Low-Hanging Fruit: Knowledge Distillation from Noisy Teachers for Open Domain Spoken Language Understanding

ECML PKDD 2024

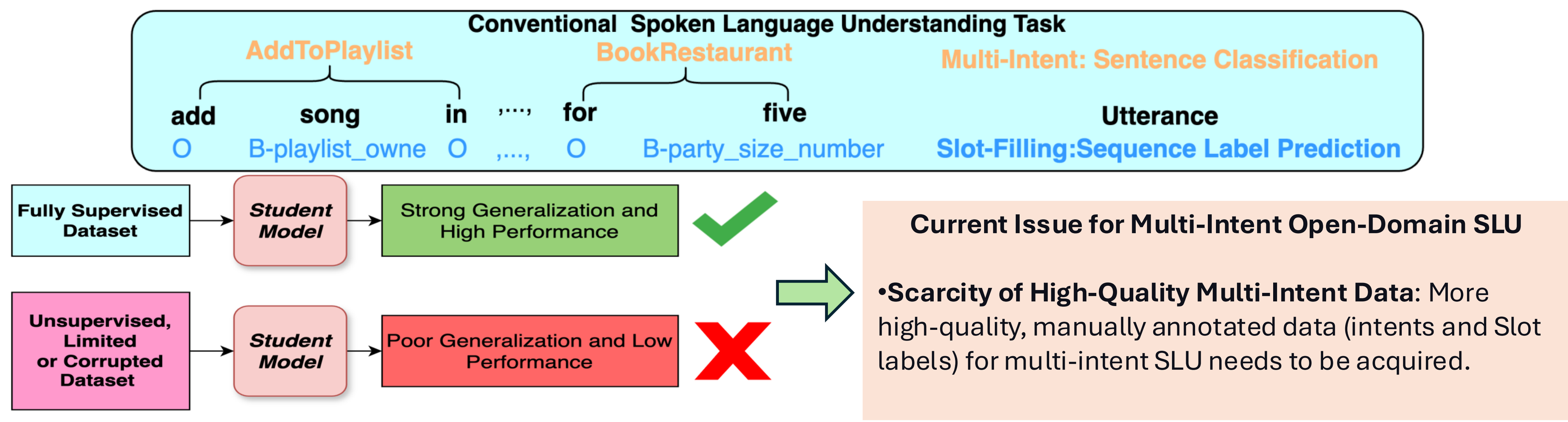
Cheng Chen [1,2] Bowen Xing [1,5,6], Ivor W Tsang [1,2,3,4]

[1]University of Technology Sydney & [2]CFAR, Agency for Science, Technology and Research, Singapore
& [3]IHPC, Agency for Science, Technology and Research, Singapore
& [4]College of Computing and Data Science, Nanyang Technological University
& [5]Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing
& [6]School of Computer and Communication Engineering, University of Science and Technology Beijing

