Prefix-to-SQL: Text-to-SQL Generation from Incomplete User Questions

Naihao Deng\textsuperscript{1} Shuaichen Chang\textsuperscript{2} Peng Shi\textsuperscript{3} Tao Yu\textsuperscript{4} Rui Zhang\textsuperscript{5}

\textsuperscript{1} The University of Michigan, Ann Arbor \textsuperscript{2} The Ohio State University
\textsuperscript{3} The University of Waterloo \textsuperscript{4} The University of Hong Kong \textsuperscript{5} Penn State University
dnaihao@umich.edu, chang.1692@osu.edu, peng.shi@uwaterloo.ca, tao.yu@yale.edu, rmz5227@psu.edu

Abstract

Existing text-to-SQL research only considers complete questions as the input, but lay-users might strive to formulate a complete question. To build a smarter natural language interface to database systems (NLIDB) that also processes incomplete questions, we propose a new task, prefix-to-SQL which takes question prefix from users as the input and predicts the intended SQL. We construct a new benchmark called PAGSAS that contains 124K user question prefixes and the intended SQL for 5 sub-tasks Advising, GeoQuery, Scholar, ATIS, and Spider. Additionally, we propose a new metric SAVE to measure how much effort can be saved by users. Experimental results show that PAGSAS is challenging even for strong baseline models such as T5. As we observe the difficulty of prefix-to-SQL is related to the number of omitted tokens, we incorporate curriculum learning of feeding examples with an increasing number of omitted tokens. This improves scores on various sub-tasks by as much as 9% RECALL and 3.9% SAVE scores on sub-task GeoQuery in PAGSAS.

1 Introduction

Text-to-SQL aims to translate natural utterances to executable SQL queries in relational databases. Effective natural language interfaces to databases (NLIDB) give lay-people access to vast amounts of data stored in relational databases (Finegan-Dollak et al., 2018). However, existing text-to-SQL research only considers complete questions as the input. Users might struggle to formulate proper questions to retrieve their desired information (Sordoni et al., 2015; McCamish et al., 2018).

To build a more user-friendly NLIDB model as shown in Figure 1, we propose a new task, prefix-to-SQL which takes the prefix of user questions and predicts the intended SQL for the database system. For the second stage of SQL-to-text, we collect canonical rules for converting the intended SQL queries to question texts. Users can quickly choose one of the suggested question completions or alternatives and thus do not have to type the whole question by themselves (Bhatia et al., 2011; Shokouhi, 2013; Cai and De Rijke, 2016). The rules guarantee that the suggested question matches the SQL query (Yao et al., 2019). Thus the executed SQL query is what the user intends.

We focus on prefix-to-SQL in Figure 1 in this work. To simplify and better define prefix-to-SQL, we make three assumptions: (1) Examples from the original text-to-SQL dataset represent how database users are “likely” to query the database. We consider these SQL queries as “correct” SQL predictions for prefix-to-SQL because we think users would have similar intentions to the existing database users. (2) We consider SQL queries that are not in the original text-to-SQL dataset as “incorrect” because they are “unlikely” SQL queries when users query the database. In other words, we do not want our models to cover all possible combinations of tables, columns, and values for SQL. Instead, models need to learn how users are “likely” to query the database in their daily usage. (3) Following query auto-completion works in information retrieval (Sordoni et al., 2015; Bhatia et al., 2011), we consider the question prefix to be
the primary incomplete question type and work on prefix-to-SQL in this work.

For prefix-to-SQL, we build the PAGSAS benchmark, which consists of 5 sub-tasks and 124K examples from Advising (Finegan-Dollak et al., 2018), GeoQuery (Zelle and Mooney, 1996a), Scholar (Iyer et al., 2017), ATIS (Price, 1990; Dahl et al., 1994) and Spider (Yu et al., 2018) (Examples in Table 1). To evaluate the performance of our baseline models on PAGSAS, we present RECALL and MRR scores for how many SQL predictions are correct and how high the model ranks the first correct SQL query, respectively. Additionally, we propose a new metric, SAVE which measures how much user effort can be saved by models from the user’s perspective.

Our results show that both generative and retrieval models can predict SQL queries that match the user’s intention, but the scores are much lower than the original text-to-SQL task (Finegan-Dollak et al., 2018; Dong and Lapata, 2018a; Yin and Neubig, 2018; Cao et al., 2019; Rubin and Berant, 2021; Huang et al., 2021; Cao et al., 2021; Shi et al., 2020; Zhao et al., 2021; Yu et al., 2021). PAGSAS is challenging even for strong baseline models such as T5. By analysis of models’ performance, we find prefix-to-SQL differs from the original text-to-SQL task. For the original text-to-SQL task, short SQL queries are considered easier than longer ones (Finegan-Dollak et al., 2018). In contrast, models’ performance is negatively correlated with the number of omitted tokens for prefix-to-SQL. Thus, prefix-to-SQL poses new challenges compared to the existing text-to-SQL.

Based on our analysis, we adopt curriculum learning by feeding examples with an increasing number of omitted tokens. This significantly improves the model’s performance on various sub-tasks in PAGSAS. On sub-task GeoQuery, T5 with curriculum learning improves RECALL and SAVE by as much as 9% and 3.9%, respectively.

In summary, our contributions are three-fold:

- We introduce a new task prefix-to-SQL with the PAGSAS dataset that predicts SQL queries based on their question prefix. We design a new metric SAVE and evaluate baselines’ performance with metrics RECALL, MRR and SAVE on PAGSAS.

- Experiments show that PAGSAS is challenging even for strong text-to-SQL baseline models such as T5. We analyze the baseline results and demonstrate that prefix-to-SQL poses new challenges to existing models.

- We propose curriculum learning based on number of omitted tokens in the prefix, which improves scores on various sub-tasks by as much as 9% RECALL scores on sub-task GeoQuery in PAGSAS.

