CareCorpus+ for Environment: Extended Classification and Augmentation of Caregiver Strategies to Create Enabling Environments for Young Children with Disabilities

Anonymous ACL submission

Abstract

Caregivers play an essential role in supporting the participation of children with disabilities in everyday activities. Environmental strategies, such as rearranging routines or spaces, are central to this support, yet they are scarcely documented and difficult to model systematically. To address this gap, we present 1,848 environment-focused caregiver strategies from pediatric rehabilitation contexts, manually annotated into five clinically grounded subcategories capturing activity demands, social contexts, physical environments, policies, and resources. Using this dataset, we evaluate the learnability of these categories through a multi-class classification task and examine the effects of manual and synthetic data augmentation. BERT-based models achieve improvements of up to 32% in macro-F_1, supporting the utility of the dataset for modeling caregiver strategies in low-resource rehabilitation settings.

1 Introduction

Pediatric rehabilitation aims to help children engage more actively in daily life, beyond improvements in clinical functions. *Participation* captures this real-world outcome by measuring children's attendance and involvement in everyday activities (Imms et al., 2017, 2016; World Health Organization, 2007). A major determinant of participation is the child's environment, which includes not only physical spaces, but also the people, expectations, rules, and resources that shape what children can do (Albrecht and Khetani, 2017; Di Marino et al., 2018; Khetani et al., 2018b, 2020, 2018a). Caregivers often support participation by adapting these conditions to reduce barriers, such as by rearranging furniture to create space for play, dimming lights to reduce sensory overload, or setting timers to ease transitions between activities (Kaelin et al., 2021; Jarvis et al., 2019b).

These caregiver strategies are often collected as free-text descriptions through web-based tools such as the Young Children's Participation and Environment Measure (YC-PEM) (Khetani et al., 2015) to capture practical, family-generated solutions that can be adapted to similar participation challenges. However, as these strategies grow, they become difficult to organize and search manually, especially for families awaiting rehabilitation services after hospitalization. This creates need for organization of this free-text knowledge into searchable and reusable categories for rehabilitation services.

Recent NLP work has begun to classify caregiver-generated strategies (Valizadeh et al., 2024; Farzana et al., 2024; Kaelin et al., 2021). However, existing work treats environmental strategies as a single broad category (Valizadeh et al., 2024; Farzana et al., 2024; Kaelin et al., 2021), even though "environment" can refer to very different kinds of support. For example, a child who becomes overwhelmed by noise in a classroom may need a physical or sensory adaptation, while a child who cannot get to a community activity may need a resource-based support such as transportation. Grouping both under a single environmental label makes it harder to retrieve strategies that match a child's specific participation barrier.

This limitation is partly due to resource scarcity; fine-grained examples of caregiver-generated rehabilitation strategies are difficult to collect and annotate. To address this gap, we build on the YC-PEM (Khetani et al., 2015) environmental structure, which distinguishes subdomains such as (1) demands of activities, (2) relationships/attitudes, (3) physical/sensory features, (4) weather/safety/policies, and (5) resources. These categories are clinically grounded and aligned with the World Health Organization's framework (World Health Organization, 2007, 2001; World Health Organization and The World Bank, 2011) for how environmental factors shape functioning and participation. We make the following contributions:

- We introduce **fine-grained labeling schema for environmental caregiver strategies**, applying clinically grounded subcategories to capture how caregivers modify their child's environment to support participation.
- We formulate a fine-grained classification task over these strategies to **evaluate the learnability of the proposed category structure**, providing empirical baselines for future work on caregiver strategy modeling.
- We show that both manual and synthetic augmentation of caregiver strategies improve fine-grained classification performance, **highlighting augmentation as a practical approach** for low-resource caregiver strategy modeling.

The remainder of the paper describes the dataset and annotation scheme, followed by task formulation, augmentation methods, experimental results, and discussion.

2 Related Work

Prior work has classified pediatric rehabilitation caregiver strategies into broad participation-focused categories, including *Environment/Context*, *Sense of Self*, *Preferences*, and *Activity Competence* (Farzana et al., 2024; Valizadeh et al., 2024). However, these datasets have not further separated the Environment/Context class into meaningful subcomponents. This limitation is important because caregivers most commonly describe environmental strategies when supporting children's participation across settings (Jarvis et al., 2019b; Kaelin et al., 2021; Khetani and Lucero, 2024). A finer-grained representation of this frequent strategy type can support more individualized participation-focused care planning by distinguishing what kind of environmental support is needed.

Environmental subdomains defined within YC-PEM (Khetani et al., 2015) provide a clinically grounded structure for this refinement. These subdomains map onto established ICF environmental categories (World Health Organization, 2007), supporting both clinical interpretability and mutually exclusive annotation. Our work builds on this structure to extend existing caregiver strategy resources with fine-grained Environment/Context labels.

Few existing datasets are suitable for modeling environmentally focused caregiver strategies in pediatric rehabilitation. Prior clinical NLP work has linked free-text clinical documents to ICF codes (Kukafka et al., 2006; Newman-Griffis et al., 2021), while behavioral datasets have examined child engagement from video-recorded interactions (Chorianopoulou et al., 2017). These resources are valuable, but they do not capture caregiver-generated environmental supports for children's everyday participation (Jarvis et al., 2019b; Kaelin et al., 2021; Khetani and Lucero, 2024). The closest related resource is CareCorpus+ (Farzana et al., 2024), which contains caregiver strategies collected through YC-PEM across home, daycare/preschool, and community contexts. Our dataset extends this resource by adding clinically grounded fine-grained labels to the Environment/Context subset. In this sense, our contribution is primarily a resource extension designed to support more precise modeling of caregiver strategy data.