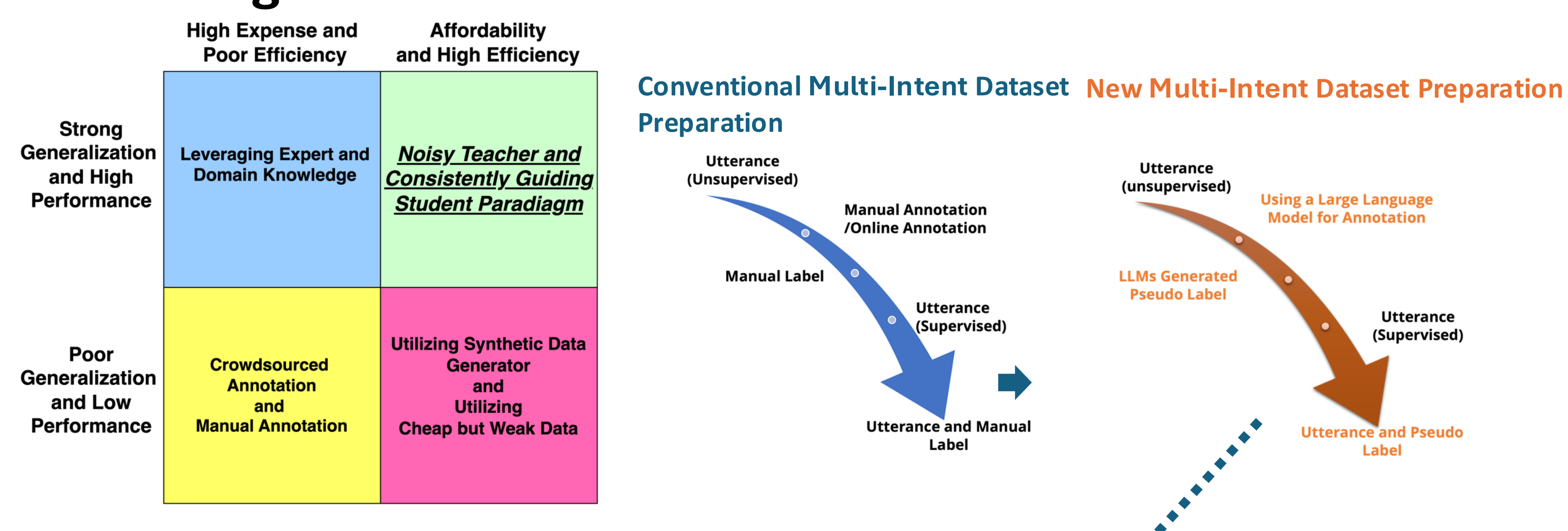
https://github.com/858006908cc/Low-Hanging-Fruit-Knowledge-Distillation-from-Noisy-Teachers-for-ODS-Language-Understanding.git