### 2 Related Work

**Question Auto-completion** Our task is inspired by the task of question auto-completion (QAC), but target at predicting SQL queries based on the incomplete question. QAC refers to that when the user gives a prefix, the user interface proposes alternative ways of extending the prefix to a full question query (Cai and De Rijke, 2016). This task is also known as type-ahead (Xiao et al., 2013; Cai et al., 2014; Li et al., 2009, 2011) or auto-complete suggestion (Jain and Mishne, 2010). Early work on QAC primarily focused on word prediction (Vanderheiden and Kelso, 1987; Swifflin et al., 1987; Darragh and Witten, 1991), while sentence completion received more attention later (Grabski and Scheffer, 2004; Bickel et al., 2005; Nandi and Jagadish, 2007). The task of sentence completion refers to that when the user gives a sentence’s initial fragment, the system identifies the remaining part of the sentence that the user intends to write (Cai
Table 2: Statistics for each domain in PAGSAS on question splits (“Q” columns) and SQL query splits (“S” columns). “#” means “the number of”, “µ” means “the average number of”. Because each SQL query corresponds to a complete question, # complete questions = # SQL. Note that µ omitted tokens ≠ µ tokens in complete questions - µ tokens in prefix, because we group prefix as described in Section 3.1.

3 Task Definition and Dataset Construction

We will focus on prefix-to-SQL as presented in Figure 1. For the second stage SQL-to-text, we present some of the rules in Table 6 in Appendix A.1.

3.1 Dataset Construction

Construction Methodology Following query auto-completion in information retrieval (Park and Chiba, 2017; Bar-Yossef and Kraus, 2011; Krishnan et al., 2020), we generate question prefixes from the first to the last token of each question from the existing text-to-SQL datasets. Note that prefixes spanning to the last token represent the complete question. We include this because users might go ahead and type their entire question in actual usages of our system.

For a question prefix $q^{pre}$, we assume the complete question $q^{cmp}$ represents the user’s intention, and the corresponding gold SQL $s^{gold}$ is the SQL representation of the user’s intention. Note that a question prefix can extend to multiple complete questions, so we group gold SQL queries if their corresponding complete questions share the same question prefix. We treat these SQL queries as the gold SQL predictions for the prefix following our first and second assumptions in Section 1. Therefore, each example in our dataset contains a question prefix $q^{pre}$ and its corresponding list of $L$ gold SQL queries $S^{gold}$ that matches the user’s intention:

$$S^{gold} = \{s^{gold}_1, s^{gold}_2, \ldots, s^{gold}_L\}$$ (1)

Source Text-to-SQL Datasets We construct PAGSAS by Prefixes for Advising, GeoQuery, Scholar, ATIS, and Spider because these large-scale datasets contain user questions collected in different domains and their corresponding SQL queries are manually annotated. As these datasets
contain different domain knowledge (Suhr et al., 2020), we treat them as different sub-tasks and perform experiments on each of them separately. We exclude data that is annotated as “exclude” by Finegan-Dollak et al. (2018).

For the original datasets other than Spider, we adopt the question split and the SQL query split following Finegan-Dollak et al. (2018). Question split splits data based on their complete questions, and SQL query split splits data based on their corresponding SQL templates. Note that question split is a common real-world setting as users tend to concern information about certain fields. For instance, in the original GeoQuery dataset, 66.25% of SQL queries include the column “state_name” from table “state” while only 0.11% SQL queries (1 SQL query) include the column “country_name” under the table “river”. But still, people might come up with questions or type in question prefixes that correspond to an unseen SQL. To evaluate the model under this challenge, we also include SQL query split in our experimental settings.

The original Spider dataset is a large cross-domain text-to-SQL dataset that consists of 10,181 questions and 5,693 unique complex SQL queries across 138 domains (Yu et al., 2018). The primary evaluation for the original Spider dataset is in the cross-database setting, where models are evaluated on examples for databases not seen during training. One of the primary challenges in this setting is the generalization to new database schemas, while in our task, it does not make sense to suggest questions or predict the corresponding SQL query for a completely unseen domain. Following Shaw et al. (2021), we adopt a setting similar to an alternative setting called the example split in the original dataset (Yu et al., 2018) where the databases are shared between train and test examples. Similar to Shaw et al. (2021), we identify examples from training set databases that contain more than 50 examples to ensure sufficient coverage over table and column names in the training data. We then generate two new training, validation, and test splits consisting of 2789 training, 493 validation, and 1094 test examples across 51 databases: a random split (question split) and a split based on SQL template (SQL query split).

Then we use the aforementioned construction methodology to construct PAGSAS for both question split and SQL query split.

**Dataset Analysis** Table 2 reports the statistics for PAGSAS. Figure 2a, 2b show the stacked distribution of the number of omitted tokens from complete questions and the number of omitted entities for question split; (c) the number of omitted tokens (vertical axis) v.s. the number of omitted entities (horizontal axis) for question split. The horizontal axis represents (a) the number of omitted tokens, (b) the number of entities in the omitted text for each sub-task in PAGSAS. The vertical line in (a) and (b) represents the corresponding number of examples.

**Dataset Analysis** Table 2 reports the statistics for PAGSAS. Figure 2a, 2b show the stacked distribution of the number of omitted tokens from complete questions and the number of omitted entities for question split; (c) the number of omitted tokens (vertical axis) v.s. the number of omitted entities (horizontal axis) for question split. The horizontal axis represents (a) the number of omitted tokens, (b) the number of entities in the omitted text for each sub-task in PAGSAS. The vertical line in (a) and (b) represents the corresponding number of examples.

