Evaluating the Text-to-SQL Capabilities of Large Language Models

Anonymous ACL submission

Abstract

We perform an empirical evaluation of Text-to-SQL capabilities of the Codex language model. We find that, without any finetuning, Codex is a strong baseline on the Spider benchmark; we also analyze the failure modes of Codex in this setting. Furthermore, we demonstrate on the GeoQuery and Scholar benchmarks that a small number of in-domain examples provided in the prompt enables Codex to perform better than state-of-the-art models finetuned on such few-shot examples. We provide code here.

1 Introduction

Translating natural language questions to SQL queries (Text-to-SQL) is an important business problem which has seen significant research interest. A common approach to this task involves training a model to produce a SQL query when given a question, a database schema, and possibly database content as inputs. A clear trend in this area is to finetune models pretrained on natural language; notably, performance significantly improves as larger pretrained models are used (Shaw et al., 2021; Scholak et al., 2021).

Recent results from the broader field demonstrate that simply scaling training data and model size for generative language models brings advanced capabilities, such as few-shot learning without finetuning (GPT-3, Brown et al., 2020) and code generation (Codex, Chen et al., 2021). In this work we study if such models are already competitive Text-to-SQL solutions without any further finetuning on task-specific training data, evaluating Codex and GPT-3 models of different sizes with varied prompts on Text-to-SQL benchmarks.

We find that Codex achieves a competitive performance of up to 67% execution accuracy on the Spider development set. We analyze the predicted queries that automatic evaluation judged as wrong and find that many of them would be judged correct by humans, whereas others could likely be fixed within the no-finetuning paradigm. Lastly, using GeoQuery and Scholar benchmarks we show that adapting Codex to a specific domain by prompting it with few examples can be more effective than fine-tuning a smaller language model on the same examples.

2 Experimental Setup

Models

Our evaluation focuses on the models accessible via the OpenAI API: GPT-3 (in the ascending ada, babbage, curie and davinci sizes) and Codex (in the ascending cushman-codex and davinci-codex sizes). These are generative language models which perform next-token prediction during training and inference; GPT-3 is trained on a diverse set of sources from the internet, and Codex is further finetuned on code from GitHub. We compare GPT-3 and Codex against methods from Shaw et al. (2021) using the T5 encoder-decoder model. Starting from public checkpoints pretrained on Common Crawl, the T5 model is finetuned on

Table 1: Best Spider development set performance across models, as measured by percentage of predictions which are valid SQL (VA), execution accuracy (EX), test-suite accuracy (TS). Models marked with * use database content. T5 results are from Scholak et al. (2021), BRIDGE v2 results are from Lin et al. (2020).

<table>
<thead>
<tr>
<th>Model</th>
<th>VA</th>
<th>EX</th>
<th>TS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finetuned</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5-base</td>
<td>72.7</td>
<td>57.9</td>
<td>54.5</td>
</tr>
<tr>
<td>T5-large</td>
<td>84.1</td>
<td>67.2</td>
<td>61.4</td>
</tr>
<tr>
<td>T5-3B</td>
<td>87.6</td>
<td>71.4</td>
<td>65.7</td>
</tr>
<tr>
<td>T5-3B*</td>
<td>88.2</td>
<td>74.4</td>
<td>68.3</td>
</tr>
<tr>
<td>T5-3B + PICARD*</td>
<td>97.8</td>
<td>79.1</td>
<td>71.7</td>
</tr>
<tr>
<td>BRIDGE v2*</td>
<td>97.8</td>
<td>79.1</td>
<td>71.7</td>
</tr>
<tr>
<td>Inference-only</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT-3 ada</td>
<td>33.8</td>
<td>2.3</td>
<td>0.3</td>
</tr>
<tr>
<td>GPT-3 babbage</td>
<td>48.8</td>
<td>5.7</td>
<td>3.9</td>
</tr>
<tr>
<td>GPT-3 curie</td>
<td>70.9</td>
<td>12.6</td>
<td>8.3</td>
</tr>
<tr>
<td>GPT-3 davinci</td>
<td>65.0</td>
<td>26.3</td>
<td>21.7</td>
</tr>
<tr>
<td>Codex cushman*</td>
<td>86.3</td>
<td>63.7</td>
<td>53.0</td>
</tr>
<tr>
<td>Codex davinci*</td>
<td>91.6</td>
<td>67.0</td>
<td>55.1</td>
</tr>
</tbody>
</table>

See Appendix A.2 for a discussion on parameter counts.
Spider to predict the output SQL, conditioned on the question and schema. The 3B parameter T5 model is currently the state-of-the-art on Spider when combined with constrained inference using the PICARD algorithm (Scholak et al., 2021). We also compare to BRIDGE v2 (Lin et al., 2020), a sequence-to-sequence model based on BERT.

