CSE 5525: Assignment 3

Abhiram Rustagi

1 Data Statistics and Processing (8pt)

Instructions: Use Table 1 and **??** to describe the data statistics before and after any pre-processing respectively. Use the T5 tokenizer to report the statistics. For the statistics after pre-processing, if you did different pre-processing for different models, you need to indicate them separately. The gray text in each row is there to guide you and should be removed in your submitted report. Depending on your pre-processing, some numbers may be identical across tables.

Statistics Name	Train	Dev
Number of examples	4225	466
Mean sentence length	23.0973	23.0708
Mean SQL query length	217.3725	211.053
Vocabulary size (natural language)	796	470
Vocabulary size (SQL)	556	396

Table 1: Data statistics before any pre-processing. You need to at least provide the statistics listed above, and can add new entries.

Note:
For table 2, I had similar statistics as the data is not being modified.

2 T5 Fine-tuning and Training From Scratch (8pt)

Instructions: Use Table 2 and Table 3 to describe your data processing steps (if any) and the implementation details, respectively for the fine-tuned T5 model, and the T5 model trained from scratch. The gray text in each row is there to guide you and should be removed in the submitted report. Be clear enough that we can replicate your approach in PyTorch using only your descriptions.

Design choice	Description
Data processing	Tokenization using T5's built-in tokenizer with normalization (lowercasing, SQL keyword standardization).
Tokenization	Used T5 tokenizer without modifications.
Architecture	Fine-tuned the full T5-small model.
Hyperparameters	Learning rate: 5e-4, Batch size: 16, Optimizer: AdamW, Scheduler: Cosine, Max epochs: 10, Patience epochs: 4, Max new tokens: 512.

Table 2: Details of the best-performing T5 model configurations (fine-tuned)

Note:
Table 3 would have similar design choices, as nothing was done explicitly to differentiate from the fine tuned model.

Design choice	Description
Data processing	Tokenization using T5's built-in tokenizer with normalization (lowercasing, SQL keyword standardization).
Tokenization	Used T5 tokenizer without modifications.
Architecture	Fine-tuned the full T5-small model.
Hyperparameters	Learning rate: 5e-4, Batch size: 16, Optimizer: AdamW, Scheduler: Cosine, Max epochs: 10, Patience epochs: 4, Max new tokens: 512.

Table 3: Details of the best-performing T5 model configurations (from scratch)

3 Large Language Model (LLM) Prompting (14pt)

3.1 In-Context Learning (ICL)

Instructions: Provide in Table 4 the instruction prompt(s) that you used for ICL.

If the prompt you used for zero- and few-shot prompting is identical, except for the examples, there's no need to repeat it. If you made small modifications between zero- and few-shot, please provide them separately. For all entries, you need to specify the corresponding values of *k*.

Shot	Prompt
all	<instructions> You are an SQL expert that translates user requests into SQL queries for a flight database. Here is the schema: [shortened schema] Please generate ONLY the SQL query, and do not repeat the prompt. </instructions> <user_request> [Prompt example] </user_request> <response> [Response (SQL) example] </response> <user_request> [Prompt example] </user_request> <response> [Response (SQL) example] </response> <user_request> [Actual prompt] </user_request> <response>

Table 4: Instruction prompts used for zero- and/or few-shot prompting.

Example selections: Please provide a clear, detailed, and succinct description of how you selected the examples when $k > 0$.

There is random selection of examples when $k > 0$

3.2 Best Prompt and Ablation Study

Instructions: Report the best prompt you used in Table 5. If the best prompt you used is the same as the one specified in Table 4, you can just copy the best prompt and label it. If it is different (e.g. you designed another prompt that is better), you should clearly describe how you created it and what are the methods you used in the caption.

You will also need to clearly and succintly describe in Table 6 the ablation experiments that you performed by removing different parts of the prompt. For that, you need to first highlight the parts of the prompt that you ablated for each experiment in a distinct color[1], as shows the placeholder example in Table 4, and second, provide the description by referring to the highlighted part. When reporting your results in Table 7, you will need to refer to your ablations variants.

Prompt
The prompt is the same as that given in table 4.

Table 5: The best prompt. Similar prompt as in table 4. The best value is when $k = 3$

Color	**Description**
ForestGreen	I always encountered an out-of-memory error for the GPU when trying this. I then sought advice from my friend Steven on how to proceed. Initially, I tried using the prompt with and without the schema, but I kept getting the exact same prompt returned. It turned out that the issue was due to token size limitations and the model having trouble recognizing the end of the prompt. That's why I experimented with using an end marker. Switching to tags provided more specificity and clarity about what I wanted the model to do.
Blue	I tested two different versions of the schema: A shortened version where the "ents" field had only entity names and their field names, the "defaults" field included only entity names, and the "links" field listed entity names and their linked entities. This version worked the best. The full scheme, where I kept getting the out of memory error.

Table 6: Ablation variants.

[1] `https://www.overleaf.com/learn/latex/Using_colors_in_LaTeX`

4 Results and Analysis (20pt)

Quantitative Results: Use Table 7 to report your test and development set results. Your test results should match with the results on gradescope. For the development set, you should also report results from experiments you conducted to arrive at your final configuration. When reporting experiments, you should replace "variant" with brief and meaningful descriptions of whatever hyperparameter or setting that you varied. For ICL, you should specify what is the parameter k used, and what the full model corresponds to. For T5, the full model refers to the best model you described in Section 2. For T5, if you experimented with different design choices, you can add rows specifying the variants and what you tried. The text in gray is only for example purpose, and should be removed and replaced with your own choices. You may add more rows if needed.