**Dataset Analysis** Table 2 reports the statistics for PAGSAS. Figure 2a, 2b show the stacked distribution of the number of omitted tokens from complete questions and the number of omitted entities for question split; (c) the number of omitted tokens (vertical axis) v.s. the number of omitted entities (horizontal axis) for question split. The horizontal axis represents (a) the number of omitted tokens, (b) the number of entities in the omitted text for each sub-task in PAGSAS. The vertical line in (a) and (b) represents the corresponding number of examples.
3.2 Prefix-to-SQL Task Definition and Evaluation

Given the database schema \( c \) and a question prefix \( q^{\text{pre}} \), the model generates a list of top \( K \) SQL queries that match possible complete questions:

\[
S_K^\text{gold} = \{s_1^\text{gold}, s_2^\text{gold}, \ldots, s_K^\text{gold}\}
\]

Based on our first and second assumptions in the Section 1, we will only consider complete questions in the original text-to-SQL dataset as the gold SQL queries.

**Evaluation Metrics** We use RECALL to evaluate how many predicted SQL queries match the gold SQL queries. The equation is given as follows:

\[
\text{RECALL} = \frac{|S^\text{gold} \cap S^\text{sg} \text{ K}|}{|S^\text{gold}|}
\]

Additionally, we report mean reciprocal rank (MRR) (Schütze et al., 2008) scores to calculate how high the model will rank the first correctly predicted SQL query.

But neither RECALL nor MRR can measure how much effort users can save using the system. Thus, we propose \( \text{SAVE} \) to measure how well the model can save user effort. For each complete question \( q^{\text{cmp}} \), \( \text{SAVE} \) calculates which token (token \( t \)) the model can start predicting the correct SQL query corresponding to \( q^{\text{cmp}} \). Tokens after \( t \) in \( q^{\text{cmp}} \) are tokens the user does not need to type. When the model does not predict anything correct, \( \text{SAVE} \) becomes 0. Formally, we define \( \text{SAVE} \) as:

\[
\text{SAVE} = \frac{\text{len}(q^{\text{cmp}})}{\text{len}(q^{\text{cmp}})} - \min \{\text{len}(q^{\text{cmp}})\}_{s_j \in S}
\]

for \( |S| > 0 \) and \( \text{SAVE} = 0 \) for \( |S| = 0 \). In Eq 3, \( S \) is the set of SQL queries that are correctly predicted, given each prefix \( q_i^{\text{pre}} \) of the complete question \( q^{\text{cmp}} \). \( s_j \) is one of the correctly predicted SQL queries in \( S \). Function \( \text{len}() \) calculates the number of tokens.

For all the aforementioned metrics, we use the exact match result to judge the correctness.

**Metrics’ Ceiling Scores** There exist certain examples in PAGSAS where there are more than 5 gold SQL queries so that models cannot achieve 100% RECALL. The maximum average RECALL@5 and RECALL@10 scores a model can achieve in PAGSAS are 97% to 99% for each sub-task. For MRR, the maximum score a model can achieve will be 100% across all sub-tasks if the first SQL prediction is among the gold SQL queries. For \( \text{SAVE} \), the maximum score the model can achieve theoretically is 100%, but in practice, it is impossible to reach 100% because models cannot predict the correct corresponding SQL query with a zero-length prefix.

4 Baseline Models

We regard prefix-to-SQL task either as a SQL generation task or a SQL retrieval task. Thus, we experiment with baselines from both SQL generation models and retrieval models.

**Generation-based Model** We experiment with seq2seq models using a Bi-LSTM encoder and a LSTM decoder (Hochreiter and Schmidhuber, 1997) directly on question prefixes to generate SQL. We name the seq2seq model \( \text{PF-S2S} \). Seq2seq with attention (\( + \text{ATTN} \)), as well as seq2seq with attention and copy mechanism (\( + \text{ATTN + COPY} \)) are also evaluated. \( \text{T5} \) (Raffel et al., 2020) is a pre-trained sequence-to-sequence model based on the Transformer architecture (Vaswani et al., 2017). Following Shaw et al. (2021); Hazoom et al. (2021), we use T5 as our baseline model. Unlike the traditional setting where one question has a single gold SQL, a question prefix can match multiple gold SQL queries in prefix-to-SQL. Thus, our models optimize the sum of log-likelihood of all gold SQL queries following Jin and Ghahramani (2003):

\[
L = \sum_{q^{\text{pre}}} \log P(s_i^{\text{gold}}|q^{\text{pre}})
\]

We also use a two-stage model (QAC-S2S) that first uses the GPT2 language model (Radford et al., 2019) to auto-complete question prefixes to form complete questions. Then seq2seq with attention and copy mechanism translates the complete question into SQL queries. The two stages are trained separately, and we select the top K SQL predictions during testing.

**Retrieval-based Model** We finetune the RoBERTa (Liu et al., 2019) model to generate the embeddings for both question prefixes and historical SQL queries (SQL queries in the training set). We use the dot-product of their embeddings to represent the similarity between question prefix and SQL query. Historical SQL queries with the K highest similarity scores will be retrieved for
a question prefix during the inference. We use EMB-RTR to denote this model.

Additionally, we train a RoBERTa-based classification model (CLS-RTR) to distinguish relevant SQL queries from irrelevant SQL queries. The model predicts whether the SQL matches the intention of the question prefix directly. During inference, we rank all historical SQL queries and retrieve top K SQL queries based on the predicted probability by CLS-RTR.

5 Baseline Results and Analysis

Overall Results. Table 3 shows scores in terms of the three metrics on PAGSAS. Models that achieve good RECALL scores also achieve good MRR and SAVE scores, which indicates that models predicting the most number of correct gold SQL queries also rank the correct SQL queries higher and save more user efforts.

We do not report scores for retrieval models on SQL query split because retrieval models fail to retrieve unseen SQL queries. But in question split where the same SQL query might appear in training, CLS-RTR achieves a RECALL@5 of 60%, outperforming generative models on sub-task Spider (the highest RECALL@5 for generative models is 44%). Although CLS-RTR outperforms EMB-RTR on all the sub-tasks, CLS-RTR requires much more running time than EMB-RTR because it needs to run the RoBERTa classification model on all historical SQL queries with the given question prefix. In contrast, EMB-RTR can calculate and cache historical SQL embeddings in advance.