Because caregiver-generated rehabilitation data are difficult to collect and annotate at scale, this setting remains low-resource. Data augmentation has been explored for coarse-grained strategy classification (Farzana et al., 2024), but its role in fine-grained environmental strategy modeling remains unexplored. Manual augmentation offers greater clinical control and can better preserve caregiver intent, but it is costly to scale. Synthetic augmentation is more scalable, but raises concerns about whether generated strategies preserve family voice and clinical meaning. We therefore examine both manual and synthetic augmentation as resource expansion strategies, using classification experiments to evaluate whether the resulting data support learnable fine-grained category distinctions.

3 Data Annotation

3.1 Data Source

Our data are derived from the CareCorpus+ dataset (Farzana et al., 2024), a large corpus of 3,062 caregiver-generated strategies collected through the Young Children's Participation and Environment Measure (YC-PEM) (Khetani et al., 2015). Each strategy describes an action that caregivers take (or could take) to help their child participate in daily activities across home, daycare/preschool, and community settings. Strategies were manually categorized into five clinically grounded constructs: *Environment/Context*, *Sense of Self*, *Activity Competence*, *Preferences*, and *Non-Strategy*, which align with established pediatric rehabilitation frameworks for child participation (Imms et al., 2017; Kaelin et al., 2023; Valizadeh et al., 2024).

In this study, we focus on the Environment/Context subset, the largest and most diverse class within CareCorpus+ (n=1,848). To construct a clinically interpretable resource, we extend this subset with fine-grained labels that capture distinct ways caregivers modify everyday conditions to support participation.

3.2 Annotation Process

We define a fine-grained annotation scheme for Environment/Context strategies to construct a clinically interpretable resource for caregiver strategy modeling. This refinement captures distinct ways caregivers modify a child's environment to support participation and enables automated classification of specific environmental strategy types.

Annotations were conducted by two trained undergraduate research assistants majoring in computer science and neuroscience, as supervised by a research-engaged pediatric occupational therapist and a computer science PhD student. The annotation guideline, developed under domain-expert supervision, outlined five Environment/Context subtypes, each corresponding to a label (1-5). Each strategy was assigned a single label based on its dominant functional intent:

- **1: Demands of the Activity:** Represents the physical (e.g., walking), cognitive (e.g., attention), or social (e.g., taking turns) effort required for a child to engage in an activity.
- **2: Relationships / Attitudes:** Captures strategies emphasizing the support, participation, or attitudes of other people (e.g., "Mom sings while child brushes teeth.").
- **3: Physical / Sensory:** Relates to the physical setup, layout, or sensory qualities of the environment (e.g., "Rearranged furniture to create more play space.").
- **4: Weather / Safety / Policies:** Encompasses rules, environmental conditions, and safety considerations that shape participation (e.g., "Avoid crossing busy streets.").
- **5: Resources:** Refers to tangible or intangible supports that make participation possible (e.g., "Therapist visits on Mondays.").

If a strategy did not fit into any of the above categories, annotators were told to assign the additional label, **0: Misclassified**. The annotators independently labeled a total of 1,848 strategies over seven weeks, averaging around 260 strategies per week. After each annotation round, they met with the supervising researchers, including a domain expert, to discuss and resolve any disagreements. This iterative adjudication process clarified ambiguous cases and aligned interpretation between annotators. Figure 1 summarizes the final label distribution.

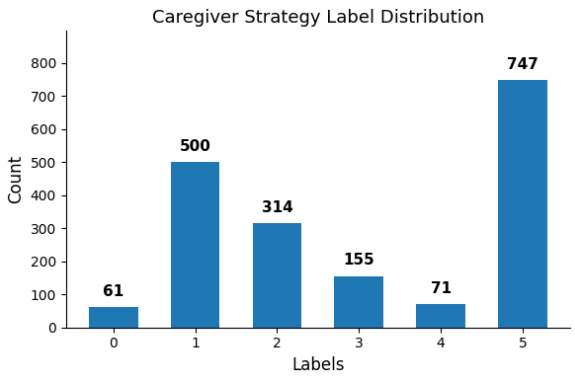

Figure 1: Distribution of annotated caregiver strategy labels. The numbers above each bar indicate the count of strategies per category. Labels correspond to: 0–Misclassified, 1–Demands of the Activity, 2–Relationships/Attitudes, 3–Physical/Sensory, 4–Weather/Safety/Policy, and 5–Resources.

Label (#)	Agreement (%)
Demands of the Activity (1)	79.5
Relationships/Attitudes (2)	83.9
Physical/Sensory (3)	70.4
Weather/Safety/Policy (4)	65.7
Resources (5)	64.8
Overall	72.8

Table 1: Inter-annotator agreement by label. Agreement (%) reflects the proportion of annotator pairs who assigned the same label. Overall Agreement was computed as the class size-weighted average across all rows.

3.3 Inter-Annotator Agreement

To ensure consistency, we computed agreement for all labels, as shown in Table 1. Agreement values ranged from 64.8% to 83.9%, with an overall agreement of 72.8%. We also computed Cohen's κ (Cohen, 1960) to complement the overall score, which yielded a substantial agreement of $\kappa = 0.72$.