### Zero-Shot Experiments
We use the Spider benchmark (Yu et al., 2019) for cross-domain Text-to-SQL. We report performance using percentage of development set predictions which are valid (executable) SQLite SQL, execution accuracy, and test-suite execution accuracy. The latter metric was proposed by Zhong et al. (2020) to measure semantic equivalence of SQL queries written in different styles, which is essential when comparing Codex to models trained on Spider. We address concerns around possible memorization of Spider data by Codex in Appendix A.5.

### Few-Shot Experiments
We re-purpose the question-splits of the GeoQuery and Scholar datasets (Zelle and Mooney, 1996; Iyer et al., 2017; Finegan-Dollak et al., 2018) to perform experiments in a few-shot setting. The examples in these datasets are grouped by query templates. Examples corresponding to the same template have the same SQL query structure, but may have different English questions and SQL literals. To define the few-shot task, we first sort the templates by their frequency in the training set. In the n-shot setting we then use one random example for each of the n most frequent templates.

### Prompts
We use six prompt structures in our experiments (examples provided in Appendix C).

- **Question** provides no database information and just includes the question as a SQL comment. **API Docs** follows the style of the Text-to-SQL example in Codex documentation and includes a schema in a comment style which does not conform to SQLite standards. **Select X** includes in comments the results of executing a `SELECT * FROM T LIMIT X` query on each table, including schemas via column headers. **Create Table** includes the CREATE TABLE commands for each table, including column type and foreign key declarations. **Create Table + Select X**\(^2\) is a combination of the preceding two prompt formats. Finally, **Fewshot** additionally includes question-query pairs.

### 3 Zero-Shot Results
We present results for different model sizes in Table 1 and for different prompt styles in Table 2. Full results are available in Table 4 in Appendix B.

**Codex provides a strong baseline for Text-to-SQL tasks** In Table 1 the best performing model (davinci-codex, Create Table + Select 3) achieves 67% execution accuracy and 56.5% test suite execution accuracy on Spider. This is comparable to the performance of the BRIDGE v2 (Lin et al., 2020) model which achieved a (then) state-of-the-art 68% execution accuracy in November 2020.

**Prompt design is critical for performance** As seen in Table 2, providing the question alone results in a low 8.3% execution accuracy. There is a progressive improvement to 56.8% as schema information is introduced in API Docs, to 59.9% when valid SQL and foreign key information is used in Create Table, and to 67.0% when database content is introduced with Create Table + Select 3.

**More database content can harm performance** In Table 2 we observe that for the Select Limit X prompts there is a negligible change in performance when adding more rows. By contrast, Create Table + Select Limit X prompt accuracy peaks with 3 rows before significantly decreasing in performance as more rows are added.

**Diminishing returns for Codex model size** While GPT-3 performance significantly benefits from increased model size, the davinci-codex model does not perform drastically better than cushman-codex. Full results in Table 4 in Appendix B show cushman-codex generally being within 1 percentage point of davinci-codex for the

<table>
<thead>
<tr>
<th>Prompt</th>
<th>VA</th>
<th>EX</th>
<th>TS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>14.0</td>
<td>8.3</td>
<td>8.2</td>
</tr>
<tr>
<td>API Docs</td>
<td>83.8</td>
<td>56.8</td>
<td>47.5</td>
</tr>
<tr>
<td>Select 1</td>
<td>86.3</td>
<td>60.9</td>
<td>52.0</td>
</tr>
<tr>
<td>Select 3</td>
<td>85.8</td>
<td>60.3</td>
<td>52.2</td>
</tr>
<tr>
<td>Select 5</td>
<td>85.2</td>
<td>60.5</td>
<td>51.5</td>
</tr>
<tr>
<td>Select 10</td>
<td>86.0</td>
<td>60.8</td>
<td>51.2</td>
</tr>
<tr>
<td>Create Table</td>
<td>89.8</td>
<td>59.9</td>
<td>50.0</td>
</tr>
<tr>
<td>+ Select 1</td>
<td>92.5</td>
<td>64.8</td>
<td>53.7</td>
</tr>
<tr>
<td>+ Select 3</td>
<td>91.6</td>
<td>67.0</td>
<td>55.1</td>
</tr>
<tr>
<td>+ Select 5</td>
<td>91.0</td>
<td>65.3</td>
<td>53.9</td>
</tr>
<tr>
<td>+ Select 10</td>
<td>91.2</td>
<td>63.3</td>
<td>52.4</td>
</tr>
</tbody>
</table>

