March 28, 2024 → Experiments with open source text-to-sql models

After deciding to use text-to-sql models for our task, I first experimented with different open source models to find the one that performs close to GPT and carried out 2 more in depth qualitative evaluation of CodeS-7B model, which was performing better than other open source models.

▼ Selecting open source model

Dataset

I prepared a small dataset of 60 question-answer pairs to test different models. The dataset includes only a question and the answer. Here is a sample of the dataset:

Question	Answer
What is the family name of "chiara" who owns "casa in 2 affittanze, porcion" in "corte nova"?	stella
What is the total number of distinct property types listed?	3231
What are the locations of the properties owned by "nicolo" "contarini"?	['fondamenta del bagatin', 'giù dal ponte bagatin', ...]

The notebook for preparing the dataset: `./text_to_sql/experiment_0/prepare_test_db.ipynb`

The dataset: `./text_to_sql/experiment_0/test_db.csv`

Models

By going through some paper, I decided to test 3 models that are fine-tuned for text-to-sql task:

- sqlcoder-7b-2
- codes-7b
- nsql-llama-2-7B

After running an inference using all these 3 models, I executed the query on the database, and computed the exact matching and went through the answers to see the pros and cons of each model. For each of the models, I computed the following:

- Execution Error → Wrong SQL syntax
- Wrong → SQL syntax is correct, but does not answer the question
- True → True SQL query and answer

Below, I will report the summary and some limitations of each model:

1. sqlcoder-7b-2

Execution Error	Wrong	True
5	18	**37**

Limitations:

- It always uses `ILIKE` SQL key, even if we want the exact match, resulting in a wrong answer to the question.
    - Example: *What is the total number of properties held by individuals sharing the last name "pasqualigo"?*

      ```
      SELECT COUNT(c.Property_Type) AS total_properties
      FROM catastici c
      WHERE c.Owner_Family_Name ILIKE '%pasqualigo%';
      ```
- Cannot handle `'` that comes in the text correctly
- Hard to understand the question when it is not clear
    - Example: *property in "casa" in "calle del zadio"*
        - may not understand *"calle del zadio"* is the location
- Uses `COUNT` with multitple arguments → Execution error

2. codes-7b

Execution Error	Wrong	True
9	11	**40**

Limitations:

- Fails to identify First and Family names correctly
    - Example: *"pier alvise" "barbaro"*
- Converted lower case to upper case
    - Example: *"carlo"* to *"Carlo"*
- Confuses the column names
    - Example: *Rent_Income* with *Property_Type*

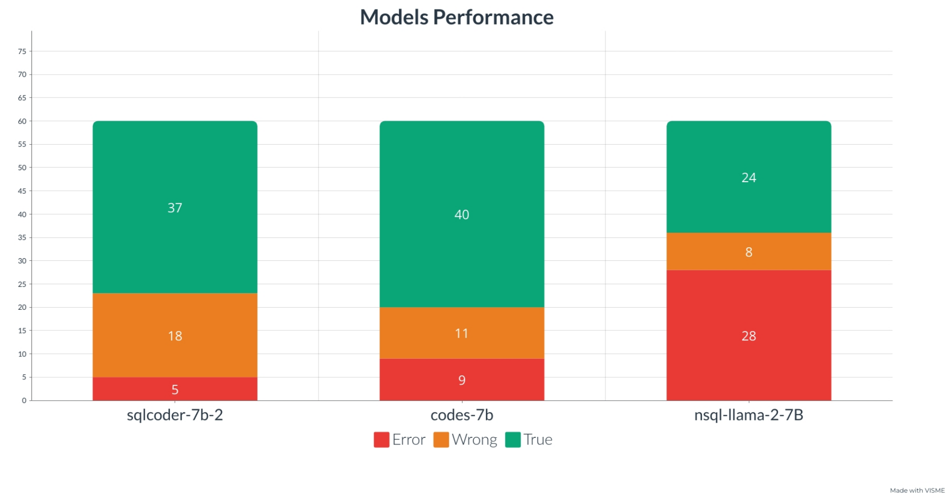
3. nsql-llama-2-7B

Execution Error	Wrong	True
28	8	**24**

Limitations:

- Generally the ones in the other 2 models 🙂

Summary



The notebook of these evaluations: `./text_to_sql/experiment_0/experiment.ipynb`

CodeS-7B model experiments

Based on the above experiment results and the limitations, we decided to use CodeS-7B model and did 2 experiment on it with more diverse dataset.