3.4 Dataset Summary

Among the five annotated subtypes, *Resources* (n=747 strategies) and *Demands of the Activity* (n=500) were the most prevalent, indicating that

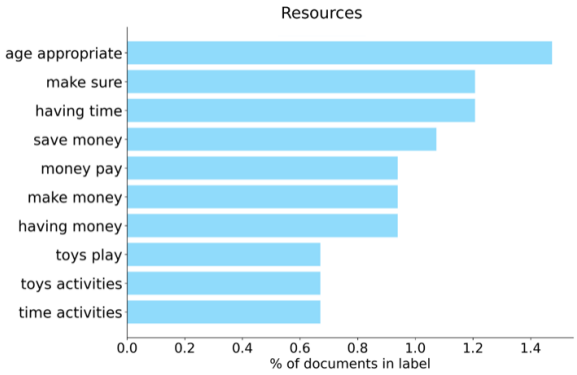

Figure 2: Most frequent bigrams in *Resources* strategies.

caregivers actively leverage tangible supports, services, or activity-related demands to facilitate participation. *Weather/Safety/Policy* strategies were relatively rare (n=71), suggesting that environmental constraints or enforced rules were less frequently emphasized. Linguistic patterns also varied across labels. For example, in *Resources* (Figure 2), frequent bigrams included "age appropriate," "make sure," and "having time," showing how caregivers secure or plan supports such as materials, time, or information. Meanwhile, *Demands of the Activity* strategies often contained verbs like "show," "practice," or "read," describing modeling and skill-building approaches. Overall, the distribution illustrates that caregivers tend to adopt concrete, resource-oriented strategies in daily routines. Additional bigrams are provided in Appendix A.

4 Data Augmentation

Because caregiver strategies are costly to elicit and annotate at scale, we examine manual and synthetic augmentation as practical strategies for expanding this low-resource dataset while preserving the meaning of caregiver-generated strategies.

4.1 Manual Data Augmentation

To investigate the feasibility of augmenting environmentally focused caregiver strategies, we developed a manual augmentation protocol that draws on commonly referenced text augmentation categories described by Wei and Zou (2019), adapted for a human-guided approach. The protocol includes four rewriting techniques: Selective Deletion (D), Selective Replacement (R), Selective Addition (A) and Word Shuffling (S). Descriptions and representative examples of each technique are provided in Table 2, with the full version available in Appendix B. The manual augmentation process was initially piloted by two individuals: one with a neuroscience background and one with a computer science background. The individuals were instructed to focus their augmentations on two underrepresented classes (*Physical/Sensory* and *Weather/Safety/Policy*) identified in Figure 1. The pilot phase sought to guide the protocol and identify how different forms of manual augmentation could be implemented in the main study.

After the pilot phase, a pediatric rehabilitation clinical expert manually augmented 1,786 strategies following the same protocol. Among the four techniques, Selective Replacement (R) was used most frequently (n=1,732; 41.9%), whereas Selective Addition (A) was used least frequently (n=118; 2.9%). These patterns were consistent in home, daycare/preschool, and community settings.

4.2 Synthetic Data Augmentation

To further explore whether large language models can effectively generate synthetic caregiver strategies, we employed GPT-5.1 (OpenAI, 2025) to rewrite training examples that mirror the human augmentation process. We implemented synthetic strategy generation through a two-step process, involving an initial strategy generation step followed by an evaluation and selection step to refine the set of strategies included in the augmented output.

4.2.1 Strategy Generation

Each caregiver strategy was rewritten using four different rewriting styles drawn from the human augmentation guideline: Selective Deletion (D), Selective Replacement (R), Selective Addition (A), and Word Shuffling (S). Each method was prompted separately to prevent instruction interference and ensure higher-quality outputs, following the idea of task decomposition by Khot et al. (2022). The prompt formats for each rewriting style are shown in Table 2. Each prompt begins with a common instruction asking the model to rewrite the original strategy according to a specified method, followed by the method's definition and a real example from the dataset to guide the model. This process was iterated for all training-set strategies (excluding label 5 which is *Misclassified/not an EC strategy*), yielding 5,716 synthetic strategies. Table 3 provides manually and synthetically generated examples for each caregiver strategy category.

Shared Instruction: Rewrite the following caregiver strategy sentence using the specified method while preserving the original meaning.	
Approach	**Augmentation Guideline**
D	Remove unnecessary words or verbs to isolate the main action or object. Example: "Rocking him while singing a song and then putting him down awake" → "Rocking while singing."
R	Replace a key adjective or verb with a natural synonym that preserves meaning. Example: "Shut off TV" → "Turn off TV."
A	Add one realistic, relevant action or detail that complements the strategy. Example: "Wheelchair accessibility" → "Ensure wheelchair accessibility."
S	Reorder clauses or phrases while keeping the same meaning. Example: "When home alone allow more freedom" → "Allow more freedom when home alone."

Table 2: Summary of the four augmentation approaches used for both manual and synthetic data generation. Approaches are **D**=Selective Deletion, **R**=Selective Replacement, **A**=Selective Addition, and **S**=Word Shuffling. The definitions were adapted from the human annotation guideline and used directly as prompt instructions for GPT-based synthetic strategy generation. See the full human annotation guideline in Appendix B.