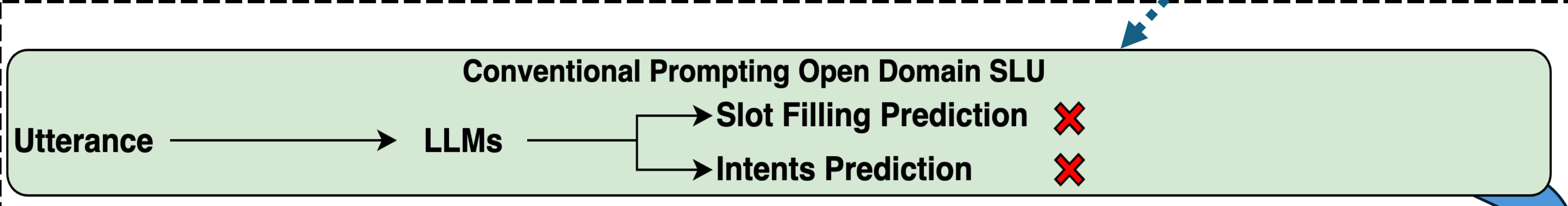
Introduction

- **Background**



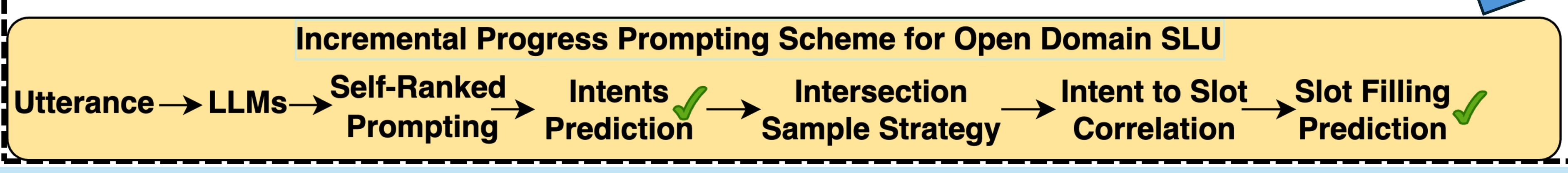
- **Challenges**



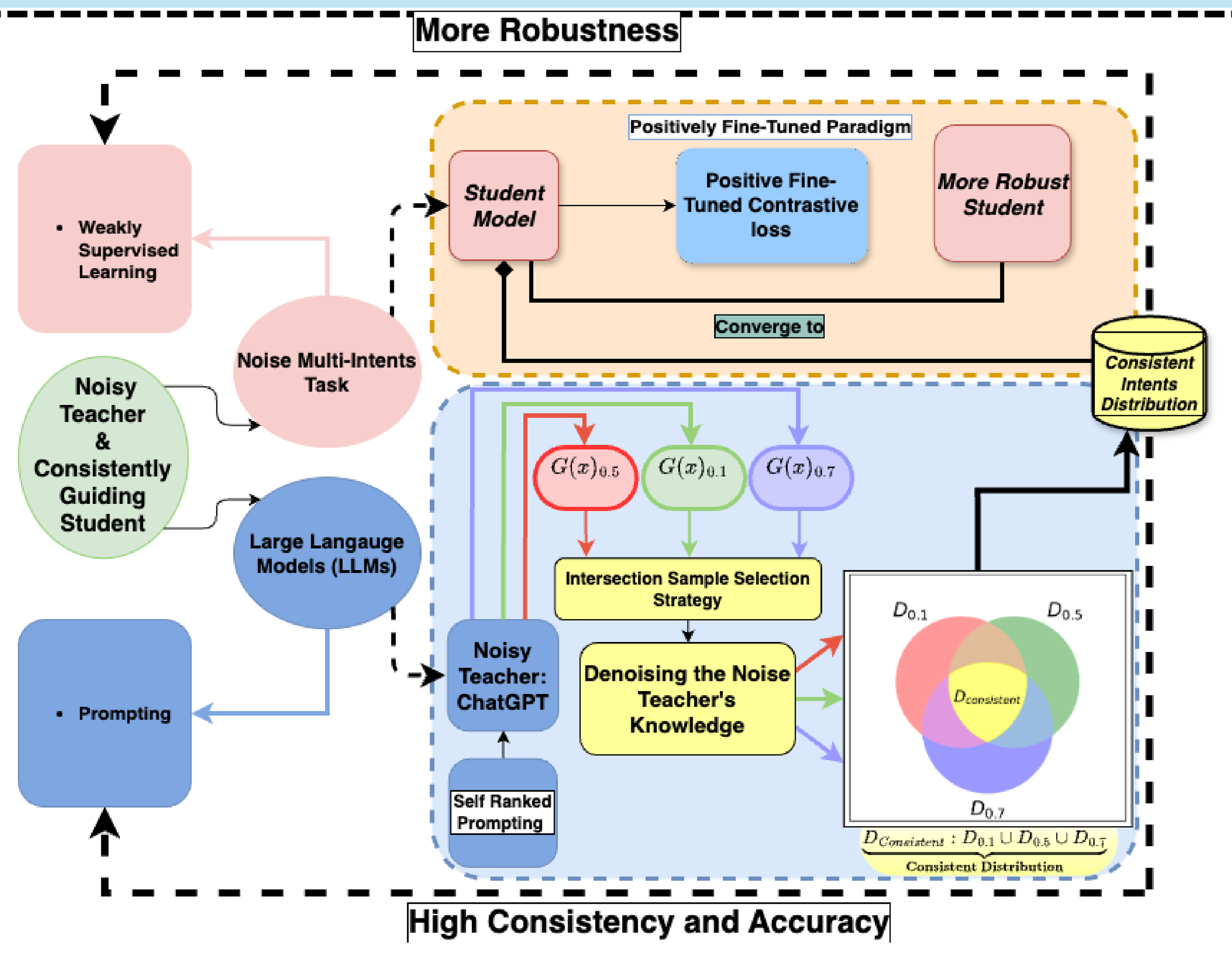


Noisy Teacher Generated Output

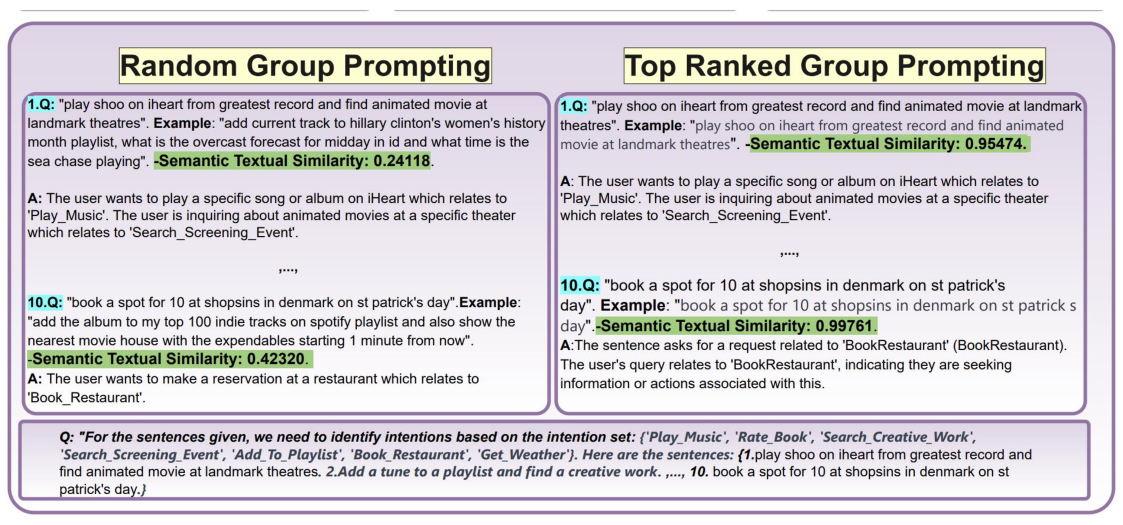
1. **Inconsistency**: The output manifests inconsistency and randomness issues.