---

\(^2\)Only the davinci-codex model can evaluate Create Table + Select X prompts with more than 1 row, due to it’s expanded 4096-token prompt window compared to the 2048-token window of all other models. In addition, GPT-3 models preprocess whitespace tokens less efficiently than Codex models, and therefore cannot evaluate Create Table + Select X prompts at all.
14%  **Shortcuts**  
**NL** What is the number of car models created by the car maker American Motor Company?  
**Gold** SELECT count(*) FROM CAR_MAKERS AS T1 JOIN MODEL_LIST AS T2 ON T1.Id = T2.Maker WHERE T1.FullName = 'American Motor Company';  
**Pred** SELECT COUNT(Model) FROM model_list WHERE Maker = 1;  

**Gold** SELECT City, Country FROM AIRPORTS WHERE AirportName = "Alton"  
**Pred** SELECT City, Country FROM airports WHERE AirportCode = 'ALN';  

---

8%  **SELECT Extra Columns**  
**NL** List names of conductors in descending order of years of work.  
**Gold** SELECT Name FROM conductor ORDER BY Year_of_Work DESC  
**Pred** SELECT Name, Year_of_Work FROM conductor ORDER BY Year_of_Work DESC;  

---

5%  **SELECT Convention**  
**NL** What are all the makers and models?  
**Gold** SELECT Maker , Model FROM MODEL_LIST;  
**Pred** SELECT DISTINCT car_makers.Maker, model_list.Model FROM car_makers JOIN model_list ON car_makers.Id = model_list.Maker;  

---

3.1 Error Analysis  
We focus our error analysis on the davinci-codex model with Create Table + Select 3 prompt, and present a breakdown of prediction types in Table 3 and examples of errors in Figure 1. Our error categories were chosen to surface the most interesting Codex-specific behaviours we observed amongst the errors made. We randomly selected and annotated 100 predictions which were valid SQL yet were judged incorrect by test-suite evaluation.

We first consider **Semantic Incorrect** behaviours, which Spider evaluation and the human annotator both view as incorrect predictions.  
**Shortcut** errors are where Codex made use of either specific table values or “world knowledge” from GPT-3 pretraining, while the ground-truth query contained the exact literals from the question.  
**GROUP BY Convention** errors are where Codex incorrectly groups on a non-primary-key column (such as a name or title column).

We also consider **Ambiguous Correct** behaviours which are semantically different from the gold query and are therefore judged as incorrect by Spider evaluation, but which the human annotator viewed as being an acceptable SQL translation of the given question.  
**SELECT Convention** errors are where Codex selects a different column than the per-database convention of the gold queries (such as name instead of ID).  
**SELECT Extra Columns** errors are where Codex includes additional useful columns in its query beyond what the gold query includes.  
**Argmax** errors are where Codex differs from the gold query in how a min/max resolution (such as “youngest singer”) is handled for ties.

We observe in Table 3 that a significant 31% of valid yet erroneous predictions are penalized by Spider evaluation as being incorrect though a human annotator viewed them as acceptable solutions. Future work could be to investigate to what extent one can control the behaviour of Codex. This could allow to fix these ambiguous errors, either by prompt design or using a few examples.

---

**Table 3: Breakdown of prediction annotations over Spider development set for the davinci-codex model with Create Table + Select 3 prompt.**  
% is percentage of all predictions, E% is percentage of manually annotated erroneous queries (see Section Section 3.1 for details).  

<table>
<thead>
<tr>
<th>Annotation</th>
<th>%</th>
<th>E%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-Suite Correct</td>
<td>55.1</td>
<td>–</td>
</tr>
<tr>
<td>Semantic Incorrect</td>
<td>25.2</td>
<td>69</td>
</tr>
<tr>
<td>– Shortcuts</td>
<td>5.1</td>
<td>14</td>
</tr>
<tr>
<td>– GROUP BY Convention</td>
<td>1.5</td>
<td>4</td>
</tr>
<tr>
<td>– Other</td>
<td>18.6</td>
<td>51</td>
</tr>
<tr>
<td><strong>Ambiguous Correct</strong></td>
<td>17.5</td>
<td>51</td>
</tr>
<tr>
<td>– SELECT Extra Columns</td>
<td>2.9</td>
<td>8</td>
</tr>
<tr>
<td>– SELECT Convention</td>
<td>1.8</td>
<td>5</td>
</tr>
<tr>
<td>– Argmax</td>
<td>1.5</td>
<td>4</td>
</tr>
<tr>
<td>– Other</td>
<td>5.1</td>
<td>14</td>
</tr>
<tr>
<td>Invalid SQL</td>
<td>8.4</td>
<td>–</td>
</tr>
<tr>
<td>– Ambiguous column name</td>
<td>1.9</td>
<td>–</td>
</tr>
<tr>
<td>– No such column</td>
<td>4.5</td>
<td>–</td>
</tr>
</tbody>
</table>
4 Few-Shot

We investigate whether Codex can perform few-shot Text-to-SQL. As described in Section 2, we re-purpose the GeoQuery and Scholar datasets in a few-shot setting. It is well known that models trained on Spider transfer poorly to other single-database Text-to-SQL datasets (Suhr et al., 2020) in a zero-shot setting. Studying few-shot Text-to-SQL on GeoQuery and Scholar should show to what extent models are able to leverage a small amount of examples to effectively adapt to a new domain.

