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

Shahla Farzana[1], Ivana Lucero[1], Vivian Villegas[1], Vera C. Kaelin[1,2],
Mary A. Khetani[1], Natalie Parde[1]
University of Illinois Chicago[1], Umea University[2]
{sfarza3, ilucer3, vvilleg2, mkhetani, parde}@uic.edu, vera.kaelin@umu.se

Introduction

- Caregivers' strategies data are helpful to designing meaningful pediatric rehabilitation for the >50 million young children experiencing disability worldwide[1,2], but manually classifying caregiver strategies when documented in free text form is not scalable.
- Prior work to establish benchmarks for their automated classification were constrained by smaller, homogenous, and imbalanced data sources[3,4].
- We introduce CareCorpus+ as a larger and more balanced deidentified data source with 3,062 caregiver strategies and non-strategies for young children, across a broader age range and diverse rehabilitation contexts.
- We use CareCorpus+: 1) to examine the reproducibility and generalizability of prior findings, and 2) to propose novel data augmentation techniques to generate and filter caregiver strategies, enabling inclusion of synthetic data to strengthen model performance.

CareCorpus+ Dataset

Data Collection

Caregiver Strategies and Non-Strategies			Non-Strategies
Dataset A[5]: 93 caregivers of children with developmental disability/delay, aged 0-5 years, accessing rehabilitation.	Dataset B[6]: 39 caregivers of children, aged 0-3 years, enrolled in early intervention for rehabilitation.	Dataset C[7]: 53 caregivers of critically ill children, aged 0-4 years, from hospital until 6 months post-discharge.	Public health forums: Caregivers of children with reported health issues, aged 0-5 years

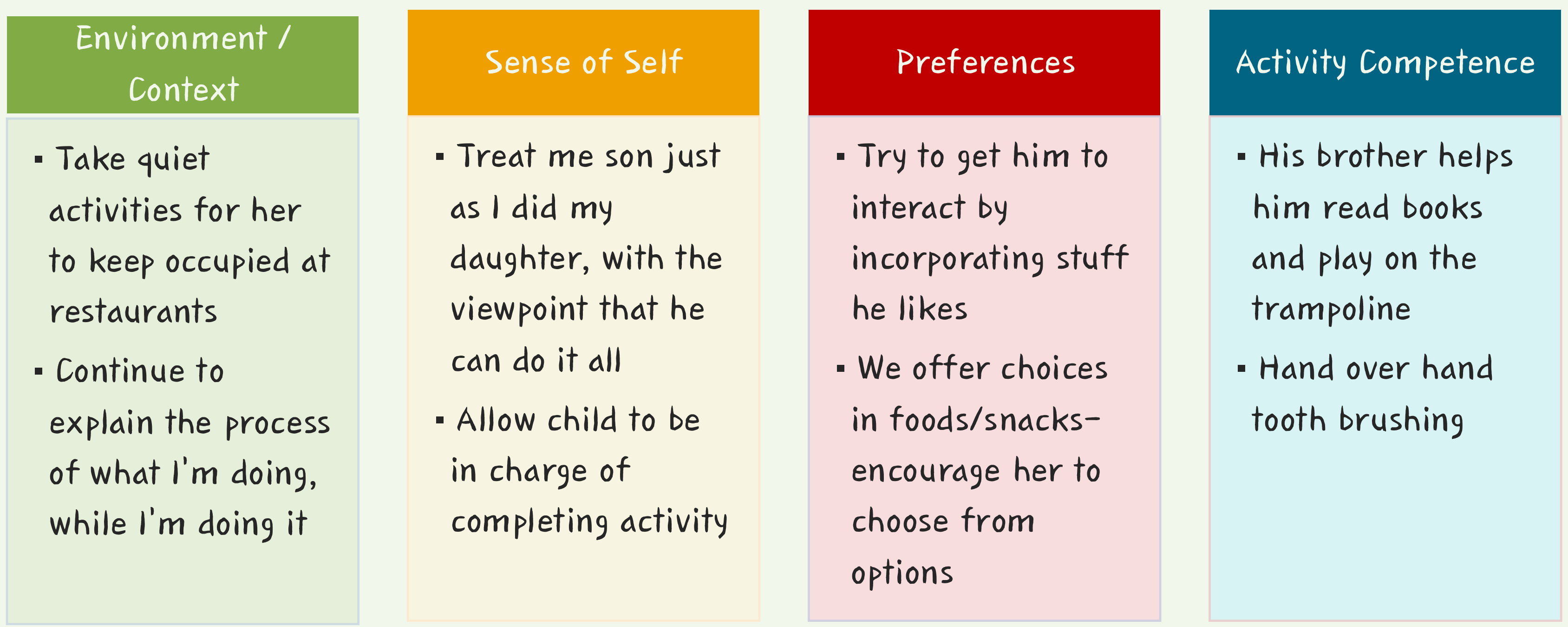
Table 1. Data included in CareCorpus+

Data Annotation

- Two trained annotators independently annotated 50-250 strategies per week (March-August 2023).
- Annotators met with an adjudicator to settle discrepancies, seeking additional feedback from other key informants as needed.