Experiment 1

In this experiment, I created 500 questions for evaluation and did inference of CodeS model in the **few-shot** setting, as explained in their paper and GitHub.

Code and dataset: `./text_to_sql/experiment_1/`

Dataset

Using GPT model, I created 100 diverse questions about the dataset, and created the SQL query of each question, and the answer to the questions by executing the SQL query on the database. So that we have a set of Question, True Answer, True SQL. Then, for each of the question, I created 4 more variations, resulting in 500 questions in total, every 5 question having the same SQL query and Answer, but 5 ways of rephrasing the questions. Here is a sample of the dataset:

Question	True SQL	True Answer

Can you tell me the number of properties located on "la calle vicina al campiel dal panizza in arzere"?	SELECT COUNT("Property_Type")
	FROM catastici
	WHERE "Property_Location" = 'la calle vicina al campiel dal panizza in arzere'
		14

What is the property count on "la calle vicina al campiel dal panizza in arzere"?	SELECT COUNT("Property_Type")
	FROM catastici
	WHERE "Property_Location" = 'la calle vicina al campiel dal panizza in arzere'
		14

Could you provide the total of properties found in "la calle vicina al campiel dal panizza in arzere"?	SELECT COUNT("Property_Type")
	FROM catastici
	WHERE "Property_Location" = 'la calle vicina al campiel dal panizza in arzere'
		14

Question	True SQL	True Answer

How many buildings can one find on "la calle vicina al campiel dal panizza in arzere"?	SELECT COUNT("Property_Type")
	FROM catastici
	WHERE "Property_Location" = 'la calle vicina al campiel dal panizza in arzere'
		14

How many properties are there in "la calle vicina al campiel dal panizza in arzer	SELECT COUNT("Property_Type")
	FROM catastici
	WHERE "Property_Location" = 'la calle vicina al campiel dal panizza in arzere'
		14

We have **500** questions in total.

Experiment results

Execution Error	Wrong	True
24	266	**230**

We have a mismatch in the answer in 266 cases, however, in most of the cases, even if the answers don’t match, the generated query and the answer are correct. It is usually because of

- additional soring in the answer
- returning additional column
- multiple ways of answering the question

In some cases, I even realized the ground truth SQL query generated by GPT was wrong.

So, the actual number of wrong answers were ~100, i.e. it had around **380 True answer**

Experiment 2

After the 1st experiment I realized that there were some issues with the dataset, such as

- Wrong Ground Truth SQL
- Ambiguous Questions

So, I spent some time to clean the dataset, to make sure the Ground Truth SQL queries are correct and the Questions are clear, and came up with more clean dataset for evaluation.

Further, depending on the SQL query length to answer the given question, I also selected the **100 Hard questions**, that usually requires nested loops of query to answer.

Example: *How many individuals are there without real estate in the area recognized as the most populated property location?*

Code and dataset: `./text_to_sql/experiment_2/`

Experiment results

Execution Error	Wrong	True
18	246	230

I also went through the Wrong answers counting how many of them are actually wrong, as a result I found out that 162 of were actually correct, but different than the ground truth for the above reasons, and in some cases both answers were acceptable.

Example: for the question *What is the range of rent incomes in "calle de franchi"?*

- True answer is MIN and MAX of Rent Income
- Generated answer is MAX-MIN of Rent Income

So, if we update the table we had:

Execution Error	Wrong	True
18	84	**392**

Observed Limitations of CodeS-7B model

- Cannot handle statistical questions, i.e. medina, variation, standard deviation, ... (of rent price for example)
- Confuses the feature names, e.g. property type with rent price
- Limits the number of outputs when not asked explicitly
- SQL syntax inconsistency when using `COUNT`
- Gets confused when answering H*ard questions* (like above) that require multiple nested queries and some statistics as well.

To Do

1. Clean the evaluation dataset further to make sure all the question variants are also non-ambiguous, and all the Ground Truth SQL are correct
2. Run another experiment using the Fine-tuned CodeS-7B checkpoint as explained in their paper
   - in the paper they have further fine-tuned this model on a famous text-to-SQL benchmark, and show that it generalizes to other dataset as well.

Further Ideas

- Use another LLM before the text-to-sql model
  - to paraphrase the question in a more clear way, so that that text-to-sql model can answer it more easily.
  - to breakdown the complex queries into simpler queries and run them separately (to be verified if it works)