**Baseline** The baseline is a T5-3B model that was finetuned on Spider, reaching 71% exact-match accuracy on Spider validation set. The model is then further finetuned on the new domain – GeoQuery or Scholar. The learning rate for domain-specific-finetuning was selected in the 20-shot setting among \([0.1, 0.2, 0.5, 1, 2] \cdot 10^{-5}\), based on the best validation set performance after 300 steps. We use batch-size 1024, such that all the few-shot examples fit in the same batch.

**Codex** Building on the Create Table + Select X prompt, we append \(n\) question-query examples to the input in an \(n\)-shot setting. An example of this prompt is provided in Figure 11. All samples are generated using greedy decoding, with temperature 0. Note that for a given \(n\)-shot setting, the baseline and Codex use the same set of support examples. These examples are in the prompt for Codex, and used to finetune the baseline on the new domain. Given the limited window-size of API models, on GeoQuery we can feed up to 40 support examples to davinci-codex, and up to 10 examples to cushman-codex and GPT-3 models. On Scholar the queries are longer and the schema more complex – we fit only 10 examples in the prompt of davinci-codex, 5 for cushman-codex, and none at all for GPT-3 models.

4.1 Results

Figure 2 shows test-suite accuracies on the Scholar and GeoQuery datasets. The baseline reaches 85.7% test-set performance when trained on the complete GeoQuery training set (549 examples). Respectively, it reaches 87.2% test accuracy when trained on the whole Scholar training set (499 examples). This simple baseline is a very competitive model when considering the entire datasets. However Figure 2 shows that it is largely beaten by Codex in few-shot settings. In a zero-shot setting, both davinci-codex and cushman-codex already beat the baseline on GeoQuery. We speculate that Codex performs well here because it uses the same argmax convention as the GeoQuery dataset, which is different than the convention used in Spider. With up to 40 examples in the prompt, davinci-codex outperforms a T5-3B model finetuned on these same examples by a large margin, whereas GPT-3 davinci performs quite poorly on this task. On the other hand, the T5 model outperforms Codex in a zero-shot setting on Scholar. In 5 and 10-shot settings, Codex shows better adaptation from these few samples and beats the T5 baseline.

5 Conclusion

We demonstrated that generative language models trained on code provide a strong baseline for Text-to-SQL. We also provided analysis of failure modes for these models, which we hope guides further prompt design (whether few-shot or through natural language instructions) in this setting. Finally, we showed that prompt-based few-shot learning with these models performs competitively with finetuning-based few-shot learning of smaller models. A clear direction for future work is to evaluate the benefits of finetuning with Codex models.
References


Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingming Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2019. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task.


A API Details

At time of writing, the OpenAI API was accessible at https://openai.com/api/. The example from which our API Docs prompt draws from can be found at https://beta.openai.com/examples/default-sql-translate.

A.1 Hyperparameters

We sample 200 tokens from GPT-3 and Codex with temperature 0, with the following strings used as stop tokens to halt generation: “--”, “

”, “;”, “#”.

A.2 Parameter Counts

Parameter counts for OpenAI API models are not openly available. Gao (2021) evaluated API GPT-3 models across a variety of language modelling tasks to compare to published results in Brown et al. (2020), finding that “Ada, Babbage, Curie and Davinci line up closely with 350M, 1.3B, 6.7B, and 175B respectively”. We presume that the davinci-codex model is the same size as the GPT-3 davinci model; cushman-codex is a new model name so we can only guess that it is of a similar (but not the same) size to GPT-3 curie. Nevertheless these remain guesses which should not be relied on.

A.3 Model Versioning

The exact models served through the OpenAI API may vary over time. We verified that for each model type, only a single model version was used to generate results. These versions are ada:2020-05-03, babbage:2020-05-03, curie:2020-05-03, davinci:2020-05-03, cushman-codex:2021-08-03, davinci-codex:2021-08-03.

A.4 Fine-tuning

In Table 4 we include preliminary results from fine-tuning GPT-3 models on the Spider training set. We used the full training set, and the default fine-tuning settings of 4 epochs, a batch size of 8, and a learning rate multiplier of 0.1. We did not perform a hyperparameter sweep due to the significant cost this would incur.