QAC-S2S achieves a good RECALL@5 (37%) on sub-task GeoQuery (same as RECALL@5 for T5). However, QAC-S2S performs poorly on sub-task Spider (< 1% RECALL@5) because the second stage is doing the original text-to-SQL on complete questions, and seq2seq models perform poorly on the original Spider (16.0% reported by (Yu et al., 2018)) compared to 71% on the original GeoQuery dataset reported by (Finegan-Dollak et al., 2018)). The cascade of errors in the two stages result in the poor performance of QAC-S2S on sub-task Spider.

T5 is a strong baseline as it achieves the highest RECALL@5 scores on sub-tasks Advising and Scholar (45% and 36%, respectively). In terms of RECALL@5 scores, T5 outperforms variants of PF-S2S (22% on Advising, 14% on Scholar) and QAC-S2S (8% on Advising, 11% on Scholar) by a large margin on Advising and Scholar. So are the cases for MRR and SAVE (For MRR@5 and SAVE@5, T5 achieves 39% and 29% on Advising while the second-highest are 19% and 12%, respectively). We use T5 as a strong baseline for our later curriculum learning setting because the performance of PF-S2S is comparable to QAC-S2S in most cases, and there is no clear indication that auto-completing prefix will benefit our task.

Generalization to Unseen SQL queries As discussed in Section 3.1, SQL query split is a challenging setting because of the unseen SQL templates

Table 3: RECALL@5 (R), MRR@5 (M) and SAVE@5 (S) in percentage for each subtask in PAGSAS on question split. We omit the results for sub-tasks Advising and Scholar where the same SQL query might appear in training, (S) in percentage generative models on SQL query split.

Table 4: RECALL@5 (R), MRR@5 (M) and SAVE@5 (S) in percentage generative models on SQL query split. We omit the results for sub-tasks Advising and Scholar where the same SQL query might appear in training, (S) in percentage generative models on SQL query split.

Table 3 shows scores in terms of the three metrics on PAGSAS. Models that achieve good RECALL scores also achieve good MRR and SAVE scores, which indicates that models predicting the most number of correct gold SQL queries also rank the correct SQL queries higher and save more user efforts.

We do not report scores for retrieval models on SQL query split because retrieval models fail to retrieve unseen SQL queries. But in question split where the same SQL query might appear in training, CLS-RTR achieves a RECALL@5 of 60%, outperforming generative models on sub-task Spider (the highest RECALL@5 for generative models is 44%). Although CLS-RTR outperforms EMB-RTR on all the sub-tasks, CLS-RTR requires much more running time than EMB-RTR because it needs to run the RoBERTa classification model on all historical SQL queries with the given question prefix. In contrast, EMB-RTR can calculate and cache historical SQL embeddings in advance.

QAC-S2S achieves a good RECALL@5 (37%) on sub-task GeoQuery (same as RECALL@5 for T5). However, QAC-S2S performs poorly on sub-task Spider (< 1% RECALL@5) because the second stage is doing the original text-to-SQL on complete questions, and seq2seq models perform poorly on the original Spider (16.0% reported by (Yu et al., 2018)) compared to 71% on the original GeoQuery dataset reported by (Finegan-Dollak et al., 2018)). The cascade of errors in the two stages result in the poor performance of QAC-S2S on sub-task Spider.

T5 is a strong baseline as it achieves the highest RECALL@5 scores on sub-tasks Advising and Scholar (45% and 36%, respectively). In terms of RECALL@5 scores, T5 outperforms variants of PF-S2S (22% on Advising, 14% on Scholar) and QAC-S2S (8% on Advising, 11% on Scholar) by a large margin on Advising and Scholar. So are the cases for MRR and SAVE (For MRR@5 and SAVE@5, T5 achieves 39% and 29% on Advising while the second-highest are 19% and 12%, respectively). We use T5 as a strong baseline for our later curriculum learning setting because the performance of PF-S2S is comparable to QAC-S2S in most cases, and there is no clear indication that auto-completing prefix will benefit our task.

Generalization to Unseen SQL queries As discussed in Section 3.1, SQL query split is a challenging setting because of the unseen SQL templates
Figure 3: T5’s (a) Recall@5 v.s. number of omitted tokens, (b) Recall@5 v.s. prefix length (number of tokens in prefix), (c) Recall@5 v.s. SQL length (number of tokens in SQL) on question split for each sub-task in PAGSAS. We plot Recall@5 corresponding to ≥ 50 and ≥ 100 examples for sub-tasks other than Spider and Scholar, respectively. Recall@5 is negatively correlated with number of omitted tokens and possesses no monotonic relationships with either prefix length or SQL length.

Figure 4: Save@5 v.s. omitted entity types for T5 for the top 4 most popular omitted entity types. T5 achieves similar Save@5 of around 50% to 60% on GPE, ORG, DATE, and CARDINAL on question split, and 35% to 40% on SQL query split. In other words, if users intend to ask questions including GPE (geopolitical entity), the model can auto-complete that entity for around 50% of the time if it sees the SQL template before.

Performance Analysis: Figure 4 shows T5’s Save@5 v.s. the omitted entity types. T5 achieves similar Save@5 of around 50% to 60% on GPE, ORG, DATE, and CARDINAL on question split, and 35% to 40% on SQL query split. In other words, if users intend to ask questions including GPE (geopolitical entity), the model can auto-complete that entity for around 50% of the time if it sees the SQL template before.

In Figure 3, T5’s Recall@5 scores are negatively correlated with the number of omitted tokens but possess no monotonic relationships with either the prefix length (number of tokens in the prefix) or SQL length (number of tokens in SQL queries). This differs prefix-to-SQL task from the original complete text-to-SQL task, as short SQL queries are considered easier than longer ones in the original text-to-SQL task (Finegan-Dollak et al., 2018). As discussed in Section 3.1, the number of omitted tokens possesses a positive linear relationship with the number of omitted entities as shown in Figure 2c. The hardness of the task is negatively correlated with the number of omitted tokens or the number of omitted entities as well.

