & Child Neurology*, 60.

Mary A. Khetani, James E. Graham, Patricia L. Davies, Mary C. Law, and Rune J. Simeonsson. 2015. Psychometric properties of the young children's participation and environment measure. *Archives of Physical Medicine and Rehabilitation*, 96(2):307–316.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. 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 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Beate Krieger, Christina Schulze, Jillian Boyd, Ruth Amann, Barbara Piškur, Anna Beurskens, Rachel Teplicky, and Albine Moser. 2020. Cross-cultural adaptation of the participation and environment measure for children and youth (pem-cy) into german: a qualitative study in three countries. *BMC Pediatrics*, 20(1):492.

Kenton Lee, Kelvin Guu, Luheng He, Timothy Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. *ArXiv*, abs/2102.01335.

Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y. Ng. 2006. Efficient l˜ 1 regularized logistic regression. In *AAAI*, volume 6, pages 401–408, Boston, USA.

Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023. Selective in-context data augmentation for intent detection using pointwise V-information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1463–1476, Dubrovnik, Croatia. Association for Computational Linguistics.

D. Magnusson and M.A. Khetani. 2022. Early intervention. In Heidi M Feldman, Ellen Roy Elias, Nathan J Blum, Manuel Jimenez, and Terry Stancin, editors, *Developmental and Behavioral Pediatrics*. Elsevier.

Beth M. McManus, Zachary Richardson, Margaret Schenkman, Natalie Murphy, and Elaine H. Morrato. 2019. Timing and Intensity of Early Intervention Service Use and Outcomes Among a Safety-Net Population of Children. *JAMA Network Open*, 2(1):e187529–e187529.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. ConvGQR: Generative query reformulation for conversational search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4998–5012, Toronto, Canada. Association for Computational Linguistics.

Denis Newman-Griffis, Jonathan Camacho Maldonado, Pei-Shu Ho, Maryanne Sacco, Rafael Jimenez Silva, Julia Porcino, and Leighton Chan. 2021. Linking free text documentation of functioning and disability to the icf with natural language processing. *Frontiers in Rehabilitation Sciences*, 2:1–17.

Denis R. Newman-Griffis, Max B. Hurwitz, Gina P. McKernan, Amy J. Houtrow, and Brad E. Dicianno. 2022a. A roadmap to reduce information inequities in disability with digital health and natural language processing. *PLOS Digital Health*, 1(11):1–19.

Denis R Newman-Griffis, Max B Hurwitz, Gina P McKernan, Amy J Houtrow, and Brad E Dicianno. 2022b. A roadmap to reduce information inequities in disability with digital health and natural language processing. *PLOS Digital Health*, 1(11):e0000135.

Eda Okur, Saurav Sahay, and Lama Nachman. 2022. 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. European Language Resources Association.

Bolajoko O Olusanya, Adrian C Davis, Donald Wertlieb, Nem-Yun Boo, MKC Nair, Ricardo Halpern, Hannah Kuper, Cecilia Breinbauer, Petrus J De Vries, Melissa Gladstone, et al. 2018. Developmental disabilities among children younger than 5 years in 195 countries and territories, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Global Health*, 6(10):e1100–e1121.

Bolajoko O Olusanya, Scott M Wright, Tracey Smythe, Mary A Khetani, MARISOL MORENO-ANGARITA, Sheffali Gulati, Sally A Brinkman, Nihad A Almasri, Marta Figueiredo, Lidia B Giudici, et al. 2024. Early childhood development strategy for the world's children with disabilities. *Frontiers in Public Health*, 12:1390107.

Alexandros Papangelis, Karthik Gopalakrishnan, Aishwarya Padmakumar, Seokhwan Kim, Gokhan Tur, and Dilek Hakkani-Tur. 2021. Generative conversational networks. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 111–120, Singapore and Online. Association for Computational Linguistics.