2. **Struggles with Slot Filling**: The LLMs struggle with accurately filling slots.
3. **Accuracy**: The current accuracy is unsatisfactory and requires further improvement.



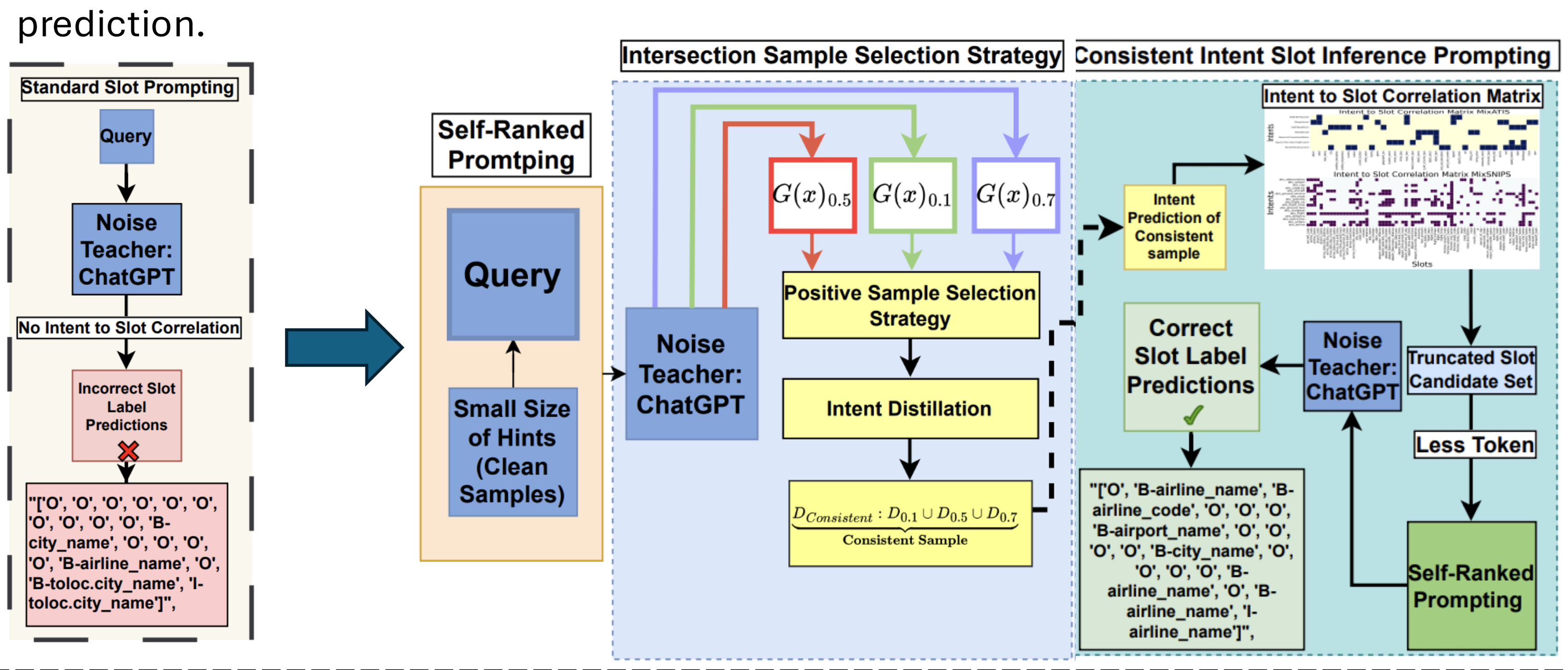
Noise Teacher and Consistently Guiding Student Framework