Class	% Agreement	κ
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

Table 2. Per-class inter-annotator agreement



Environment / Context
- Take quiet activities for her to keep occupied at restaurants
- Continue to explain the process of what I'm doing, while I'm doing it

Sense of Self
- Treat me son just as I did my daughter, with the viewpoint that he can do it all
- Allow child to be in charge of completing activity

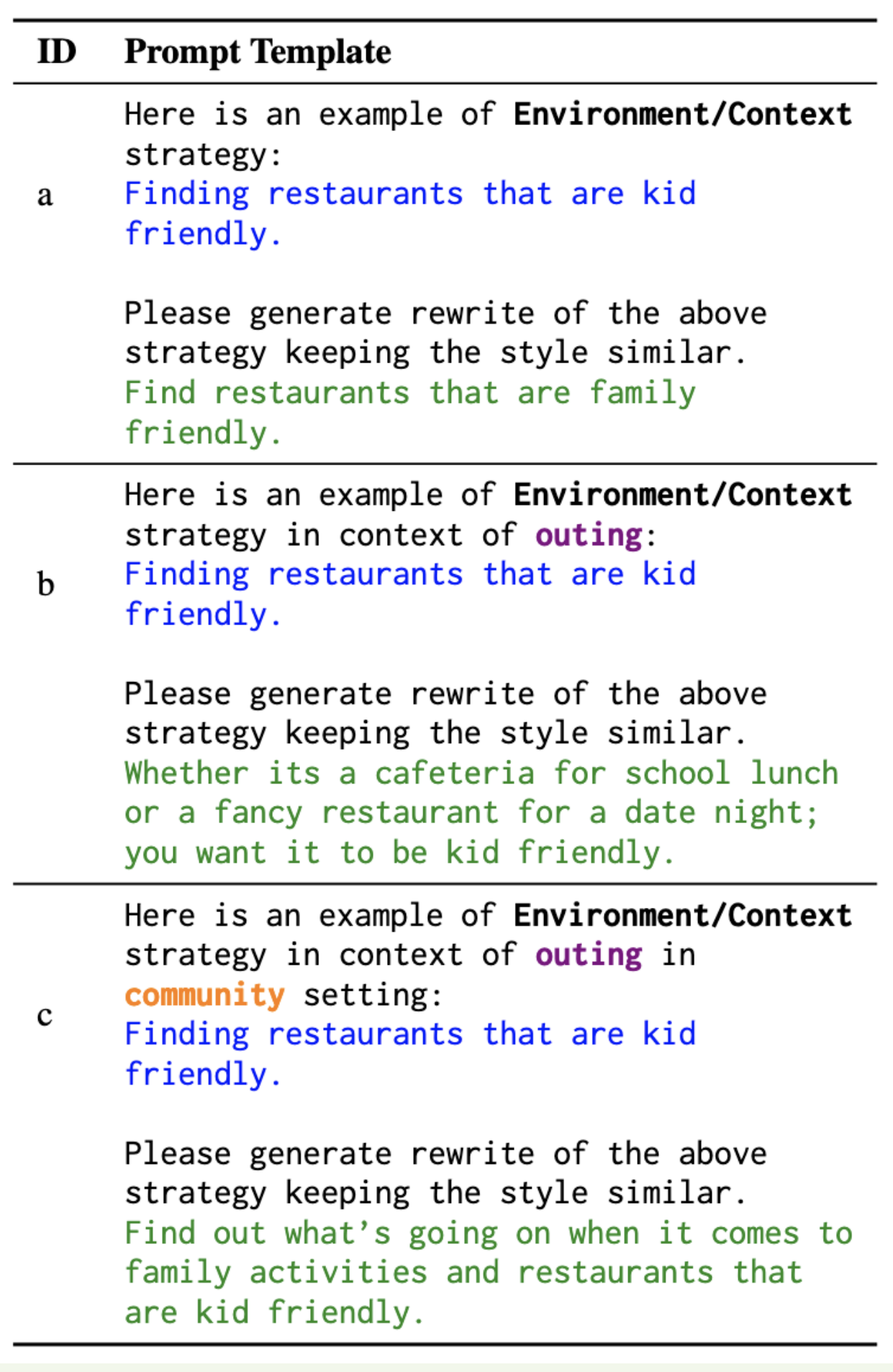
Preferences
- Try to get him to interact by incorporating stuff he likes
- We offer choices in foods/snacks- encourage her to choose from options

Activity Competence
- His brother helps him read books and play on the trampoline
- Hand over hand tooth brushing

Figure 1. Sample strategies per class

Data Augmentation

- Prompt-based strategy generation using Flan-t5-xl[8] with PVI filtering[9].
- Strategy augmentation was framed as a paraphrase task.
- Three prompt components: 1) class name, 2) broader activity context, and 3) setting.