Neethi P Pinto, Aline B Maddux, Leslie A Dervan, Alan G Woodruff, Jessica M Jarvis, Sholeen Nett, Elizabeth Y Killien, Robert J Graham, Karen Choong, Peter M Luckett, et al. 2022. A core outcome measurement set for pediatric critical care. *Pediatric Critical Care Medicine*, 23(11):893–907.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Payam Refaeilzadeh, Lei Tang, and Huan Liu. 2009. Cross-validation. *Encyclopedia of Database Systems*, 5:532–538.

Sabrin Rizk, Vera C Kaelin, Julia Gabrielle C Sim, Natalie J Murphy, Beth M McManus, Natalie E Leland, Ashley Stoffel, Lesly James, Kris Barnekow, Elizabeth Lerner Papautsky, and Mary A Khetani. 2023. Implementing an electronic patient-reported outcome and decision support tool in early intervention. *Applied clinical informatics*, 14(1):91—107.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Verónica Schiariti, Eileen Fowler, Joline E Brandenburg, Eric Levey, Sarah Mcintyre, Theresa Sukal-Moulton, Sharon L Ramey, Jessica Rose, Susan Sienko, Elaine Stashinko, Laura Vogtle, Robin S Feldman, and James I Koenig. 2018. A common data language for clinical research studies: the national institute of neurological disorders and stroke and american academy for cerebral palsy and developmental medicine cerebral palsy common data elements version 1.0 recommendations. *Developmental Medicine & Child Neurology*, 60(10):976–986.

Igor Sterner, Weizhe Lin, Jinghong Chen, and Bill Byrne. 2024. Few-shot VQA with frozen llms: A tale of two approaches. *CoRR*, abs/2403.11317.

Vanessa Tomas, Roopa Srinivasan, Vrushali Kulkarni, Rachel Teplicky, Dana Anaby, and Mary Khetani. 2022. A guiding process to culturally adapt assessments for participation-focused pediatric practice: the case of the participation and environment measures (pem). *Disability and Rehabilitation*, 44(21):6497–6509.

Mina Valizadeh, Vera Kaelin, Mary Khetani, and Natalie Parde. 2024. Carecorpus: A corpus of real-world solution-focused caregiver strategies for personalized pediatric rehabilitation service design. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Turin, Italy. European Language Resources Association.

Vivian C Villegas, Dianna L Bosak, Zurisadai Salgado, Michelle Phoenix, Natalie Parde, Rachel Teplicky, Mary A Khetani, and High Value Early Intervention Research Group Kuznicki L. Pedrow A. Howell A. 2023. Diversified caregiver input to upgrade the young children's participation and environment measure for equitable pediatric re/habilitation practice. *Journal of Patient-Reported Outcomes*, 7(1):87.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Allen Nie, Dan Jurafsky, and A. Ng. 2017. Data noising as smoothing in neural network language models. *ArXiv*, abs/1703.02573.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.

Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *ArXiv*, abs/2402.14568.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *ArXiv*, abs/1804.09541.

Wen Zhang, Taketoshi Yoshida, and Xijin Tang. 2011. A comparative study of tf* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765.

A Child and Caregiver Characteristics

We report relevant demographic characteristics of the children and caregivers represented in data sources A, B, and C in Table 7. Characteristics include child age, sex, and disability status, as well as caregiver race, ethnicity, and education level. Mean child age is reported with interquartile range in parentheses; all other characteristics are reported as frequencies with percentage in parentheses.

B Sample Caregiver Strategies

We include sample caregiver strategies in Table 8. Samples from the *Pediatric Habilitation Context* correspond to data sources A and B. Samples from the *Pediatric Rehabilitation Context* are from source C. We abbreviate strategy types as *EC* (Environment/Context), *SOS* (Sense of Self), *AC* (Activity Competence), and *P* (Preferences). While pediatric rehabilitation focuses on helping children redevelop skills they have lost, pediatric habilitation focuses on helping them develop new skills.