1. **Self-Ranked Prompting for Intent**: This approach generates a more consistent multi-intent dataset.



2. **Consistent Intent Slot Prompting for Open-Domain Slot-Filling Task**: We have incorporated refined LLM-generated multi-intent information to assist with slot prediction.



Noise Teacher and Consistently Guiding Student Framework

- **Label-wise Embedding Regularisation**

$$C_+(j) = \{c | c \in P(j), i \in I{:} M_{ci} > 1\},$$

$$M_{ic} = \begin{cases} 1 & \text{if } \vec{\boldsymbol{Y}}_i^{\,T} \cdot \vec{\boldsymbol{Y}}_c > 1 \\ 0 & \text{otherwise} \end{cases}.$$

- **Equivalent Anchors using Reliable Feature Representation.**

$$\mathcal{L}(f(x), \tau, C, A) = -\sum_{j \in P} \frac{1}{|C_+(j)|} \sum_{c \in C_+(j)} \log \frac{\exp(z_j^\top z_c / \tau)}{\sum_{a \in A(j)} \exp(z_j^\top z_a / \tau)},$$

- **Label Consistency Regularisation using Intersection Sample Prior.**

$$\mathcal{L}_{\text{ISPL}} = -\frac{1}{K} \sum_{j=1}^{K} \left[\vec{\boldsymbol{Y}}_{0.3_j} \log(\sigma(V_{m_j})) + (1 - \vec{\boldsymbol{Y}}_{0.3_j}) \log(1 - \sigma(V_{m_j})) \right],$$

$$T_{ij} = \boldsymbol{Y}_{0.3+ij} \cdot m_{ij}, \quad \forall i \in \{1, \ldots, n\},\ j \in \{1, \ldots, k\},$$

$$m_{i,j} = exp(\alpha \cdot \max(\lambda_{i,j} \cdot N_j)) \quad, \forall j \in \{1, \ldots, k\},$$

- The goal is to guide the student model to treat each class uniformly, avoiding biases caused by skewed frequencies of noisy multi-partial labels that do not reflect the true data distribution.

Experiments

Table 1. Comparison of Self-Ranked Prompting and Chain of Thought Prompting in the Intent Prediction Task for MixSNIPS and MixATIS

t	0.1	0.3	0.5	0.7	Average
MixATIS - Chain of Thought Prompting					
Accuracy Ratio	0.31	0.30	0.30	0.29	0.30
Subset Ratio	0.52	0.52	0.52	0.50	0.51
MixATIS - Self-Ranked Prompting					
Accuracy Ratio	0.37	0.38	0.38	0.34	0.37
Subset Ratio	0.57	0.58	0.58	0.55	0.57
MixSNIPS - Chain of Thought Prompting					
Matching Ratio	0.40	0.41	0.44	0.46	0.43
Subset Ratio	0.45	0.46	0.48	0.50	0.47
MixSNIPS - Self-Ranked Prompting					
Accuracy Ratio	0.66	0.68	0.68	0.69	0.68
Subset Ratio	0.75	0.76	0.77	0.77	0.76

- **Self-Ranked Prompting**: Providing more relevant examples helps improve the output accuracy of LLMs.