4.2.2 Evaluation and Selection

Each synthetically rewritten version across D, R, A, and S was evaluated using three combined criteria: semantic similarity (sim), fluency (perplexity, ppl), and GPT-based self-evaluation score (gpt, 1–5 scale). This evaluation setup was designed to mirror the human selection process used during manual augmentation, where annotators selected the most appropriate rewritten strategy based on meaning preservation and naturalness. Semantic similarity measures how well the core meaning of the original caregiver strategy is preserved, while fluency, captured through perplexity, reflects the readability and naturalness of the generated output. The GPT-based self-evaluation allows the model to make its own judgment by scoring each rewritten version on a 1–5 scale using the prompt:
`Rate each version on a scale from 1 to 5 based on how well it preserves the original meaning while sounding natural and fluent.`

- `1 = Very poor (meaning changed or sentence unnatural)`
- `2 = Weak (partially preserves meaning but awkward phrasing)`
- `3 = Fair (mostly correct meaning but minor awkwardness)`
- `4 = Good (meaning well-preserved, natural phrasing)`
- `5 = Excellent (fully preserves meaning, very natural and fluent)`

Considering semantic, linguistic, and model-based self-evaluation together, our framework holistically assesses the overall quality of each rewritten strategy. We normalized all metrics to the [0,1] range and computed a weighted composite score for each version v as:

$$\text{score}_v = \alpha \cdot \text{sim}_v^{\text{norm}} + \beta \cdot (1 - \text{ppl}_v^{\text{norm}}) + \gamma \cdot \text{gpt}_v^{\text{norm}} \quad (1)$$

where $\alpha = 0.5$, $\beta = 0.3$, and $\gamma = 0.2$ represent the respective weights for semantic similarity, fluency, and GPT-based self-evaluation. We chose these weights to prioritize semantic preservation, followed by fluency, while treating GPT-based self-evaluation as a lower-weight auxiliary signal due to potential bias. The version with the highest composite score across D, R, A, and S was selected as the final synthetic rewrite, thereby narrowing the synthetic examples to 1,429.

Across all synthetic strategies generated, R was the most frequently chosen method (909 strategies), followed by D (396), A (119), and S (5). Based on the combined evaluation criteria, replacement-based rewrites appeared to have the best balance of meaning preservation and fluency within our framework. This trend is broadly similar to the distribution observed in the human-augmented dataset, where R and D were most commonly selected.

Category	Strategy
Demands of the Activity	Original: "Show how to do the activity and allow him to copy." → Manual: "Rocking while singing." (*Deletion*) → Synthetic: "Demonstrate how to do the activity and permit him to copy." (*Replacement*)
Relationships/Attitudes	Original: "Sing song together." → Manual: "Chant songs together." (*Replacement*) → Synthetic: "Sing songs together when needed." (*Addition*)
Physical/Sensory	Original: "Make easy for him to access his toys/clothes." → Manual: "Easy access to toys/clothes." (*Deletion*) → Synthetic: "Create easy for him to access his toys/clothes." (*Replacement*)
Weather/Safety/Policy	Original: "Control the amount of TV or tablet time to make time for these other activities." → Manual: "Control the TV or tablet time for other activities." (*Deletion*) → Synthetic: "Control the amount of TV or tablet time to create time for these other activities." (*Replacement*)
Resources	Original: "Exercise programs - for young children to help them keep their senses in balance." → Manual: "Help young children keep their senses in balance through exercise programs." (*Substitution*) → Synthetic: "Exercise programs - for young children to help them keep their senses in balance to support development." (*Addition*)

Table 3: Examples of manual and synthetic augmentations for each caregiver strategy category. Each rewrite is followed by its corresponding augmentation type (e.g., Deletion, Replacement, Addition, or Substitution).

5 Strategy Classification

We use automated caregiver strategy classification to evaluate whether the proposed fine-grained labels are learnable from caregiver strategy text and to establish initial benchmarks for this dataset. We frame this as a multi-class text classification problem, where the goal is to predict the correct strategy label for each caregiver strategy.

5.1 Experimental Setup

We framed our task as a multi-class text classification problem, where the goal was to predict the correct strategy label for each caregiver strategy. Each strategy was assigned a single label corresponding to its primary functional intent, consistent with the mutually exclusive annotation scheme described in Section 3.2. The dataset was divided into training (80%) and test (20%) sets using a stratified split to preserve the original class distribution. Model performance was evaluated on the fixed test set using macro-averaged accuracy, precision, recall, and F_1 scores. We evaluate a range of standard baselines and pretrained models to establish initial empirical benchmarks for this dataset and assess the learnability of the fine-grained category structure:

- **Most Frequent Class:** A simple baseline that predicts the most frequent label observed in the training data. This serves as a performance floor that reflects class imbalance.
- **Random:** A stochastic baseline that assigns labels uniformly at random across classes, representing performance without learned data.
- **Naive Bayes:** A probabilistic model based on word frequency distributions with a multinomial likelihood, assuming conditional independence among features given a label.
- **Logistic Regression:** A linear classifier trained on TF–IDF feature representations of the input text, modeling the probability of each class through a softmax layer.
- **BERT:** A transformer-based model (Devlin et al., 2019) fine-tuned on the training data to learn contextual embeddings of the input text. We used *bert-base-uncased* from HuggingFace[1] with a linear classification head, trained for three epochs with a batch size of 16, a learning rate of $2e{-}5$.
- **BioClinicalBERT:** A domain-adapted variant of BERT pre-trained on biomedical and clinical corpora (Alsentzer et al., 2019). The

[1] https://huggingface.co/docs/transformers

model was fine-tuned on the training data under the same setup as BERT to examine whether domain-specific pre-training improves contextual understanding of the input caregiver strategy text.

- **LLaMA:** A large instruction-tuned language model (Touvron et al., 2023) evaluated without task-specific fine-tuning. We loaded Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024) in 4-bit quantized format and tested zero-shot and few-shot prompting conditions (See Appendix C for full prompting details).