A.5 Memorization

The Spider development set is available on GitHub, and is therefore possibly in the training set of Codex. We believe that this does not manifest as memorization for our results however, for the following reasons.

Evaluation data on Spider’s repo is formatted differently to our prompts. Most related is the dev.sql file, which contains evaluation question-query pairs in the following format:

```
Question 1: ...
SQL: ...
...
```

This resembles but isn’t identical to our “Question” prompt. We prompted Codex with verbatim fragments of this file and generations failed to replicate any file contents. Our “Question” prompt has very poor performance - hardly an indication of memorization from dev.sql. Furthermore, most of Codex’s performance is due to including in the prompt the schemas (see Table 2), which are not present in dev.sql.

As well, Codex prediction style is very different to evaluation gold queries. Gold queries make use of a consistent table aliasing strategy (using T1, T2, etc.) which we never see with Codex (see Figure 3 for example comparisons).

Furthermore, in Table 4 we reported performance for all models on spider-realistic (Deng et al., 2021), a modification of the spider evaluation set that removes column name references in questions. We observe a similar trend in performance across models as on spider (the consistent performance drop on spider-realistic is expected due to the difficulty of the updated dataset). Memorization cannot account for the performance observed, as spider-realistic is not publicly available on GitHub.

Finally, Ziegler (2021) studied memorization in Copilot, a derivative of the Codex models, and found that “Copilot can quote a body of code verbatim, but that it rarely does so, and when it does, it mostly quotes code that everybody quotes, and mostly at the beginning of a file". Spider evaluation data is rare on GitHub, and we use long contexts in our prompts that significantly differ from the files on GitHub.

A.6 Choice of Spider Evaluation Set

We chose not to evaluate on the held-out test set of Spider, as this could not be done offline - it would instead require sending these held-out examples through the API to OpenAI, which risks inadvertently leaking them for retraining of Codex.
### B  Additional Tables and Figures

<table>
<thead>
<tr>
<th>Engine</th>
<th>Prompt</th>
<th>VA</th>
<th>EX</th>
<th>TS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPT-3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ada</td>
<td>Question</td>
<td>1.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Docs</td>
<td>3.4</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>1 Row</td>
<td>40.1</td>
<td>1.1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Schema</td>
<td>33.8</td>
<td>2.3</td>
<td>0.3</td>
</tr>
<tr>
<td>babbage</td>
<td>Question</td>
<td>4.4</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>Docs</td>
<td>22.5</td>
<td>1.0</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>1 Row</td>
<td>56.0</td>
<td>5.1</td>
<td>3.9</td>
</tr>
<tr>
<td></td>
<td>Schema</td>
<td>48.8</td>
<td>5.7</td>
<td>3.9</td>
</tr>
<tr>
<td>curie</td>
<td>Question</td>
<td>9.0</td>
<td>2.9</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>Docs</td>
<td>25.2</td>
<td>7.4</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>1 Row</td>
<td>70.6</td>
<td>10.8</td>
<td>7.6</td>
</tr>
<tr>
<td></td>
<td>Schema</td>
<td>70.9</td>
<td>12.6</td>
<td>8.3</td>
</tr>
<tr>
<td>davinci</td>
<td>Schema</td>
<td>65.0</td>
<td>26.3</td>
<td>21.7</td>
</tr>
<tr>
<td><strong>Finetuned GPT-3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ada</td>
<td>Schema</td>
<td>27.5</td>
<td>20.2</td>
<td>19.1</td>
</tr>
<tr>
<td>babbage</td>
<td>Schema</td>
<td>47.2</td>
<td>34.8</td>
<td>31.9</td>
</tr>
<tr>
<td>curie</td>
<td>Schema</td>
<td>66.9</td>
<td>51.3</td>
<td>46.9</td>
</tr>
<tr>
<td><strong>Codex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cushman</td>
<td>Question</td>
<td>11.3</td>
<td>8.5</td>
<td>8.3</td>
</tr>
<tr>
<td></td>
<td>Docs</td>
<td>83.8</td>
<td>53.2</td>
<td>43.5</td>
</tr>
<tr>
<td></td>
<td>1 Row</td>
<td>84.7</td>
<td>59.6</td>
<td>48.5</td>
</tr>
<tr>
<td></td>
<td>3 Rows</td>
<td>82.9</td>
<td>60.3</td>
<td>49.4</td>
</tr>
<tr>
<td></td>
<td>5 Rows</td>
<td>83.6</td>
<td>61.5</td>
<td>50.4</td>
</tr>
<tr>
<td></td>
<td>Schema</td>
<td>88.3</td>
<td>62.1</td>
<td>53.1</td>
</tr>
<tr>
<td></td>
<td>+ 1 Row</td>
<td>86.3</td>
<td>63.7</td>
<td>53.0</td>
</tr>
<tr>
<td>davinci</td>
<td>Question</td>
<td>14.0</td>
<td>8.3</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>Docs</td>
<td>83.8</td>
<td>56.8</td>
<td>47.5</td>
</tr>
<tr>
<td></td>
<td>1 Row</td>
<td>86.3</td>
<td>60.9</td>
<td>52.0</td>
</tr>
<tr>
<td></td>
<td>3 Rows</td>
<td>85.8</td>
<td>60.3</td>
<td>52.2</td>
</tr>
<tr>
<td></td>
<td>5 Rows</td>
<td>85.2</td>
<td>60.5</td>
<td>51.5</td>
</tr>
<tr>
<td></td>
<td>10 Rows</td>
<td>86.0</td>
<td>60.8</td>
<td>51.2</td>
</tr>
<tr>
<td></td>
<td>Schema</td>
<td>89.8</td>
<td>59.9</td>
<td>50.0</td>
</tr>
<tr>
<td></td>
<td>+ 1 Row</td>
<td>92.5</td>
<td>64.8</td>
<td>53.7</td>
</tr>
<tr>
<td></td>
<td>+ 3 Rows</td>
<td>91.6</td>
<td>67.0</td>
<td>55.1</td>
</tr>
<tr>
<td></td>
<td>+ 5 Rows</td>
<td>91.0</td>
<td>65.3</td>
<td>53.9</td>
</tr>
<tr>
<td></td>
<td>+ 10 Rows</td>
<td>91.2</td>
<td>63.3</td>
<td>52.4</td>
</tr>
</tbody>
</table>

