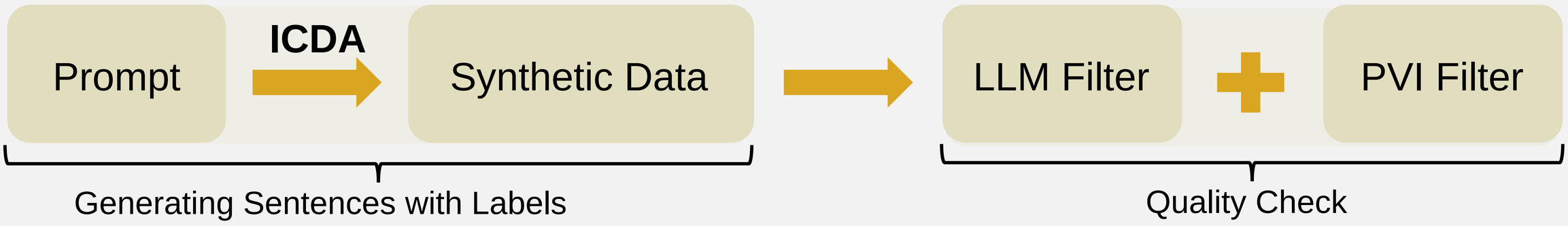
In-Context Data Augmentation for Slot Filling

Yun-Fan Chen, Yen-Ting Lin, Yun-Nung (Vivian) Chen



Idea

- LLMs augment scarce data
- LLM and PVI control data quality

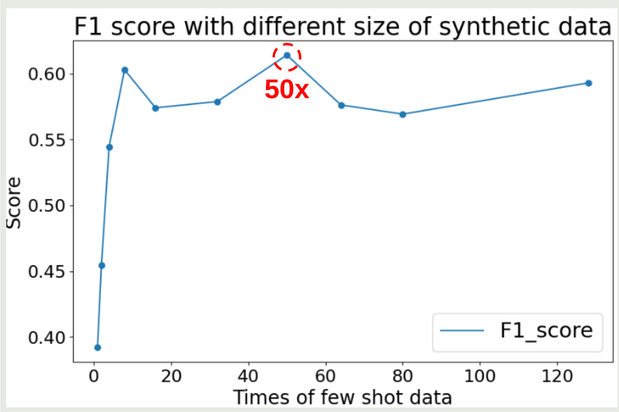
→ smaller model, lower cost, better controllability

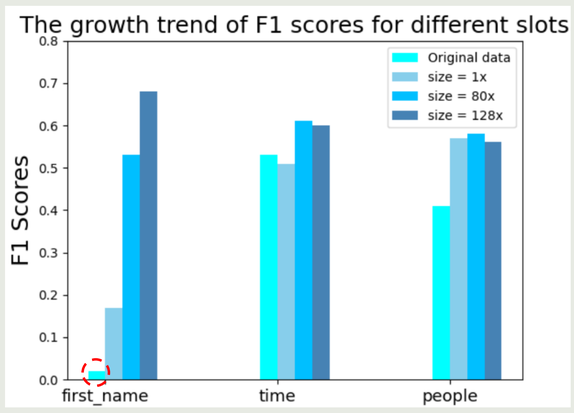
ICDA: In-Context Data Augmentation

Prompt Template:
Generate examples include some of these slot: ['people', 'date', 'time', 'first_name', 'last_name']. Here is an example sentence:
Example1: There will be <people> 5 adults and 1 child </people>.
Example2: We will require and outside table to seat <people> 9 people </people> on <date> August 23rd </date>.
:

Output (Synthetic Data):
We would like to booked the table for <time> 8 pm </time> on <date> Saturday, February 12 </date>.

Results on Different Synthetic Data Sizes





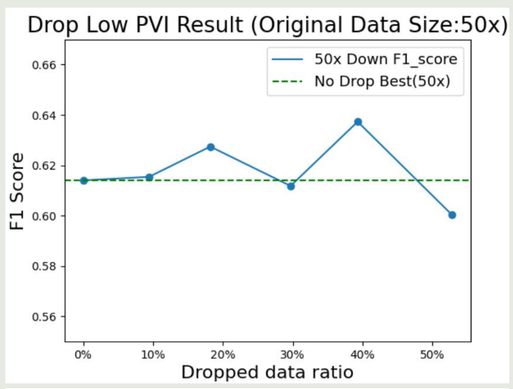
- ✓ 50x data performs best
- ✓ Slots with poor scores are improved most

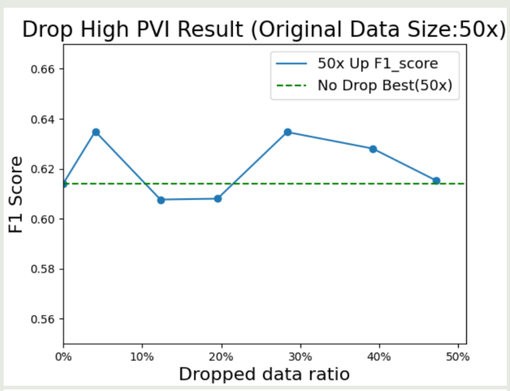
Results on Different Filtering Ratios

- Higher PVI → model already knows
- Lower PVI → model does not know (noisy or informative)