We compared strategy classification performance across baselines and transformer-based models under three data configurations: the original annotated dataset, the original dataset combined with manually augmented strategies, and the original dataset combined with synthetically generated strategies. This setup allowed us to examine the relative effects of expert-authored and automatically generated augmentation on classification performance. Manual augmentation represents a clinically grounded approach to dataset expansion, whereas synthetic augmentation provides a scalable alternative for low-resource settings where expert annotation is costly. All fine-tuning experiments were conducted on a single NVIDIA A100 GPU.

5.2 Results

Model performance across all data configurations is summarized in Table 4. Results for BERT and BioClinicalBERT are averaged across three runs, and macro-averaged metrics are reported to account for class imbalance across strategy types.

Overall, BERT consistently performed the best across all datasets, aligning with prior findings from Farzana et al. (2024) in coarser-grained tasks. Its performance further improved after augmentation, reaching the best overall performance of F_1=0.863 on the *Original + Manual Augmentation* dataset. BioClinicalBERT followed closely, particularly after augmentation (F_1=0.739 on the *Original + Synthetic Augmentation* dataset). Logistic Regression and Naive Bayes achieved moderate performance, ranging from 0.422 to 0.668 F_1, while the Most Frequent and Random baselines remained lowest as expected. Prompting LLaMA performed comparably to Logistic Regression and Naive Bayes, with a zero-shot F_1=0.489 and a slightly improved few-shot F_1=0.591.

Augmenting the data generally improved classification performance, especially for transformer-based models. After manual augmentation (*Original + Manual Aug*), the top-performing model, BERT, improved from an F_1 score of 0.656 to 0.863, an increase of approximately 32%. BioClinicalBERT exhibited a similar trend (0.632 $\rightarrow$ 0.830, an increase of 31%), showing that domain-specific pretraining also benefits from additional data. Naive Bayes and Logistic Regression achieved modest gains (roughly 10–15%).

After synthetic augmentation (*Original + Synthetic Aug*), Naive Bayes showed a minor decrease (0.424 $\rightarrow$ 0.422 in F_1) compared to the *Original* dataset, but all other models improved in a pattern consistent with manual augmentation. Although the overall gains were smaller in *Original + Synthetic Aug*, BERT and BioClinicalBERT both reached an F_1 score of 0.739, which is approximately 13%-17% higher than their original performance. These results suggest that while synthetically-generated rewrites do not yet match the quality of human-generated ones, they meaningfully enhance model robustness.

6 Error Analysis

We conducted a qualitative error analysis on BERT fine-tuned on the *Original + Synthetic Augmentation* dataset, the best-performing GPT-augmented model, to identify where synthetic data contributed to misclassification and how future augmentation could be refined. The analysis was performed on the fixed test set (358 samples), of which 70 (approximately 19.5%) were misclassified.

Most misclassifications occurred between *Demands of the activity* $\leftrightarrow$ *Resources* (25 cases) and *Demands of the activity* $\leftrightarrow$ *Relationships/attitudes* (10 cases). Many of these cases were also challenging for human annotators to separate during the initial data annotation phase. For instance, "Watching a YouTube video or reading a book on the topic" was labeled *Demands of the activity* but predicted as *Resources*, likely because of its references to tangible objects (e.g., video and book) that resemble strategies in *Resources*. During manual annotation, this example was discussed and labeled *Demands of the activity* since the emphasis was placed slightly more on the action than the material itself. Conversely, "Once something doesn't work, try a new strategy" was labeled *Resources* but predicted as *Demands of the activity*, as the model prioritized

Data	Model	Accuracy	Precision	Recall	F_1
Original	Most Frequent	0.419	0.084	0.200	0.118
	Random	0.182	0.184	0.200	0.164
	Naive Bayes	0.616	0.855	0.397	0.424
	Logistic Regression	0.654	0.817	0.465	0.508
	BERT	0.743	0.737	**0.626**	**0.656**
	BioClinicalBERT	**0.760**	**0.792**	0.609	0.632
	LLaMA (zero-shot)	0.528	0.548	0.511	0.489
	LLaMA (few-shot)	0.643	0.598	0.597	0.591
Original + Manual Augmentation	Most Frequent	0.418	0.084	0.200	0.118
	Random	0.206	0.205	0.210	0.184
	Naive Bayes	0.700	0.873	0.485	0.529
	Logistic Regression	0.763	0.866	0.607	0.668
	BERT	**0.894**	**0.876**	**0.853**	**0.863**
	BioClinicalBERT	0.874	0.849	0.815	0.830
Original + Synthetic Augmentation	Most Frequent	0.419	0.084	0.200	0.118
	Random	0.201	0.216	0.221	0.183
	Naive Bayes	0.617	0.848	0.395	0.422
	Logistic Regression	0.668	0.829	0.478	0.522
	BERT	0.791	0.751	**0.728**	**0.739**
	BioClinicalBERT	**0.802**	**0.769**	0.717	**0.739**

Table 4: Performance of models on caregiver strategy classification with different training data variants. “Original” is the original, non-augmented training set; “Manual Augmentation” adds manually paraphrased examples; and “Synthetic Augmentation” includes synthetically generated examples. All metrics are macro-averaged to account for class imbalance. LLaMA models were evaluated only on the original dataset, without augmented training, as they serve as large pre-trained baselines rather than fine-tuned models to maintain the same evaluation setup.

the verb “try” (the action) rather than interpreting the “strategy” as the main resource. These borderline cases reveal the need for explicit tie-breaking criteria in the annotation protocol and for targeted inclusion of ambiguous examples in data augmentation, to help models learn finer distinctions between potentially overlapping strategy categories.