Table 4: Performance on Spider across all evaluated models and prompts, as measured by percentage of predictions which are valid/executable SQL (VA), execution accuracy (EX), test-suite accuracy (TS). Main results are on the Spider development set, results in parantheses are on Spider-Realistic (Deng et al., 2021), a modified subset of the Spider development set with explicit references to column names removed from questions.
14% Shortcuts
Q What are the template ids with template type description "Presentation".
Gold SELECT T2.template_id FROM Ref_template_types AS T1 JOIN Templates AS T2 ON T1.template_type_code = T2.template_type_code WHERE T1.template_type_description = "Presentation"
Pred SELECT Template_ID FROM Templates WHERE Template_Type_Code = 'PPT';

8% SELECT Extra Columns
NL Which city has the most frequent destination airport?
Gold SELECT T1.City FROM AIRPORTS AS T1 JOIN FLIGHTS AS T2 ON T1.AirportCode = T2.DestAirport GROUP BY T1.City ORDER BY count(*) DESC LIMIT 1
Pred SELECT City, COUNT(*) AS 'Number of Destinations' FROM airports JOIN flights ON airports.AirportCode = flights.DestAirport GROUP BY City ORDER BY COUNT(*) DESC LIMIT 1;

5% SELECT Convention
NL Which airlines have a flight with source airport AHD?
Gold SELECT T1.Airline FROM AIRLINES AS T1 JOIN FLIGHTS AS T2 ON T1.uid = T2.Airline WHERE T2.SourceAirport = 'AHD'
Pred SELECT Airline FROM flights WHERE SourceAirport = 'AHD';

4% GROUP BY Convention
NL Which makers designed more than 3 car models? List full name and the id.
Gold SELECT T1.FullName , T1.Id FROM CAR_MAKERS AS T1 JOIN MODEL_LIST AS T2 ON T1.Id = T2.Maker GROUP BY T1.Id HAVING count(*) > 3;
Pred SELECT car_makers.FullName, car_makers.Id FROM car_makers JOIN model_list ON car_makers.Id = model_list.Maker GROUP BY car_makers.FullName HAVING COUNT(model_list.Model) > 3;

4% Argmax
NL Give the name, year of independence, and surface area of the country that has the lowest population.
Gold SELECT Name, SurfaceArea, IndepYear FROM country ORDER BY Population LIMIT 1
Pred SELECT Name, IndepYear, SurfaceArea FROM country WHERE Population = (SELECT MIN(Population) FROM country);

51% Other - Semantic Incorrect
NL Which countries have either English or Dutch as an official language?
Pred SELECT * FROM countrylanguage WHERE IsOfficial = 'T' OR IsOfficial = 'F';

4% Argmax
NL Show the name and the release year of the song by the youngest singer.
Gold SELECT song_name , song_release_year FROM singer ORDER BY age LIMIT 1
Pred SELECT Name, Song_release_year FROM singer WHERE Age = (SELECT MIN(Age) FROM singer);

NL What is the maximum miles per gallon of the car with 8 cylinders or produced before 1980 ?
Gold select max(mpg) from cars_data where cylinders = 8 or year < 1980
Pred SELECT MAX(MPG) FROM cars_data WHERE Cylinders = 8 AND Year < 1980;

