Supplementary Material for Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task

Tao Yu **Rui Zhang** **Kai Yang** **Michihiro Yasunaga**
Dongxu Wang **Zifan Li** **James Ma** **Irene Li**
Qingning Yao **Shanelle Roman** **Zilin Zhang** **Dragomir R. Radev**
Department of Computer Science, Yale University
{tao.yu, r.zhang, k.yang, michihiro.yasunaga, dragomir.radev}@yale.edu

1 SQL Hardness Criteria

For all different evaluation metrics, we would get the scores on all SQL-question pairs. Also, we would like to know the scores on SQL with different hardness levels.

We first define:

- SQL components 1: WHERE, GROUP BY, ORDER BY, LIMIT, JOIN, OR, LIKE, HAVING
- SQL components 2: EXCEPT, UNION, INTERSECT, NESTED
- Others: number of AGG > 1, number of select columns > 1, number of where conditions > 1, number of group by clauses > 1, number of group by clauses > 1 (no consider col1-col2 math equations etc.)

Then different hardness levels are determined as follows.

- Easy: if SQL key words have ZERO or exact ONE from [SQL components 1] and SQL do not satisfy any conditions in [Others] above. AND no word from [SQL components 2].
- Medium: SQL satisfies no more than two rules in [Others] and do not have more than one word from [SQL components 1]. AND no word from [SQL components 2]. Or, SQL has exact 2 words from SQL components 1 and less than 2 rules in [Others]. AND no word from [SQL components 2]
- Hard: SQL satisfies more than two rules in [Others], with no more than 2 key words in [SQL components 1] and NO word in [SQL components 2]. Or, SQL has $2 <$ number key words in [SQL components 1] $<= 3$ and satisfies no more than two rules in [Others] but NO word in [SQL components 2]. Or, SQL has no more than 1 key word in [SQL components 1] and NO rule in [Others], but exact one key word in [SQL components 2].
- Extra Hard: All others left.
- All: just use all SQL-question pairs to compute different scores listed below.

For all SQL-question pairs labeled with different hardness levels, you are going to compute scores based on below different evaluation metrics.