$\lvert\mathbb{C}_{\text{train}}\rvert$	16		32		64		128	
Domain	$\mathbb{S}$	$\mathbb{C}$	$\mathbb{S}$	$\mathbb{C}$	$\mathbb{S}$	$\mathbb{C}$	$\mathbb{S}$	$\mathbb{C}$
BERT2SEQ	82.8 ±1.0	37.7 ±10.3	82.8 ±0.8	57.4 ±7.1	82.4 ±0.2	71.1 ±2.7	81.8 ±0.9	75.8 ±2.0
+TS (Token-level Sup.)	82.9 ±0.5	47.1 ±4.0	82.5 ±0.7	65.1 ±1.8	83.1 ±0.4	72.1 ±0.9	82.3 ±0.6	77.5 ±1.5
+SS (Span-level Sup.)	83.3 ±0.7	**54.9** ±3.4	**83.4** ±0.6	**67.5** ±2.0	82.8 ±0.6	**76.0** ±1.3	82.6 ±0.3	**78.7** ±0.9
COARSE2FINE (DL18)	82.5 ±0.8	44.7 ±4.9	83.0 ±1.0	60.0 ±4.2	82.5 ±0.4	72.4 ±1.4	83.0 ±0.9	75.0 ±0.9
+TS (Token-level Sup.)	83.0 ±0.3	51.0 ±4.6	82.9 ±0.9	64.2 ±1.8	82.6 ±0.6	74.0 ±0.5	82.8 ±0.4	78.1 ±0.9
+SS (Span-level Sup.)	83.1 ±0.4	**54.2** ±3.0	83.1 ±0.5	**66.6** ±1.6	**83.5** ±0.9	74.8 ±1.1	82.9 ±0.4	78.2 ±0.5

Table 1: TEST. accuracies and standard deviation on the SMCALFLOW-CS Compositional Skills dataset w.r.t. the size of compositional examples included in the training set. We report both the results on the in-domain single-skill examples ($\mathbb{S}$) as well as the compositionally generalized multi-skill examples ($\mathbb{C}$). Results are averaged over five random random seeds. **Bold** results have *p*-values ≤ 0.01 when comparing to other systems in the same category using paired permutation test.

Split	MCD$_1$			MCD$_2$			MCD$_3$		
	C	R	All	C	R	All	C	R	All
T5-BASE	**55.8** ±4.8	77.4 ±4.7	**62.4** ±4.5	34.8 ±2.9	29.4 ±2.5	33.0 ±2.4	**21.6** ±8.6	34.4 ±2.8	23.0 ±1.7
+ TS	44.9 ±4.7	**86.4** ±2.4	57.7 ±3.4	32.4 ±3.1	32.7 ±1.4	32.5 ±2.1	14.3 ±1.5	36.6 ±1.7	22.0 ±0.7
+ SS	48.2 ±4.4	80.5 ±2.2	58.2 ±2.8	34.8 ±2.3	**36.4** ±2.8	**35.4** ±1.6	14.6 ±2.1	**40.1** ±3.5	23.8 ±1.0

Table 2: TEST. accuracies on CFQ MCD splits with 95% confidence interval, for **C**onjunctive, **R**ecursive, and **A**ll the samples. **Bold** results have *p*-values ≤ 0.01 when comparing to other systems in the same category using paired permutation test.

Model	DEV.	TEST
Oren et al. (2020)	28.9	34.4
+ Token-level Sup.	31.2 ±1.2	34.5 ±0.9
+ Span-level Sup.	31.1 ±0.6	35.0 ±2.0

Table 3: Accuracies and standard deviation on the ATIS text-to-SQL *program template* split. Results averaged over five random runs.

Model	DEV.	TEST
Oren et al. (2020)	78.4	74.5
+ Token-level Sup.	76.7 ±0.6	72.5 ±1.6
+ Span-level Sup.	78.4 ±0.8	74.0 ±0.5

Table 4: Accuracies and standard deviation on the ATIS text-to-SQL *standard i.i.d.* split. Results averaged over five random runs.

Here we present updated experiment results with standard deviation. For SMCALFLOW-CS and CFQ, we run with more (five) random seeds (three was used in the original submission). For completeness, on SMCALFLOW-CS we also include test accuracies on in-domain single-skill examples ($\mathbb{S}$), which have the same compositional patterns as the training single-skill samples. On CFQ, we follow Furrer et al. (2020) and report 95% confidence intervals.