14% Other - Ambiguous Correct
NL Find the first name and age of students who have a pet.
Gold SELECT DISTINCT T1.fname , T1.age FROM student AS T1 JOIN has_pet AS T2 ON T1.stuid = T2.stuid
Pred SELECT Fname, Age FROM Student WHERE StuID IN (SELECT StuID FROM Has_Pet);

NL How many countries speak both English and Dutch?
Pred SELECT COUNT(*) FROM countrylanguage WHERE Language = 'Dutch' AND CountryCode IN (SELECT CountryCode FROM countrylanguage WHERE Language = 'English');

NL What are the Asian countries which have a population larger than that of any country in Africa?
Gold SELECT Name FROM country WHERE Continent = "Asia" AND population > (SELECT min(population) FROM country WHERE Continent = "Africa")
Pred SELECT Name FROM country WHERE Continent = 'Asia' AND Population > (SELECT MAX(Population) FROM country WHERE Continent = 'Africa');

Figure 3: Additional examples of error types, as made by davinci-codex model with Create Table + Select 3 prompt. NL stands for natural language question. Percentage indicates the percentage of errors which are of the given type.
10-shot examples

what states does the missouri river run through

SELECT RIVERalias0.TRAVERSE FROM RIVER AS RIVERalias0 WHERE RIVERalias0.RIVER_NAME = "missouri" ;

what is the size of texas

SELECT STATEalias0.AREA FROM STATE AS STATEalias0 WHERE STATEalias0.STATE_NAME = "texas" ;

what are the major cities in texas

SELECT CITYalias0.CITY_NAME FROM CITY AS CITYalias0 WHERE CITYalias0.STATE_NAME = "texas" ;

what is the capital of pennsylvania

SELECT STATEalias0.CAPITAL FROM STATE AS STATEalias0 WHERE STATEalias0.STATE_NAME = "pennsylvania" ;

what is the biggest city in nebraska

SELECT CITYalias0.CITY_NAME FROM CITY AS CITYalias0 WHERE CITYalias0.STATE_NAME = "nebraska" ;

what is the biggest city in nebraska

SELECT CITYalias0.CITY_NAME FROM CITY AS CITYalias0 WHERE CITYalias0.STATE_NAME = "nebraska" ;

which state is kalamazoo in

SELECT CITYalias0.CITY_NAME FROM CITY AS CITYalias0 WHERE CITYalias0.CITY_NAME = "kalamazoo" ;

name all the rivers in colorado

SELECT RIVERalias0.RIVER_NAME FROM RIVER AS RIVERalias0 WHERE RIVERalias0.TRAVERSE = "colorado" ;

what states border missouri

SELECT BORDER_INFOalias0.BORDER FROM BORDER_INFO AS BORDER_INFOalias0 WHERE BORDER_INFOalias0.STATE_NAME = "missouri" ;

how many people live in new mexico

SELECT STATEalias0.POPULATION FROM STATE AS STATEalias0 WHERE STATEalias0.STATE_NAME = "new mexico" ;

- what state has the smallest population

SELECT STATEalias0.STATE_NAME FROM STATE AS STATEalias0 WHERE STATEalias0.POPULATION = (SELECT MIN(STATEalias1.POPULATION) FROM STATE AS STATEalias1) ;

- what is the population of the state with the largest area

SELECT STATEalias0.POPULATION FROM STATE AS STATEalias0 WHERE STATEalias0.AREA = (SELECT MAX(STA}
C  Example Prompts

What is Kyle's id? | network_1 | highschooler : id, name ( Kyle ), grade | friend : student_id, friend_id | likes : student_id, liked_id

Figure 5: Example input for baseline T5 models.

-- Using valid SQLite, answer the following questions.
-- What is Kyle’s id?
SELECT

Figure 6: Example prompt for Question.

### SQLite SQL tables, with their properties:
#
# Highschooler(ID, name, grade)
# Friend(student_id, friend_id)
# Likes(student_id, liked_id)
#
### What is Kyle’s id?
SELECT

Figure 7: Example prompt for API Docs.
CREATE TABLE Highschooler(
    ID int primary key,
    name text,
    grade int)

CREATE TABLE Friend(
    student_id int,
    friend_id int,
    primary key (student_id,friend_id),
    foreign key(student_id) references Highschooler(ID),
    foreign key (friend_id) references Highschooler(ID)
)

CREATE TABLE Likes(
    student_id int,
    liked_id int,
    primary key (student_id, liked_id),
    foreign key (liked_id) references Highschooler(ID),
    foreign key (student_id) references Highschooler(ID)
)

-- Using valid SQLite, answer the following questions for the tables provided above.