Beyond category misclassifications, we observed that manual augmentations generally remained concise (e.g., "Chant songs together"), while synthetic augmentations were often longer with additional contextual details (e.g., "Sing songs together when needed"), increasing the likelihood of boundary-crossing predictions in ambiguous cases.

6.1 Discussion

In our experiments, both manual and synthetic augmentation improved performance, with BERT and BioClinicalBERT achieving the strongest results across data settings. This suggests that transformer-based models can learn the fine-grained distinctions well, and that augmentation can help address the scarcity of caregiver-generated rehabilitation data. Qualitative inspection revealed that synthetic rewrites can differ from manual ones in style and specificity. For example, synthetic strategies often adopted more directive, action-oriented language, such as *ensure*, which may have helped BioClinicalBERT correctly classify some procedural strategies (e.g., “Ensure wheelchair access”).

Automatic classification and expansion of environmental caregiver strategies may support large collections of caregiver-generated rehabilitation knowledge, such as free-text rehabilitation notes or goal-setting applications with strategy banks. Human-generated rewrites could be leveraged to guide and refine synthetic generation, enabling scalable and cost-efficient augmentation pipelines for future pediatric rehabilitation applications.

7 Conclusion

Overall, this work contributes a clinically grounded fine-grained dataset extension and initial modeling benchmarks for environmental caregiver strategies in pediatric rehabilitation. We extend 1,848 caregiver strategies with YC-PEM-based environmental subcategories and examine manual and synthetic augmentation for resource expansion. Classification results show that these strategies are learnable and that augmentation improves performance, supporting future NLP tools for organizing caregiver-generated rehabilitation knowledge.

Limitations

While this work demonstrates the feasibility of augmenting and classifying caregiver strategies, several limitations remain. First, our evaluation was conducted on a relatively small and domain-specific dataset, which may limit generalizability to broader healthcare contexts. Second, although both manual and synthetic augmentation improved performance, the quality of synthetic rewrites was sometimes less robust or natural-sounding compared to human-generated ones. Future work could incorporate human-in-the-loop evaluation and explore more prompting techniques to improve the quality of synthetic data.

Ethical Considerations

This study used de-identified caregiver strategy data from the CareCorpus+ dataset, which was originally collected under institutional ethics approval as described in prior work (Khetani et al., 2015; Kaelin et al., 2022; Khetani et al., 2023; Rizk et al., 2023; Choong et al., 2018; Jarvis et al., 2019a; Khetani et al., 2018a). All participants in the original studies provided informed consent prior to participation. No identifiable personal or health information was used in this research, and all analyses were conducted solely for research purposes to support the development of tools that assist caregivers in browsing and organizing strategies more efficiently. All annotators involved in providing fine-grained Environment/Context strategy labels and manual strategy augmentations completed institutional ethics training. They were compensated for their annotation and augmentation work through research assistantships paid at institutionally-defined rates set higher than local minimum wage.

References

Erin C Albrecht and Mary A Khetani. 2017. Environmental impact on young children's participation in home-based activities. *Developmental Medicine & Child Neurology*, 59(4):388–394.

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.

Karen Choong, Douglas Fraser, Samah Al-Harbi, Asm Borham, Jill Cameron, Saoirse Cameron, Ji Cheng, Heather Clark, Tim Doherty, Nora Fayed, and 1 others. 2018. Functional recovery in critically ill children, the "weecover" multicenter study. *Pediatric Critical Care Medicine*, 19(2):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 *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5055–5059, New Orleans, USA. IEEE.

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

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.

Erica Di Marino, Stephanie Tremblay, Mary Khetani, and Dana Anaby. 2018. The effect of child, family and environmental factors on the participation of young children with disabilities. *Disability and health journal*, 11(1):36–42.

Shahla Farzana, Ivana Lucero, Vivian Villegas, Vera C Kaelin, Mary Khetani, and Natalie Parde. 2024. Carecorpus+: expanding and augmenting caregiver strategy data to support pediatric rehabilitation. In *The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), Miami, Florida, USA, November 12-16, 2024.*, pages 6912–6927. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Christine Imms, Brooke Adair, Deb Keen, Anna Ullenhag, Peter Rosenbaum, and Mats Granlund. 2016. 'participation': a systematic review of language, definitions, and constructs used in intervention research with children with disabilities. *Developmental medicine and child neurology*.

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. 2019a. 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, Anupama R. Gurga, Hyunji Lim, Jessica Cameron, Jan Willem Gorter, Karen Choong, and Mary A. Khetani. 2019b. Caregiver strategy use to promote children's home participation after pediatric critical illness. *Archives of Physical Medicine and Rehabilitation*, 100(11):2144–2150.

Vera Kaelin, Vivian Villegas, Yi-Fan Chen, Natalie Murphy, Elizabeth Papautsky, Jodi Litfin, Natalie Leland, Varun Maheshwari, Beth McManus, and Mary Khetani. 2022. 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, Vivian C. Villegas, Christine Imms, and Mary A. Khetani. 2021. Participation-focused strategy use among caregivers of children receiving early intervention. *American Journal of Occupational Therapy*, 75:7501205090.

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.

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 A Khetani, Erin C Albrecht, Jessica M Jarvis, David Pogorzelski, Emmy Cheng, and Karen Choong. 2018a. Determinants of change in home participation among critically ill children. *Developmental Medicine & Child Neurology*, 60(8):793–800.