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 3. Sample prompts to generate synthetic strategies

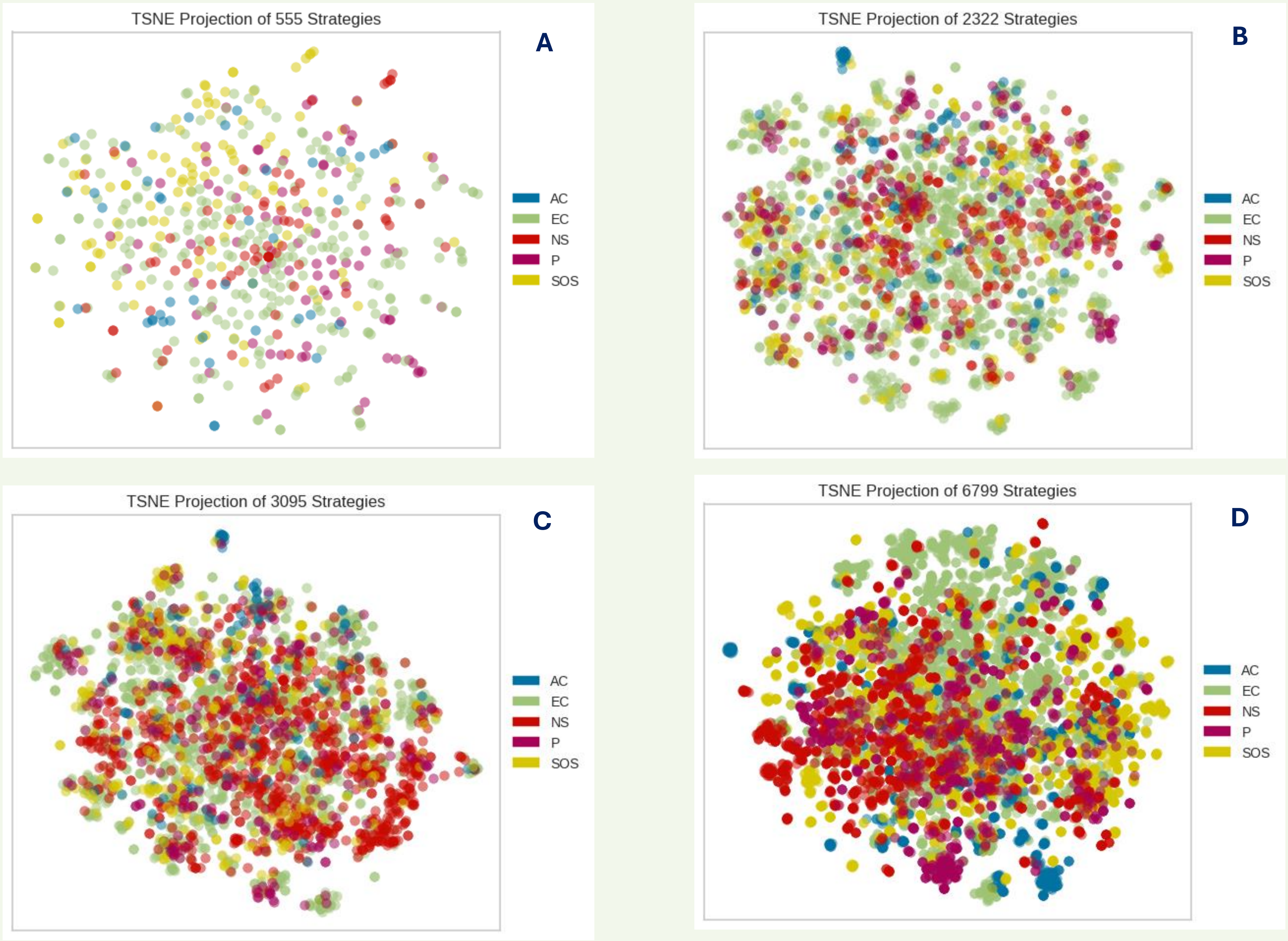


Figure 2. Visualizations of strategies by class and across four datasets: CareCorpus (A), CareCorpus+ (B), CareCorpus+NoStrategies (C), and CareCorpus+Augmentation