C Dataset Statistics

We report statistics for our CC+, CC+NS, and CC+Aug training datasets in Table 9. We use consistent validation and test sets across all conditions, and their class distributions are as follows:

- **Validation:**
  - *Environment/Context*: 212
  - *Sense of Self*: 48
  - *Preferences*: 37
  - *Activity Competence*: 22
  - *Non-Strategy*: 36
- **Test:**
  - *Environment/Context*: 229
  - *Sense of Self*: 52
  - *Preferences*: 40
  - *Activity Competence*: 25
  - *Non-Strategy*: 38

We also report average strategy length for each training set in Table 9, with standard deviation provided in parentheses.

Characteristic	**Source A (n=39)**	**Source B (n=53)**	**Source C (n=93)**
Child Age, M(IQR)	2.4 (1.9, 2.6)	-	3.2 (1.3, 4.6)
Child Sex, n (%)			
Female	12 (30.8)	25 (44.6)	30 (32.3)
Male	27 (69.2)	31 (55.4)	63 (67.7)
Child Disability Status, n (%)			
Developmental Delay/At Risk	27 (69.2)	0 (0.0)	41 (44.0)
Diagnosed Condition	12 (30.8)	52 (98.1)	52 (55.9)
Caregiver Race, n (%)			
American Indian/Alaskan Native	1 (2.6)	-	0 (0.0)
Asian	2 (5.13)	-	7 (7.5)
Black or African American	6 (15.4)	-	9 (9.7)
White	29 (74.4)	-	77 (82.8)
Caregiver Ethnicity, n (%)			
Latinx	9 (24.3)	-	12 (12.9)
Non-Latinx	27 (73.0)	-	81 (87.1)
Caregiver Education, n (%)			
High School Graduate	10 (25.6)	-	7 (0.1)
Some College/Technical Training	-	-	15 (16.2)
Associates Degree	2 (5.13)	-	13 (14.0)
College/University Degree	10 (25.6)	-	29 (31.2)
Some Graduate Coursework	3 (7.69)	-	6 (0.1)
Graduate Degree	14 (35.9)	-	29 (31.2)

Table 7: Characteristics of the participants in data sources A, B, and C. *Developmental Delay/At Risk* indicates that the child has a developmental delay or is at risk for a developmental delay. Caregiver race, ethnicity, and education were not collected for Source 2 since those demographic items are not part of the Canadian standard.

EC	***Pediatric Habilitation Context*** 1. Take quiet activities for her to keep occupied at restaurants 2. Routines and consistency so she knows what to expect ***Pediatric Rehabilitation Context*** 1. We set up play areas specifically to suit her needs 2. Continue to explain the process of what I'm doing, while I'm doing it
SOS	***Pediatric Habilitation Context*** 1. Treat my son just as I did my daughter, with the viewpoint that he can do it all 2. Letting her explore to find out what she can and cannot do on her own ***Pediatric Rehabilitation Context*** 1. Allow child to be in charge of completing activity 2. Encourage trying new/different things
P	***Pediatric Habilitation Context*** 1. Consequences of losing the toys they don't take care of 2. Try to get him to interact by incorporating stuff he likes ***Pediatric Rehabilitation Context*** 1. We offer choices in foods/snacks- encourage her to choose from options 2. Making things fun or silly, try to create a better interest
AC	***Pediatric Habilitation Context*** 1. His brother helps him read books and play on the trampoline 2. Teaching the sounds of letters and encouraging her to mimic them ***Pediatric Rehabilitation Context*** 1. Practice activities at home with child to increase confidence/participation 2. Hand over hand tooth brushing

Table 8: Sample strategies from each class, organized according to pediatric habilitation/rehabilitation context.

Data	**Class**	**Freq.**	**Length**
CC+	EC	1390	12.93 (15.63)
	SOS	320	
	P	239	
	AC	147	
	NS	227	
CC+NS	EC	1390	11.82 (13.82)
	SOS	320	
	P	239	
	AC	147	
	NS	1000	
CC+Aug	EC	2675	11.60 (9.77)
	SOS	1394	
	P	716	
	AC	718	
	NS	1296	

Table 9: Dataset statistics, including frequencies for each strategy class in the training set and average strategy length with standard deviation provided in parentheses.