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.

Mary A. Khetani and Isabella Lucero. 2024. Classifying children's participation and its family of related concepts through the participation and environment measure (pem) approach. In Melek and Aslam, editors, *ICF Turkey Conference Proceedings*, page 47. Atılım University.

Mary A Khetani, Beth M McManus, Erin C Albrecht, Vera C Kaelin, Jodi K Dooling-Litfin, Elizabeth A Scully, and High Value Early Intervention Research Group. 2020. Early intervention service intensity and young children's home participation. *BMC pediatrics*, 20(1):330.

Mary A Khetani, Beth M McManus, Kristen Arestad, Zachary Richardson, Renee Charlifue-Smith, Cordelia Rosenberg, and Briana Rigau. 2018b. Technology-based functional assessment in early childhood intervention: a pilot study. *Pilot and feasibility studies*, 4(1):65.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Rita Kukafka, Michael E. Bales, Ann Burkhardt, and Carol Friedman. 2006. Human and automated coding of rehabilitation discharge summaries according to the international classification of functioning, disability, and health. *Journal of the American Medical Informatics Association*, 13(5):508–515.

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.

OpenAI. 2025. Gpt-5.1. `https://openai.com/index/gpt-5-1/`. Large language model with enhanced instruction-following and conversational capabilities.

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 1 others. 2023. Implementing an electronic patient-reported outcome and decision support tool in early intervention. *Applied Clinical Informatics*, 14(01):091–107.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Mina Valizadeh, Vera C Kaelin, Mary A Khetani, and Natalie Parde. 2024. Carecorpus: a corpus of real-world solution-focused caregiver strategies for personalized pediatric rehabilitation service design. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, May 20-25, 2024*, pages 2871–2882. ELRA Language Resource Association.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *EMNLP*.

World Health Organization. 2001. *International Classification of Functioning, Disability and Health: ICF*. World Health Organization, Geneva.

World Health Organization. 2007. *International classification of functioning, disability and health: children and youth version: ICF-CY.* World Health Organization.

World Health Organization and The World Bank. 2011. *World Report on Disability*. World Health Organization, Geneva.

A Frequent Bigrams by Strategy Type

In Figure 6, we show the most frequent bigrams for each fine-grained strategy type.

B Manual Augmentation Samples

In the following tables, we show manual augmentation samples for each augmentation technique, for each fine-grained caregiver strategy type.

C Prompt Templates for LLaMA Baseline

We evaluated Meta-Llama-3.1-8B-Instruct in inference-only mode using both zero-shot and few-shot prompting. The prompts were designed to enforce single-label classification across the five environmental strategy categories, consistent with our annotations. For few-shot prompting, two representative examples per class were automatically selected based on their proximity to the class centroid in TF–IDF space (10 examples total). The model was not fine-tuned on augmented datasets to keep the evaluation consistent.

C.1 Zero-shot Prompt

The following template was used for zero-shot classification:

> Classify the caregiver strategy into one label. Answer with the label number only.
>
> Labels:
> 0 = Activity Demands
> 1 = Social Context
> 2 = Physical Environment
> 3 = Policies
> 4 = Resources
>
> If multiple cues appear, break ties with these priorities:
>
> If it is a rule, policy, safety condition, or weather-related constraint → 3
> Else if it changes environment layout, sensory features, or accessibility → 2
> Else if it brings or accesses equipment, services, time, or location → 4
> Else if it emphasizes another person's participation or support → 1
> Else → 0
>
> Strategy: [TEXT]

Tie-breaking rules were included to ensure mutually exclusive outputs aligned with the single-label annotation framework.

C.2 Few-shot Prompting

For few-shot prompting, two representative examples per category (one prototypical example and one boundary case) were automatically selected using TF–IDF centroid similarity within each class. This resulted in a total of 10 demonstration examples included in the prompt context prior to classification.

The same label definitions and tie-breaking rules used in the zero-shot setting were retained for few-shot prompting.

Figure 3: Top 10 most frequent bigrams in Demands of the Activity strategies.

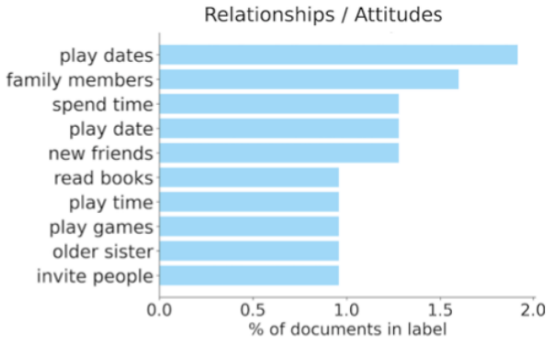

Figure 4: Top 10 most frequent bigrams in Relationships/Attitudes strategies.

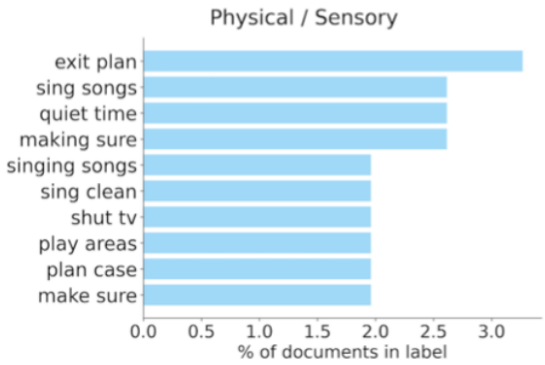

Figure 5: Top 10 most frequent bigrams in Physical/Sensory strategies.