-- What is Kyle’s id?
SELECT
```sql
CREATE TABLE Highschooler(
    ID int primary key,
    name text,
    grade int);

/*
3 example rows:
SELECT * FROM Highschooler LIMIT 3;
    ID  name  grade
  1510  Jordan  9
  1689  Gabriel  9
  1381   Tiffany  9
*/

CREATE TABLE Friend(
    student_id int,
    friend_id int,
    primary key (student_id, friend_id),
    foreign key (student_id) references Highschooler(ID),
    foreign key (friend_id) references Highschooler(ID));

/*
3 example rows:
SELECT * FROM Friend LIMIT 3;
    student_id  friend_id
     1510       1381
     1510       1689
     1689       1709
*/

CREATE TABLE Likes(
    student_id int,
    liked_id int,
    primary key (student_id, liked_id),
    foreign key (liked_id) references Highschooler(ID),
    foreign key (student_id) references Highschooler(ID));

/*
3 example rows:
SELECT * FROM Likes LIMIT 3;
    student_id  liked_id
     1689       1709
     1709       1689
     1782       1709
*/

-- Using valid SQLite, answer the following questions for the tables provided above.
-- What is Kyle’s id?
SELECT
Figure 10: Example prompt for Create Table + Select 3.
CREATE TABLE "border_info" ("state_name" text, "border" text)
/
state_name border
alabama tennessee
alabama georgia
alabama florida/

CREATE TABLE "city" ("city_name" text, "population" int DEFAULT NULL, "country_name" varchar(3) NOT NULL DEFAULT '', "state_name" text)
/
city_name population country_name state_name
birmingham 284413 usa alabama
mobile 200452 usa alabama
montgomery 177857 usa alabama/

CREATE TABLE "highlow" ("state_name" text, "highest_elevation" text, "lowest_point" text, "highest_point" text, "lowest_elevation" text)
/
state_name highest_elevation lowest_point highest_point lowest_elevation
alabama 734 gulf of mexico cheaha mountain 0
alaska 6194 pacific ocean mount mckinley 0
arizona 3851 colorado river humphreys peak 21/

CREATE TABLE "lake" ("lake_name" text, "area" double DEFAULT NULL, "country_name" varchar(3) NOT NULL DEFAULT '', "state_name" text)
/
lake_name area country_name state_name
iliamna 2675.0 usa alaska
becharof 1186.0 usa alaska
teshkepk 816.0 usa alaska/

CREATE TABLE "mountain" ("mountain_name" text, "mountain_altitude" int DEFAULT NULL, "country_name" varchar(3) NOT NULL DEFAULT '', "state_name" text)
/
mountain_name mountain_altitude country_name state_name
mckinley 6194 usa alaska
st. elias 5489 usa alaska
foraker 5304 usa alaska/

CREATE TABLE "river" ("river_name" text, "length" int DEFAULT NULL, "country_name" varchar(3) NOT NULL DEFAULT '', "traverse" text)
/
river_name length country_name traverse
mississippi 3778 usa minnesota
mississippi 3778 usa wisconsin
mississippi 3778 usa iowa/

CREATE TABLE "state" ("state_name" text, "population" int DEFAULT NULL, "area" double DEFAULT NULL, "country_name" varchar(3) NOT NULL DEFAULT '', "capital" text, "density" double DEFAULT NULL)
/
state_name population area country_name capital density
alabama 3894000 51700.0 usa montgomery 75.319149
alaska 401800 591000.0 usa juneau 0.679865
arizona 2718000 114000.0 usa phoenix 23.842105/

-- Using valid SQLite, answer the following questions for the tables provided above.
-- what is the population of austin
SELECT CITYalias0.POPULATION FROM CITY AS CITYalias0 WHERE CITYalias0.CITY_NAME = "austin";
-- which state is kalamazoo in
SELECT CITTAlias0.STATE_NAME FROM CITY AS CITTAlias0 WHERE CITTAlias0.CITY_NAME = "kalamazoo";
-- name all the rivers in colorado
SELECT RIVERalias0.RIVER_NAME FROM RIVER AS RIVERalias0 WHERE RIVERalias0.TRAVERSE = "colorado";
-- how many people live in new mexico
SELECT STATEalias0.POPULATION FROM STATE AS STATEalias0 WHERE STATEalias0.STATE_NAME = "new mexico";
-- what states border missouri
SELECT BORDER_INFOalias0.BORDER FROM BORDER_INFO AS BORDER_INFOalias0 WHERE BORDER_INFOalias0.STATE_NAME = "missouri";
-- what is the biggest city in arizona
SELECT

Figure 11: Example prompt for 5-shot. It starts with the schema and 3 rows per database (exactly as in Figure 10), followed by 5 few-shot examples, and finally the target question.