Table 2. Consistent Distribution Generation Via Intersection Sample Selection: Results for MixATIS and MIXSNIPS Datasets with Intersection Sample Selection. The 3468 of 13162 is the sample size for consistent sample distribution MixATIS. The 23186 of 39776 is the sample size for consistent sample distribution MIXSNIPS.

Dataset	Metric	$D_{\text{consistent}}$	% of the Dataset Left
MixATIS	Accuracy Ratio	**59.02%**	26.34%
MixATIS	Subset Ratio	63.41%	26.34%
MIXSNIPS	Accuracy Ratio	**80.78%**	58.29%
MIXSNIPS	Subset Ratio	87.74%	58.29%

- **Consistent Sample Selection**: Consistent samples have demonstrated significantly higher accuracy and subset ratio compared to LLM-generated predictions across all samples.

Table 3. Comparison of Random and Top Self-Ranked Prompting methods on the MixSnips and MixATIS datasets. The Group Random and Top Ranking are shown in Appendix Fig.4.

Dataset	Prompting	Accuracy Ratio	Subset Ratio
MixSNIPS	Group Random Ranking	$64.75 \pm 2.65\%$	$70.75 \pm 12.37\%$
	Group Top Ranking	$\mathbf{74.19} \pm 3.86\%$	$\mathbf{81.18} \pm 8.49\%$
MixATIS	Group Random Ranking	$37.00 \pm 10.54\%$	$56.00 \pm 1.00\%$
	Group Top Ranking	$\mathbf{39.75} \pm 0.50\%$	$\mathbf{58.75} \pm 4.03\%$

- **Effectiveness of Group Random Ranking**: Selecting top-ranked relevant examples, rather than providing random samples, helps improve the output accuracy of LLMs.

Table 4. A Comparison between Consistent Intent Slot Prompting (CISP) and Standard Slot Prompting for Slot Filling Task on MixATIS and MixSNIP.

MixSNIP	**Our (CISP)**	**BaseLine**
Total Exact Match	**3.85%**	0.57%
Average F1 Score	**24%**	7%
MixATIS	**Our (CISP)**	**BaseLine**
Total Exact Match%/Number	**6.49%**	3.15%
Average F1 Score	**35%**	26%

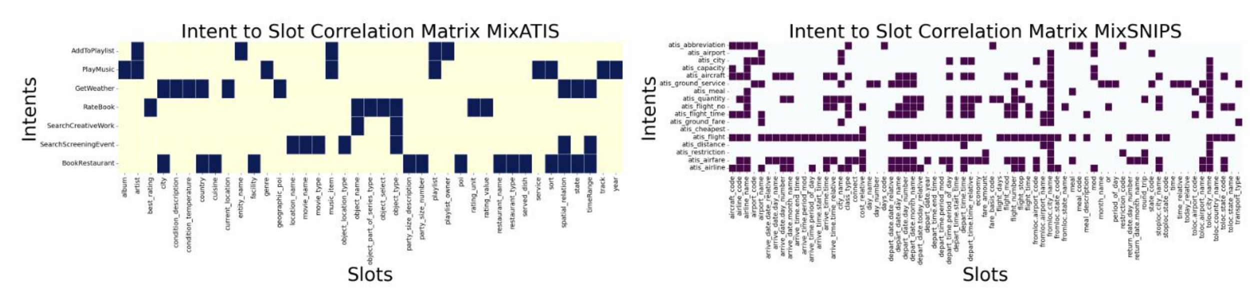


Fig. 4. Intents to Slots Correlation Matrices on MixSNIP and MixATIS

- Incorporating **the Intents-to-Slots Correlation Matrices** significantly improves both the F1 score and the exact matching ratio for slot-filling predictions with Large Language Models (LLMs).

Table 5. Comparison and Improvement of MixATIS with Different Models and PFTS Loss.