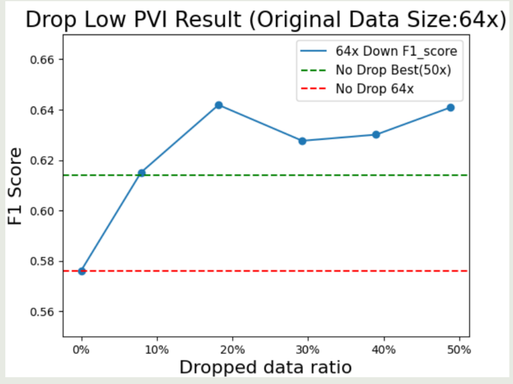
Synthetic Size

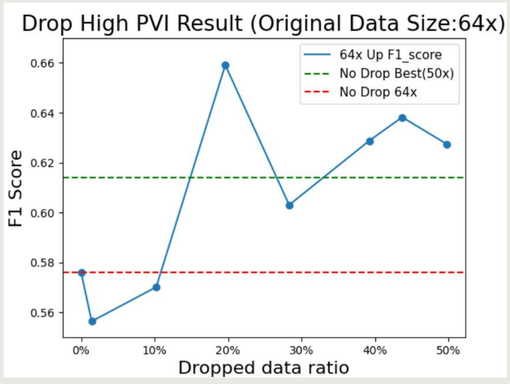
50x:



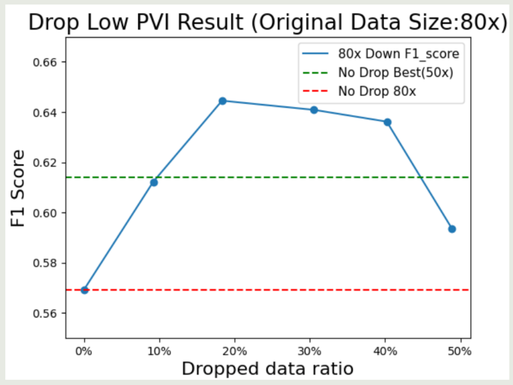


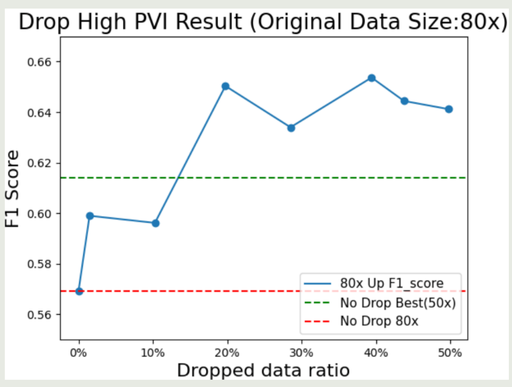
64x:





80x:





✓ PVI is more useful when generating large size of data

Four Quadrant of Synthetic Data

High PVI

Wrongly labeled data w/ high PVI

Text: John Smith ✗
Label: people
Prediction: John Smith ✗ (people)
PVI Score: 1.170

Correctly labeled data w/ high PVI

Text: July 2nd 2020 ✓
Label: date
Prediction: July 2nd 2020 ✓ (date)
PVI Score: 2.06

Wrong Data | **Correct Data**

Wrongly labeled data w/ low PVI

Text: 5 pm
Label: time ✗
Prediction: 5 pm ✓ (time)
PVI Score: -2.166

Correctly labeled data w/ low PVI

Text: October 3rd is the deadline.
Label: date ✓
Prediction: October 3rd is the deadline. — **No slot in text** ✗
PVI Score: -0.424

Low PVI

LLM Filter

Text1: We would like to booked the table for 8 pm on Saturday, February 12.

Time slot: 8 pm ✓ 8 ✗

Text2: John <first_name> and Mary <last_name> will be dining at 7 pm.

Align specified format? ✗

PVI Concept

- High PVI: stabilize prediction of the model
  - too many High PVI data → overfitting
- Low PVI: provide new information for the model
  - data with too low PVI → wrong label

PVI (Pointwise V-Information) Algorithm

PVI Formula: $PVI(x_i^* \longrightarrow y_i^*) = -\log_2 \hat{g}[\emptyset]\,(y_i^*) + \log_2 g'\,[x_i^*]\,(y_i^*)$

Original Data(x_i, y_i) → Trained model(g') → Synthetic Data (x_i^*, y_i^*) → Loss of synthetic data $log_2 g'[x_i^*](y_i^*)$

Train

Empty Data1 $(\emptyset, y_i)$ → Empty model$(\hat{g})$ → Empty Data2 $(\emptyset, y_i^*)$ → Loss of empty data2 $log_2 \hat{g}[\emptyset](y_i^*)$

Results

Model	Slot F1
Span-BERT w/ Scarce Data (64 samples)	40.0
+ Synthetic Data (50x)	57.4
+ Synthetic Data (50x) w/ LLM Filter	61.4
+ Synthetic Data (64x) w/ LLM Filter	57.6
+ Synthetic Data (64x) w/ LLM + PVI Filters	**65.9**

Takeaway

- ICDA is helpful for scarce data to train in slot filling task
- LLM can filter wrongly-labeled data
- PVI provides information about data usage
- Our method is general to diverse tasks