6 Curriculum Learning

Based on our discovery of the relationship between Recall and the number of omitted tokens, we propose the use of curriculum learning (Bengio et al., 2009) to improve the T5 performance. The setups are shown in Figure 5. We propose the scoring function as:

\[ g_{\text{score}} = \text{len}(q^{\text{pre}}) - \text{len}(q^{\text{imp}}) \]  

(5)

to score the difficulty of each example (based on Eq 4, we use \( q^{\text{pre}} \) for each SQL query in \( S^{\text{gold}} \) as a
Table 5 shows the comparison of \textsc{recall}@5, MRR@5 and \textsc{save}@5 scores between the original T5 and T5 with curriculum learning. In PAGSAS, T5 with curriculum learning improves \textsc{recall}@5 on ATIS, GeoQuery, Scholar and Spider for question split, on Advising, ATIS, GeoQuery and Scholar for SQL query split. It improves \textsc{recall}@5 for as much as 9% on SQL query split for ATIS (T5+CL achieves 24.0% while T5 achieves 14.7%); MRR@5 for as much as 16% on question split for Spider (T5+CL achieves 36.1% while T5 achieves 19.7%); \textsc{save}@5 for as much as 3.9% on question split for GeoQuery (T5+CL achieves 14.7% while T5 achieves 10.8%).

Figure 6 shows comparison of \textsc{recall}@5 for T5 and T5 + CL for sub-task GeoQuery on question split. By involving the curriculum, the model achieves better \textsc{recall}@5 scores on different levels of difficulties for prefix-to-SQL. Although T5+CL improves performances for various sub-tasks, for sub-task advising on question split and Spider on SQL query split, we find that T5+CL
Figure 7: Prefixes with the least number of tokens where both models can predict the correct gold SQL in sub-task GeoQuery. T5+CL correctly predicts the gold SQL with one less token (underlined text) in prefix than the original T5.

performs worse than T5, indicating sub-tasks Advising and Spider might have their own challenges.

Figure 7 shows an example in sub-task GeoQuery that is correctly predicted by T5+CL with one less token than T5. T5+CL successfully predict “York” as the completion for the prefix “what is the capital of New”. And this does not impair T5+CL’s performance on prefixes that end with “New”. On examples involving “New Jersey”, “New Mexico” (examples which use “Jersy” or “Mexico” as the completion for prefixes that end with “New”), T5+CL performs as good as the original T5.

7 Conclusion

In this work, we propose prefix-to-SQL and construct a benchmark PAGSAS, making the first step to build a more user-friendly NLIDB system. To better evaluate models’ performance, apart from RECALL and MRR, we propose our own metric SAVE which measures how much user effort can be saved. Experiments show that PAGSAS is challenging even for strong baseline models such as T5. Analysis shows that different from the original text-to-SQL, the difficulty of prefix-to-SQL is related to the number of omitted tokens. Based on this discovery, we incorporate curriculum learning of feeding examples with an increasing number of omitted tokens. This improves scores on various sub-tasks in PAGSAS and metrics, and by as much as 9% RECALL score and 3.9% SAVE score on sub-task GeoQuery. However, even with curriculum learning, there is a large room for improvement for current models, indicating the necessity of future research.

Acknowledgment

We would like to thank Yulong Chen for his valuable advice on the abstract and introduction. We would also like to thank Chao Chen, Yiqun Yao, Yin Lin, Lingyi Jin, Laura Biester, Andrew Lee and Nan Zhang for proofreading this paper. This work is in part supported by an Amazon Research Award.

References

Ion Androutsopoulos, Graeme D Ritchie, and Peter Thanisch. 1995. Natural language interfaces to databases-an introduction. arXiv preprint cmp-lg/9503016.

Ziv Bar-Yossef and Naama Kraus. 2011. Context-sensitive query auto-completion. In Proceedings of the 20th international conference on World wide web, pages 107–116.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In Proceedings of the 26th annual international conference on machine learning, pages 41–48.

Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. 2011. Query suggestions in the absence of query logs. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pages 795–804.

Steffen Bickel, Peter Haider, and Tobias Scheffer. 2005. Learning to complete sentences. In European Conference on Machine Learning, pages 497–504. Springer.

Ben Bogin, Matt Gardner, and Jonathan Berant. 2019. Representing schema structure with graph neural networks for text-to-sql parsing. arXiv preprint arXiv:1905.06241.

Fei Cai and Maarten de Rijke. 2016. Learning from homologous queries and semantically related terms for query auto completion. Information Processing & Management, 52(4):628–643.

Fei Cai and Maarten De Rijke. 2016. Query auto completion in information retrieval. Universiteit van Amsterdam [Host].

Fei Cai, Shangsong Liang, and Maarten De Rijke. 2014. Time-sensitive personalized query auto-completion. In Proceedings of the 23rd ACM international conference on conference on information and knowledge management, pages 1590–1608.

Ruisheng Cao, Lu Chen, Zhi Chen, Yanbin Zhao, Su Zhu, and Kai Yu. 2021. Lgesql: Line graph enhanced text-to-sql model with mixed local and non-local relations.

Ruisheng Cao, Su Zhu, Chen Liu, Jiemy Li, and Kai Yu. 2019. Semantic parsing with dual learning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 51–64, Florence, Italy. Association for Computational Linguistics.
Deborah A Dahl, Madeleine Bates, Michael K Brown, William M Fisher, Kate Hunnicke-Smith, David S Pallett, Christine Pao, Alexander Rudnick, and Elizabeth Shriberg. 1994. Expanding the scope of the atis task: The atis-3 corpus. In HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994.

John J Darragh and Ian H Witten. 1991. Adaptive predictive text generation and the reactive keyboard. Interacting with Computers, 3(1):27–50.

Li Dong and Mirella Lapata. 2018a. Coarse-toFine decoding for neural semantic parsing. Li Dong and Mirella Lapata. 2018b. Coarse-to-fine decoding for neural semantic parsing. arXiv preprint arXiv:1805.04793.