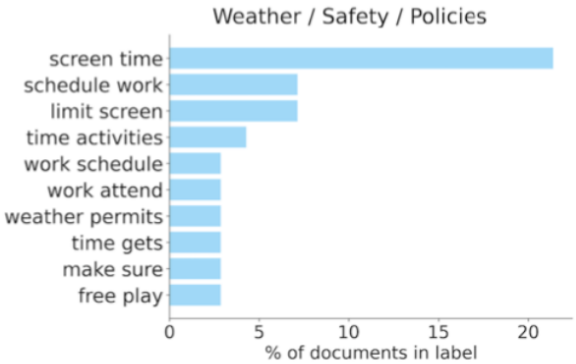

Figure 6: Top 10 most frequent bigrams in Weather/Safety/Policies strategies.

1 = Activity Demands		
Setting	**Technique**	**Example**
Home	Original	"Routines and consistency so she knows what to expect"
	D	→ "Routines so she knows what to expect"
	R	→ "Routines to help her predict what will happen"
	A	→ "Set routines so she knows what to expect"
	S	→ "So she knows what to expect, routines and consistency "
Daycare/Preschool	Original	"Preparing him for what he has to do and what is going to happen"
	D	→ "Helping him get ready for what's coming"
	R	→ "Preparing him for what to do and what will happen"
	A	→ X
	S	→ "Preparing him for what is going to happen and what he has to do"
Community	D	→ "Short event length so she doesn't fatigue "
	R	→ "Reduce the event duration to avoid fatigue"
	A	→ X
	S	→ "So she doesn't fatigue, keep the length of the event short"

Appendix Table : Manually generated examples for Class 1 (Activity Demands).

2=Relationship/Attitudes		
Setting	**Technique**	**Example**
Home	Original	"Grandparents play with him and talk to him"
	D	→ "Grandparents play and talk with him"
	R	→ "Grandparents interact and speak with him"
	A	→ "Have grandparents play and talk with him"
	S	→ "He plays with grandparents and talks with them"
Daycare/Preschool	Original	"Program is geared for interaction with other children"
	D	→ "Program geared for interaction"
	R	→ "Program structured for interaction"
	A	→ "Provide program geared for interaction"
	S	→ X
Community	Original	"Lots of physical support, holding, affection helps"
	D	→ "Physical support, holding, and affection"
	R	→ "Physical assistance, cuddling, and warmth"
	A	→ "Provide physical support, holding, and affection"
	S	→ "Holding, affection, and physical support"

Appendix Table : Manually generated examples for Classes 1-2. (Continued)

3=Physical/Sensory		
Setting	**Technique**	**Example**
Home	Original	"Breaks for quiet time away from the hubbub of crowds and noise"
	D	→ "Breaks for quiet time"
	R	→ "Breaks to calm down "
	A	→ "Have breaks for quiet time"
	S	→ "Breaks away from the hubbub of crowds and noise for quiet time"
Daycare/Preschool	Original	"Keeping him in the front, where he is less likely to be distracted"
	D	→ "Keeping him in the front for less distraction"
	R	→ "Seating him in the front to reduce distraction"
	A	→ X
	S	→ "Where he is less likely to be distracted, keep him in the front"
Community	Original	"Space to remove himself to if he becomes overwhelmed"
	D	→ "Space to remove himself if overwhelmed"
	R	→ "Space to take a break if overwhelmed"
	A	→ "Provide space for him to remove himself if overwhelmed"
	S	→ "If he becomes overwhelmed, remove him to space"

Appendix Table : Manually generated examples for Class 2 (Physical/Sensory).

4=Weather/Safety/Policy		
Setting	**Technique**	**Example**
Home	Original	"Started limiting screen time, though harder to do when weather is bad"
	D	→ "Limit screen time"
	R	→ "Restrict screen time"
	A	→ X
	S	→ "Though harder to do when weather is bad, started limiting screen time"
Daycare/Preschool	Original	"Make sure work schedule allows transporting children to activities"
	D	→ "Make sure work schedule allows transporting children"
	R	→ "Ensure work hour enables bringing children to activities"
	A	→ X
	S	→ "To transport children to activities, make sure your work schedule allows it"
Community	Original	"I also take her to parks for the same reason when weather permits"
	D	→ "Take her to parks when weather permits"
	R	→ "Visit parks with her when the weather allows"
	A	→ X
	S	→ "When weather permits, I also take her to parks for the same reason"

Appendix Table : Manually generated examples for Classes 3-4.(Continued)

5=Resources		
Setting	**Technique**	**Example**
Home	Original	"PEC system to help with steps for the routine we are trying to do"
	D	→ "PEC system to help with steps for the routine "
	R	→ "PEC system to support steps in the daily routine"
	A	→ "Have PEC system to help with steps for the routine"
	S	→ "Help with steps for the routine we are trying to do through the PEC system"
Daycare/Preschool	Original	"Specialized seating to allow [name] to sit/work with peers"
	D	→ "Specialized seating to work with peers"
	R	→ "Adaptive seating to work with peers"
	A	→ "Assign specialized seating to work with peers"
	S	→ "Allow [name] to sit/work with peers through specialized seating"
Community	Original	"Pictures of things to buy so she can shop"
	D	→ "Pictures of things to buy"
	R	→ "Images of products to purchase"
	A	→ "Have pictures of things to buy"
	S	→ "So she can shop, have pictures of things to buy"

Appendix Table : Manually generated examples for Class 5.