Dataset	Model	Loss Function	Intent Acc.	Precision	Recall	F1-Score
MixATIS						
	BERT	BCE Loss	$51.96 \pm 2.00\%$	$71.90 \pm 1.23\%$	$\mathbf{71.33} \pm 2.43\%$	$70.11 \pm 2.06\%$
	BERT	(ISPL+PFTS)	$\mathbf{54.88} \pm 3.01\%$	$\mathbf{73.19} \pm 0.51\%$	$71.52 \pm 1.58\%$	$\mathbf{70.95} \pm 1.19\%$
	Roberta	BCE Loss	$51.14 \pm 1.17\%$	$71.79 \pm 1.55\%$	$70.65 \pm 1.78\%$	$69.63 \pm 0.89\%$
	Roberta	(BCE+ISPL+PFTS)	$\mathbf{53.48} \pm 0.78\%$	$\mathbf{72.53} \pm 0.60\%$	$\mathbf{71.65} \pm 0.69\%$	$\mathbf{70.86} \pm 0.52\%$
MixSNIPS						
% Training samples						
5%	BERT	ISPL + PFTS	$\mathbf{70.72} \pm 1.14\%$	$90.91 \pm 0.88\%$	$90.76 \pm 0.52\%$	$90.77 \pm 0.29\%$
5%	BERT	BCE	$69.31 \pm 1.20\%$	$91.41 \pm 0.20\%$	$88.15 \pm 0.080\%$	$89.26 \pm 0.06\%$
10%	BERT	ISPL + PFTS	$\mathbf{79.44} \pm 0.56\%$	$95.57 \pm 0.79\%$	$90.88 \pm 0.71\%$	$93.03 \pm 0.19\%$
10%	BERT	BCE	$76.23 \pm 1.32\%$	$95.34 \pm 1.5\%$	$89.15 \pm 0.84\%$	$91.98 \pm 0.47\%$
50%	BERT	ISPL + PFTS	$\mathbf{85.73} \pm 0.55\%$	$96.50 \pm 0.45\%$	$94.76 \pm 1.57\%$	$95.03 \pm 0.05\%$
50%	BERT	BCE	$83.33 \pm 1.06\%$	$96.94 \pm 0.57\%$	$91.98 \pm 0.69\%$	$94.25 \pm 0.29\%$
100%	BERT	ISPL + PFTS	$\mathbf{87.55} \pm 0.61\%$	$96.66 \pm 0.19\%$	$94.55 \pm 0.37\%$	$95.50 \pm 0.20\%$
100%	BERT	BCE	$85.76 \pm 1.15\%$	$97.00 \pm 0.20\%$	$93.20 \pm 0.67\%$	$94.94 \pm 0.39\%$
100%	Robert	ISPL + PFTS	$\mathbf{88.93} \pm 0.12\%$	$96.75 \pm 0.23\%$	$95.40 \pm 0.18\%$	$96.01 \pm 0.16\%$
100%	Robert	BCE	$86.56 \pm 0.44\%$	$97.22 \pm 0.42\%$	$93.42 \pm 0.62\%$	$95.16 \pm 0.32\%$
100%	X-Lnet	ISPL + PFTS	$\mathbf{88.55} \pm 0.61\%$	$96.66 \pm 0.44\%$	$94.75 \pm 0.28\%$	$95.57 \pm 0.18\%$
100%	X-Lnet	BCE	$85.79 \pm 1.09\%$	$97.06 \pm 0.51\%$	$93.22 \pm 0.316\%$	$94.98 \pm 0.10\%$

Table 6. Comparison and Improvement of MixATIS and MixSNIP Datasets for Intent and Slot Filling Using Our Method (ISPL+PFTS) Vs the Baseline BCE Loss Model.

Dataset	Model	Loss Function	Intent Accuracy	Slot F1 Score
MixATIS	BERT	BCE Loss	30.37 ± 1.45	14.35 ± 0.39
	BERT	ISPL+PFTS	$\mathbf{34.82} \pm 3.76$	$\mathbf{21.33} \pm 0.58$
MixSNIP	BERT	BCE Loss	$72.26 \pm 1.15\%$	$14.90 \pm 0.82\%$
	BERT	ISPL+PFTS	$\mathbf{73.50} \pm 1.91\%$	$\mathbf{16.21} \pm 0.78\%$

- Our proposed Label Consistency Regularization, incorporating Intersection Sample Prior and Label-wise Embedding Regularization, has significantly improved intent accuracy in the student model task and **Slot F1 Score on the student model Task.** .