Catherine Finegan-Dollak, Jonathan K Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-sql evaluation methodology. arXiv preprint arXiv:1806.09029.

Alessandra Giordani and Alessandro Moschitti. 2012. Generating sql queries using natural language syntactic dependencies and metadata. In International Conference on Application of Natural Language to Information Systems, pages 164–170. Springer.

Korinna Grabski and Tobias Scheffer. 2004. Sentence completion. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’04, page 433–439, New York, NY, USA. Association for Computing Machinery.

Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and Dongmei Zhang. 2019. Towards complex text-to-sql in cross-domain database with intermediate representation. arXiv preprint arXiv:1905.08205.

Catalina Hallett. 2006. Generic querying of relational databases using natural language generation techniques. In Proceedings of the fourth international natural language generation conference, pages 95–102.

Moshe Hazoom, Vibhor Malik, and Ben Bogin. 2021. Text-to-sql in the wild: A naturally-occurring dataset based on stack exchange data.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Junyang Huang, Yongbo Wang, Yongliang Wang, Yang Dong, and Yanghua Xiao. 2021. Relation aware semi-autoregressive semantic parsing for nl2sql.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. 2017. Learning a neural semantic parser from user feedback. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 963–973, Vancouver, Canada. Association for Computational Linguistics.

Alpa Jain and Gilad Mishne. 2010. Organizing query completions for web search. In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 1169–1178.

Jyun-Yu Jiang, Yen-Yu Ke, Pao-Yu Chien, and Pu-Jen Cheng. 2014. Learning user reformulation behavior for query auto-completion. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pages 445–454.

Rong Jin and Zoubin Ghahramani. 2003. Learning with multiple labels. In Advances in neural information processing systems, pages 921–928.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Unni Krishnan, Alistair Moffat, Justin Zobel, and Bodo Billerbeck. 2020. Generation of synthetic query auto completion logs. In European Conference on Information Retrieval, pages 621–635. Springer.

Fei Li and HV Jagadish. 2014. Constructing an interactive natural language interface for relational databases. Proceedings of the VLDB Endowment, 8(1):73–84.

Guoliang Li, Shengyue Ji, Chen Li, and Jianhua Feng. 2009. Efficient type-ahead search on relational data: a tastier approach. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, pages 695–706.

Guoliang Li, Shengyue Ji, Chen Li, and Jianhua Feng. 2011. Efficient fuzzy full-text type-ahead search. The VLDB Journal, 20(4):617–640.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7999–8009.
Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Ben McCamish, Vahid Ghadakchi, Arash Termehchy, Behrouz Touri, and Liang Huang. 2018. The data interaction game. In Proceedings of the 2018 International Conference on Management of Data, pages 83–98.

Bhaskar Mitra. 2015. Exploring session context using distributed representations of queries and reformulations. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, pages 3–12.

Arnab Nandi and HV Jagadish. 2007. Effective phrase prediction. In Proceedings of the 33rd international conference on Very large data bases, pages 219–230.

Dae Hoon Park and Rikio Chiba. 2017. A neural language model for query auto-completion. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1189–1192.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.

Ana-Maria Popescu, Alex Armanasu, Oren Etzioni, David Ko, and Alexander Yates. 2004. Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability. In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, pages 141–147.

Patti Price. 1990. Evaluation of spoken language systems: The atis domain. In Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Ohad Rubin and Jonathan Berant. 2021. Smbop: Semi-autoregressive bottom-up semantic parsing.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. Introduction to information retrieval, volume 39. Cambridge University Press Cambridge.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both?

Peng Shi, Patrick Ng, Zhiguow Wang, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Cicero Nogueira dos Santos, and Bing Xiang. 2020. Learning contextual representations for semantic parsing with generation-augmented pre-training.

Milad Shokouhi. 2013. Learning to personalize query auto-completion. In Proceedings of the 36th international ACM SIGIR conference on Research and Development in Information Retrieval, pages 103–112.

Milad Shokouhi and Kira Radinsky. 2012. Timesensitive query auto-completion. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’12, page 601–610, New York, NY, USA. Association for Computing Machinery.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pages 553–562.

Alane Suhr, Ming-Wei Chang, Peter Shaw, and Kenton Lee. 2020. Exploring unexplored generalization challenges for cross-database semantic parsing. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8372–8388.

Andrew Swifflin, John Arnott, J Adrian Pickering, and Alan Newell. 1987. Adaptive and predictive techniques in a communication prosthesis. Augmentative and Alternative Communication, 3(4):181–191.

Lappoon R. Tang and Raymond J. Mooney. 2000. Automated construction of database interfaces: Integrating statistical and relational learning for semantic parsing. In 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pages 133–141, Hong Kong, China. Association for Computational Linguistics.

Gregg Vanderheiden and David Kelso. 1987. Comparative analysis of fixed-vocabulary communication acceleration techniques. Augmentative and Alternative Communication, 3(4):196–206.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2020. RAT-SQL: Relation-aware schema encoding and linking for
text-to-SQL parsers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7567–7578, Online. Association for Computational Linguistics.

DHD Warren and FCN Pereira. 1981. An efficient easily adaptable system for interpreting natural language queries (dai research paper no. 155). department of artificial. Intelligence.

Stewart Whiting and Joemon M Jose. 2014. Recent and robust query auto-completion. In Proceedings of the 23rd international conference on World wide web, pages 971–982.

Chuan Xiao, Jianbin Qin, Wei Wang, Yoshiharu Ishikawa, Koji Tsuda, and Kunihiko Sadakane. 2013. Efficient error-tolerant query autocompletion. Proceedings of the VLDB Endowment, 6(6):373–384.

Navid Yaghmazadeh, Yuepeng Wang, Isil Dillig, and Thomas Dillig. 2017. Sqlizer: query synthesis from natural language. Proceedings of the ACM on Programming Languages, 1(OOPSLA):1–26.