Dataset	Model	Acc.	F1
CC	LR	57.89	0.46
	BERT	**64.47**	**0.56**
	Bio	53.94	0.39
CC+	LR	**74.48**	**0.57**
	BERT	60.78 (0.02)	0.53 (0.01)
	Bio	48.74 (0.04)	0.44 (0.03)
CC+NS	LR	**75.26**	0.62
	BERT	72.77 (0.01)	**0.65 (0.01)**
	Bio	54.46 (0.05)	0.48 (0.04)
CC+Aug	LR	82.55	0.75
	BERT	**83.56 (0.01)**	**0.80 (0.00)**
	Bio	80.48 (0.01)	0.76 (0.01)

Dataset	Model	Acc.	F1
CC	S/NS	90.60	0.87
	ES/IS	58.06	0.53
CC+	S/NS	90.60 (0.02)	0.87 (0.00)
	ES/IS	84.97 (0.02)	0.83 (0.01)
CC+NS	S/NS	**95.02 (0.00)**	0.93 (0.00)
	ES/IS	–	–
CC+Aug	S/NS	91.78 (0.00)	0.89 (0.00)
	ES/IS	**92.18 (0.00)**	**0.91 (0.00)**

Table 3. Performance in a five-class setting and model comparison for pipelined classification tasks

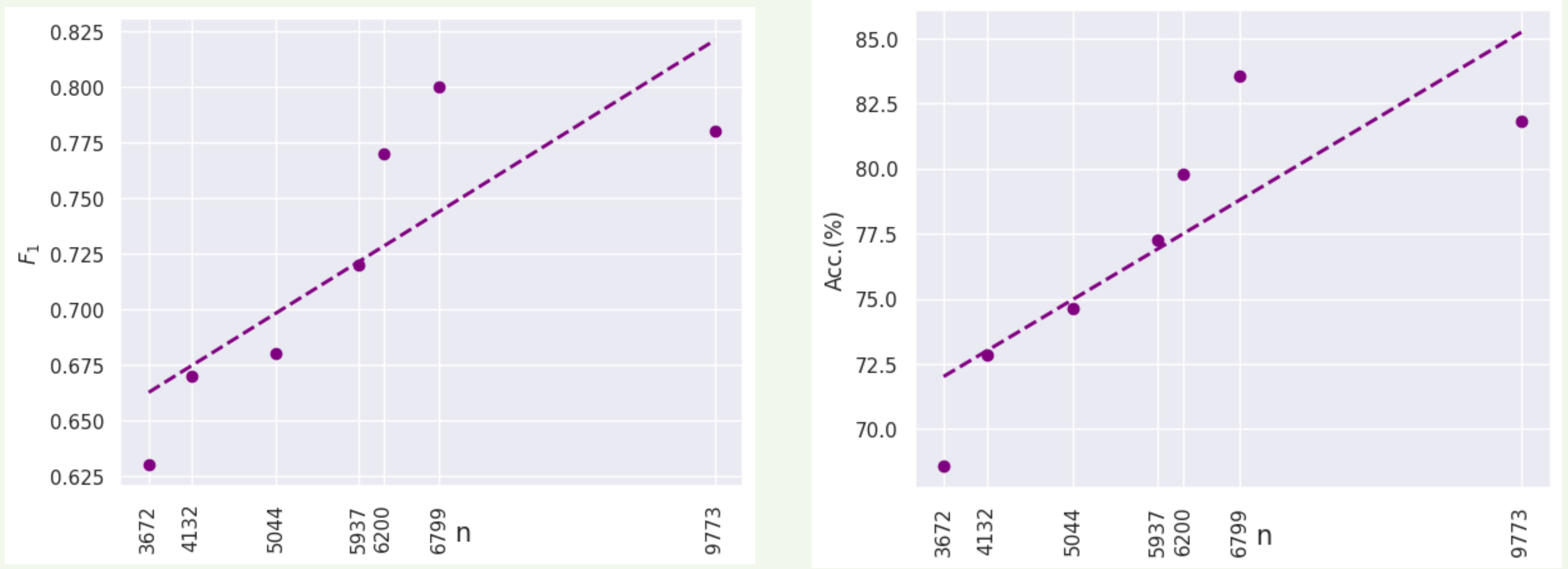


Figure 3. Performance variation with varying training instances

Discussion

- We demonstrate the value of manually curated strategies when paired with publicly available task-relevant non-strategies and a novel data augmentation approach, for replicating prior findings[3,4] and improving model performance.
  - Publicly available non-strategies support improved performance for strategy classification (22.6% relative increase in F_1)
  - Prompt-based synthetic data expansion improves model performance (50.9% relative increase in F_1).
- Results suggest inclusion of automated classification and new directions for clinically relevant and ethical applications[10] (e.g., initiating caregiver education when detecting non-strategy responses and using LLMs to consolidate strategies of similar type).