CSE 5525: Assignment 3

Kurt Wanner

1 Data Statistics and Processing (8pt)

Instructions: Use Table 1 and Table 2 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	18.10	18.07
Mean SQL query length	218.37	212.05
Vocabulary size (natural language)	792	466
Vocabulary size (SQL)	557	397

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

Statistics Name	Train	Dev
Model: T5 Fine-tuning & T5 from Scratch		
Mean sentence length	8.43	8.35
Mean SQL query length	126.14	121.96
Vocab size (NL)	727	379
Vocab size (SQL)	491	299

Table 2: Data statistics after pre-processing. You need to at least provide the statistics listed in Table 1 (except for the number of lines), and can add new entries.

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

Instructions: Use Table 3 and Table 4 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	The first step was to remove stop words. This list can be found in `data/stop_words.txt` and `data/stop_phrases.txt`. Next, proper names (i.e. houston, california, atlanta airport, etc.) were converted to their SQL counterparts (i.e. 'HOUSTON', 'CALIFORNIA', and 'ATL').
Tokenization	The default T5 tokenizer was used with additional added tokens. These additional tokens can be found in `data/tokens_nl.txt` and `data/tokens_sql.txt`.
Architecture	Because of the additional tokens, I resized the token embeddings. For fine-tuning, the entire model was tuned. To increase training speed, the model's parameters were converted to type bfloat16.
Hyperparameters	**Learning Rate:** 3e-5 **Batch Size:** 16 **Stopping Criteria:** Training periods of 15 epochs each, repeated until improvements in Record F1 slowed. **LR Scheduler:** Cosine

Table 3: Details of the best-performing T5 model configurations (fine-tuned)

Design choice	Description
Data processing	The Data Processing was identical to the fine-tuned model.
Tokenization	The tokenization was identical to the fine-tuned model.
Architecture	Similar to the fine-tuned model, the model's parameters were converted to type bfloat16 to improve training speed.
Hyperparameters	**Learning Rate:** 1e-4 **Batch Size:** 8 **Stopping Criteria:** Training periods of 15 epochs each, repeated until improvements in Record F1 slowed. **LR Scheduler:** Cosine

Table 4: 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 5 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
0	You are a very powerful text-to-sql model. Given the following SQL tables, create an SQL query for a user request. [SCHEMA FOR FLIGHT DATABASE] Create an SQL Query for the following question. Only return the output and nothing else. [Question]: [NL INSTRUCTION] [Answer]:
$k > 0$	You are a very powerful text-to-sql model. Given the following SQL tables, create an SQL query for a user request. [SCHEMA FOR FLIGHT DATABASE] [Question]: [TRAINING NL INSTRUCTION] [Answer]: [CORRESPONDING SQL QUERY] $\vdots$ Base your answer on these examples. Create an SQL Query for the following question. Only return the output and nothing else. [Question]: [NL INSTRUCTION] [Answer]:

Table 5: 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 $k > 0$.

The examples were chosen dynamically based on a simple similarity score between the given NL instruction and all training examples. The similarity score is based on the set of key words in each NL instruction. This list of key words can be found in `data/prompt_key_words.txt`.

The calculation for the similarity score is $Score(key_1, key_2) = |key_1 \setminus key_2| + |key_2 \setminus key_1|$

3.2 Best Prompt and Ablation Study

Instructions: Report the best prompt you used in Table 6. If the best prompt you used is the same as the one specified in Table 5, 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 succinctly describe in Table 7 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 5, and second, provide the description by referring to the highlighted part. When reporting your results in Table 8, you will need to refer to your ablations variants.

Prompt
You are a very powerful text-to-sql model. Given the following SQL tables, create an SQL query for a user request.
[SCHEMA FOR FLIGHT DATABASE]
[Question]: [TRAINING NL INSTRUCTION] [Answer]: [CORRESPONDING SQL QUERY]
⋮
Base your answer on these examples. Create an SQL Query for the following question. Only return the output and nothing else. [Question]: [NL INSTRUCTION] [Answer]:

Table 6: **The best prompt.** My best prompt was the prompt from Table 5

Color	Description
Red	In this experiment, the Red text was removed.
Blue	In this experiment, the Blue text was removed.
Purple	In this experiment, both the Red and Blue text was removed.

Table 7: **Ablation variants.** Put a clear description of the ablation experiments you did. If it helps to describe more clearly, you can also provide an example of the relevant part of the prompt before and after the ablation.

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

4   Results and Analysis (20pt)

Quantitative Results: Use Table 8 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: Gemma-2B		
ICL, $k = 0$	0.0	11.80
ICL, $k = 1$	0.0	11.25
ICL, $k = 2$	0.0	11.76
ICL, $k = 3$	0.0	12.33
Red Ablation, $k = 1$	0.0	11.56
Blue Ablation, $k = 1$	0.0	11.77
Purple Ablation, $k = 0$	0.0	11.80
Purple Ablation, $k = 1$	0.0	12.42
Purple Ablation, $k = 3$	0.0	12.08
T5 fine-tuned		
Full model	4.94	87.90
T5 from scratch		
Full model	4.51	70.59
Test Results		
ICL	XX.XX	XX.XX
T5 fine-tuning	XX.XX	XX.XX
T5 from scratch	XX.XX	XX.XX

Table 8: **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.
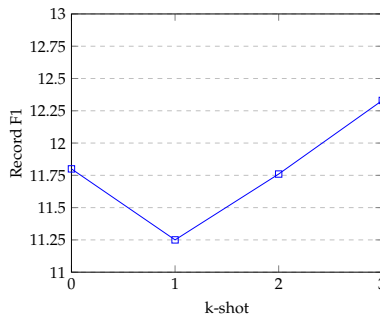
Qualitative Error Analysis: Conduct a detailed error analysis for each of the three models. Identify common error types and discuss possible reasons for these errors in Table 9.

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	Error Description	Statistics
Error name	ICL, T5 fine-tuned or T5 from scratch	Snippet from datapoint examplifying error	Describe the error in natural language	Provide statistics in the form "COUNT/TOTAL" on the prevalence of the error. TOTAL is the number of relevant examples (e.g. number of queries, for query-level error), and COUNT is the number of examples that showed this error. If an error type applies to several models, you should provide the statistics in the form "MODEL NAME: COUNT/TOTAL".

Table 9: Use this table for your qualitative analysis on the dev set, for all the three models.