Ziyu Yao, Yu Su, Huan Sun, and Wen-tau Yih. 2019. Model-based interactive semantic parsing: A unified framework and a text-to-sql case study. arXiv preprint arXiv:1910.05389.

Pengcheng Yin and Graham Neubig. 2017. A syntactic neural model for general-purpose code generation. arXiv preprint arXiv:1704.01696.

Pengcheng Yin and Graham Neubig. 2018. Tranx: A transition-based neural abstract syntax parser for semantic parsing and code generation. arXiv preprint arXiv:1810.02720.

Pengcheng Yin and Graham Neubig. 2019. Reranking for neural semantic parsing. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, and Caiming Xiong. 2021. Grappa: Grammar-augmented pre-training for table semantic parsing.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

John M Zelle and Raymond J Mooney. 1996a. Learning to parse database queries using inductive logic programming. In Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2, AAAI’96, page 1050–1055. AAAI Press.

Aston Zhang, Amit Goyal, Weiye Kong, Hongbo Deng, Anlei Dong, Yi Chang, Carl A Gunter, and Jiawei Han. 2015. adaqac: Adaptive query auto-completion via implicit negative feedback. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, pages 143–152.

Liang Zhao, Hexin Cao, and Yunsong Zhao. 2021. Gp: Context-free grammar pre-training for text-to-sql parsers. arXiv preprint arXiv:2101.09901.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. CoRR, abs/1709.00103.
A Appendices

A.1 Canonical Question Generation

We collect 118 SQL queries and question templates. Table 6 lists top 15 examples of SQL and question templates. It is worth mentioning that the listed top 15 rules can cover 80% of the question and SQL pairs from the original Spider dataset, which implies that we only need to collect a certain amount of question and SQL templates, and those templates can still cover a decent amount of possible SQL queries that users want.

Finegan-Dollak et al. (2018) classifies question and SQL templates from the original ATIS, Advising, GeoQuery, and Scholar.

A.2 Dataset Analysis

Table 7 reports statistics of the generated prefixes and their corresponding SQL queries. Figure 8 reports the distribution of prefix length (number of tokens in the prefix) as well as complete question length for the question split and SQL query split. Figure 9a and Figure 9b show stacked distribution of the number of omitted tokens and the number of omitted entities detected by SpaCy (Honnibal and Montani, 2017) on SQL query split for each sub-task of PAGSAS. Figure 9c demonstrates that there is also a linear relationship between the number of omitted tokens and the number of omitted entities for SQL query split, and more entities will be missing if there are more omitted tokens.

A.3 Model and Experiment Details

Figure 11 shows structures of our baseline models.

A.3.1 Model inputs

For variants of Pf-S2S models, T5, EMB-RTR and CLS-RTR, the models’ input is question prefix concatenated with the database schema for the prefix.

In the first stage of QAC-S2S, we feed original questions into the model directly to train the GPT-2 language model. For the second stage, we feed the original text-to-SQL dataset to the model.

A.3.2 Pf-S2S models

We implement variations of seq2seq (S2S) models by OpenNMT, a package proposed by Klein et al. (2017). We use Glove embedding (Pennington et al., 2014) for all variations of S2S models we experiment. We use bidirectional LSTM as our encoder, and use LSTM (Hochreiter and Schmidhuber, 1997) as our decoder. We have two layers with hidden size as 384 for our experiments.

A.3.3 QAC-S2S model

We finetune distilled version of GPT2 to complete questions based on the prefix. We set the learning rate to $10^{-5}$ and a warm-up ratio of the model to 0.2. The maximum sequence length is set to 256. Sequences that are longer than the maximum sequence length will be truncated. We train the S2S + ATTN + COPY model on the original text-to-SQL dataset for each of the sub-tasks in PAGSAS. The hyperparameters for the second stage are the same as in Appendix A.3.2.

During inference, GPT2 first completes the prefix. The completed questions are then fed to S2S + ATTN + COPY model. We multiply the probabilities returned by the two stages and select the top K SQL predictions.

A.3.4 T5 model

We finetune the T5-base (Raffel et al., 2020) model on each sub-task in PAGSAS. We set the learning rate to $3 \cdot 10^{-4}$ and the maximum sequence length to 512.

A.3.5 EMB-RTR model

We finetune RoBERTa-base (Liu et al., 2019) encoder model on PAGSAS. Two independent RoBERTa encoders are used to obtain the embeddings of the question prefix and SQL separately. The dot-product of question prefix and SQL embeddings is treated as their similarity score. We treat SQL queries that match a question prefix as the positive examples and sample SQL queries that do not match the prefix as the negative examples in the training process. We make the number of negative examples 5 times the number of positive examples. We adopt a learning rate of $2 \cdot 10^{-5}$ to finetune the two RoBERTa encoders. After the training process, all historical SQL embeddings are cached, so we only need to run the RoBERTa encoder once for each given question prefix. We rank all historical SQL queries by the dot-product of their embeddings with the given question prefix embedding.