System	Query EM	F1 score
Dev Results		
LLM Prompting		
Full model	16.47	18.86
Variant1 (ICL, $k = 1$)	14.45	18.75
Variant2 (ICL, $k = 0$)	11.58	11.58
Variant3 (e.g. ablating the explanation sentence in Table 5)	11.58	11.58
Variant4 (ICL, $k = 0$), full schema	ERR	ERR
T5 fine-tuned		
Full model	35.2	38.2
T5 from scratch		
Full model	32.6	35.51
Test Results		
How do you test this? I am not really sure? ICL	XX.XX	XX.XX
T5 fine-tuning	XX.XX	XX.XX
T5 from scratch	XX.XX	XX.XX

Table 7: Development and test results. Use this table to report quantitative results for both dev and test results, for all the three models.

ICL sensitivity to k: For ICL, please provide a plot of the Record F1 on the development set that the model achieved with different values of k. The x-axis should be k, and the y-axis the Record F1. The prompts and examples used for this plot should correspond to the ones you described in Subsection 3.1.

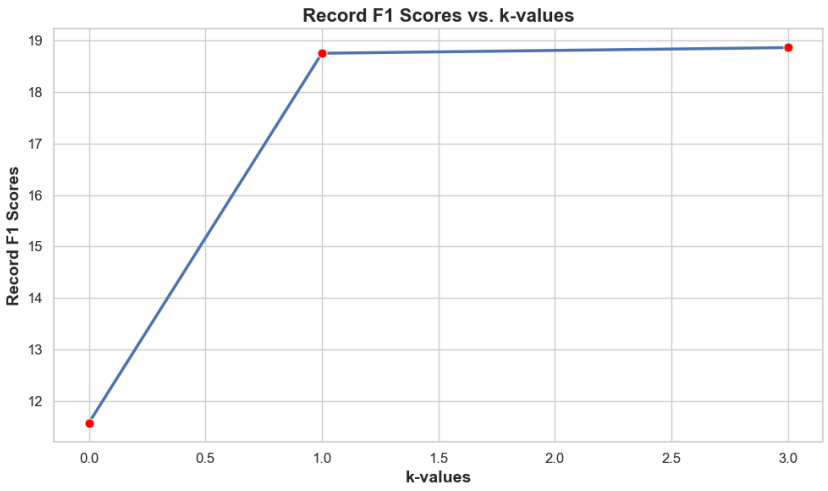

Figure 1: F1 Scores vs. k-values

Qualitative Error Analysis: Conduct a detailed error analysis for each of the three models. Identify common error types and discuss possible reasons for these errors in Table 8.

You must identify at least three classes of errors for the queries, and use examples to illustrate them. It must be clear what model makes the errors you are analyzing. If you identified the same type of error for different models, you don't need to duplicate the descriptions, but you need to clearly specify an example for each of the model, indicate the statistics for each model, and specify to which model each statistics correspond to. You may add more rows to the table.

Error Type	Relevant Models	Example of Error (All from T5 FT)	Error Description	Statistics
Unnecessary Conditions	T5 fine-tuned	INCORRECT: SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport airport_1, airport_service airport_service_1, city city_1 WHERE flight_1.to_airport = airport_1.airport_code AND airport_1.airport_code = 'MKE' AND flight_1.from_airport = airport_service_1.airport_code AND airport_service_1.city_code = city_1.city_code AND 1 = 1	Unnecessary tautologies such as "AND 1 = 1" that do not contribute to query logic	29/466
Syntax Issues (Parentheses, Commas, etc.)	T5 fine-tuned	CORRECT: SELECT DISTINCT aircraft_1.aircraft_code FROM aircraft aircraft_1 WHERE aircraft_1.aircraft_code = '734' INCORRECT: SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service airport_service_1, city city_1 WHERE flight_1.airline_code = '734' AND(flight_1.to_airport = airport_service_1.airport_code AND airport_service_1.city_code = city_1.city_code AND city_1.city_name = '734'	Unbalanced parentheses or missing commas and other syntax errors	131/466
Redundant Joins	T5 fine-tuned	INCORRECT: SELECT DISTINCT flight_1.flight_id FROM flight flight_1, airport_service airport_service_1, city city_1, airport_service airport_service_2, city city_2, days days_1, date_day date_day_1 WHERE flight_1.from_airport = airport_service_1.airport_code AND airport_service_1.city_code = city_1.city_code AND city_1.city_name = 'DENVER' AND(flight_1.to_airport = airport_service_2.airport_code AND airport_service_2.city_code = city_2.city_code AND city_2.city_name = 'BOSTON' AND flight_1.flight_days = days_1.days_code AND days_1.day_name = date_day_1.day_name AND date_day_1.year = 1991 AND date_day_1.month_number = 8 AND date_day_1.day_number = 9)	Includes unnecessary joins that do not change the result of the query	42/466
Incorrect or Mismatched Conditions	T5 fine-tuned	CORRECT: SELECT DISTINCT aircraft_1.aircraft_code FROM aircraft aircraft_1 WHERE aircraft_1.basic_type = 'F28' INCORRECT: SELECT DISTINCT capacity_1.flight_id FROM capacity_1, 'F28'	Differences in condition clauses, such as time ranges, city names, or flight_id references	460/466

Table 8: This table presents a qualitative analysis of errors found in queries generated by the T5 fine-tuned model on the development set, covering various error types.