A.3.6 CLS-RTR model

We use pre-trained RoBERTa-base (Liu et al., 2019) as our classification model. The model takes the
| #  | SQL-template | Question-template |
|----|-------------|-------------------|
| 1  | {SELECT0} {FROM} WHERE {COLUMN0} {OP0} {VALUE0} | Find {SELECT0} whose {COLUMN0} {OP0} {VALUE0}. |
| 2  | {SELECT0} {FROM} | Find {SELECT0}. |
| 3  | {SELECT0} {FROM} WHERE {COLUMN0} {OP0} {VALUE0} AND {COLUMN1} {OP1} {VALUE1} | Find {SELECT0} whose {COLUMN0} {OP0} {VALUE0} and {COLUMN1} {OP1} {VALUE1}. |
| 4  | {SELECT0} {FROM} GROUP BY {COLUMN0} | For each {COLUMN0} what is {SELECT0}. |
| 5  | {SELECT0} {FROM} GROUP BY {COLUMN0} ORDER BY {AGG0} ( * ) {SC0} LIMIT {VALUE0} | What {COLUMN0} are the top {VALUE0} {AGG0} {SC0}. List {SELECT0}. |
| 6  | {SELECT0} {FROM} ORDER BY {COLUMN0} {SC0} LIMIT {VALUE0} | List {SELECT0} that are the top {VALUE0} ranked by {COLUMN0} {SC0}. |
| 7  | {SELECT0} {FROM} GROUP BY {COLUMN0} HAVING {AGG0} ( * ) {OP0} {VALUE0} | For {COLUMN0} whose {AGG0} {OP0} {VALUE0}. List {SELECT0}. |
| 8  | {SELECT0} {FROM} WHERE {COLUMN0} {OP0} ( {SELECT1} {FROM} ) | Find {SELECT0} that have {COLUMN0} {OP0} {SELECT1}. |
| 9  | {SELECT0} {FROM} ORDER BY {COLUMN0} {SC0} | Show {SELECT0} with {COLUMN0} {SC0}. |
| 10 | {SELECT0} {FROM} WHERE {COLUMN0} {OP0} {VALUE0} INTERSECT {SELECT1} {FROM} WHERE {COLUMN1} {OP1} {VALUE1} | Show {SELECT0} ( or {SELECT1} ) that have both {COLUMN0} {OP0} {VALUE0} and {COLUMN1} {OP1} {VALUE1}. |
| 11 | {SELECT0} {FROM} WHERE {COLUMN0} {OP0} {VALUE0} OR {COLUMN1} {OP1} {VALUE1} | Find {SELECT0} whose {COLUMN0} {OP0} {VALUE0} or {COLUMN1} {OP1} {VALUE1}. |
| 12 | {SELECT0} {FROM} WHERE {COLUMN0} NOT IN ( {SELECT1} {FROM} ) | Find {SELECT0} whose {COLUMN0} is not in {SELECT1}. |
| 13 | {SELECT0} {FROM} ORDER BY {COLUMN0} | What are {SELECT0} ordered by {COLUMN0}. |
| 14 | {SELECT0} {FROM} WHERE {COLUMN0} {OP0} {VALUE0} AND ( {COLUMN1} {OP1} {VALUE1} ) {OP2} {VALUE2} | Find {SELECT0} whose {COLUMN0} {OP0} {VALUE0} and ( {COLUMN1} {OP1} {VALUE1} ) {OP2} {VALUE2}. |
| 15 | {SELECT0} {FROM} WHERE {COLUMN0} {OP0} {VALUE0} GROUP BY {COLUMN1} | For each {COLUMN1} whose {COLUMN0} {OP0} {VALUE0} list {SELECT0}. |

Table 6: Top 15 SQL and canonical question templates.
concatenation of a question prefix and a SQL candidate as input to predict whether the SQL matches the question prefix intention. Given a question prefix during inference, the classification model needs to be run on all historical SQL queries to select the top $K$ SQL queries, which makes the running time significantly larger than all the other models. Similar to EMB-RTR, we choose the learning rate as $2 \cdot 10^{-5}$ and the ratio of negative examples over positive examples is set to 5.

### A.4 Supplementary Results

Table 8 reports models’ RECALL@10, MRR@10 and SAVE@10 scores in percentage for each sub-task in PAGSAS on question split. Table 9 reports the scores for generative models on all sub-tasks except Scholar in PAGSAS on SQL query split. Because models only perform $0 \sim 1\%$ on all metrics for sub-task Scholar, we do not report the results in Table 9.
Table 7: Statistics for PAGSAS on question splits ("Q" columns) and SQL query splits ("S" columns) on train, validation (dev) and test split.

| Model       | Advising | ATIS | GeoQuery | Scholar | Spider |
|-------------|----------|------|----------|---------|--------|
|             | Q        | S    | Q        | S       | Q      | S      |
| PF-S2S      | 18       | 14   | 9        | 13      | 8      | 9      |
| + ATTN      | 19       | 14   | 9        | 12      | 8      | 9      |
| + COPY      | 25       | 19   | 15       | 11      | 9      | 10     |
| QAC-S2S     | 11       | 7    | 8        | 11      | 6      | 11     |
| T5          | 49       | 39   | 32       | 12      | 8      | 6      |
| EMB-RTR     | -        | -    | -        | -       | -      | -      |
| CLS-RTR     | 23       | 15   | 11       | 8       | 4      | 5      |
|             |          |      |          |         | 22     | 13     |
|             |          |      |          |         | 23     | 15     |
|             |          |      |          |         | 63     | 52     |
|             |          |      |          |         | 61     |        |

Table 8: RECALL@10 (R), MRR@10 (M) and SAVE@10 (S) in percentage generative models on SQL query split. We omit the results for sub-tasks Scholar because models perform 0–1% for all metrics. We use "-" to denote scores < 1%.

Figure 10: Percentage of types of entities detected by SpaCy (Honnibal and Montani, 2017) in the omitted text for question split and SQL query split in PAGSAS.

Figure 12 is similar to Figure 3 but is on SQL query split. Figure 12 also shows that T5’s RECALL@5 scores are negatively correlated with the length to complete (number of omitted tokens) but possess no monotonic relationships with either the prefix length (number of tokens in the prefix) or SQL length (number of tokens in SQL queries).
Figure 11: The framework of our generation-based and retrieval-based models. Generation models predict SQL queries based on the input sequence directly, while retrieval models predict a score for each historical SQL and then rank the scores to recommend SQL queries.

Figure 12: T5’s (a) RECALL@5 v.s. number of omitted tokens, (b) RECALL@5 v.s. prefix length (number of tokens in prefix), (c) RECALL@5 v.s. SQL length (number of tokens in SQL) on SQL query split for each sub-task in PAGSAS. We plot RECALL@5 corresponding to ≥ 50 and ≥ 100 examples for sub-tasks other than Spider and Scholar, respectively. RECALL@5 is negatively correlated with number of omitted tokens and possesses no monotonic relationships with either prefix length or SQL length.