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	10.96	10.91
Mean SQL query length	60.9	58.90
Vocabulary size (natural language)	1134	548
Vocabulary size (SQL)	759	438

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	18.10	18.07
Mean SQL query length	218	212
Vocab size (NL)	792	466
Vocab size (SQL)	557	397

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	No data processing occurred for either the natural language instructions or the SQL instructions.
Tokenization	For the natural language instructions, the default T5 tokenizer was used without change. For SQL instructions, a start-of-sequence token was inserted at the front of each instruction.
Architecture	I fine-tuned the entire model. To increase training speed, the model's parameters were converted to type bfloat16.
Hyperparameters	**Learning Rate:** 1e-5 **Batch Size:** 8 **Stopping Criteria:** 2 consecutive epochs with diminishing validation performance. **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:** 2 consecutive epochs with diminishing validation performance. **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	Convert the following natural language instruction into its equivalent SQL instruction. Ensure your response is a syntactically valid SQL instruction. Only return the output and nothing else. [Instruction]: [NL INSTRUCTION] [Answer]:
$k > 0$	[Instruction]: what evening flights do you have available from baltimore to philadelphia [Answer]: [CORRESPONDING SQL QUERY] [Instruction]: how much is coach flight from pittsburgh to atlanta [Answer]: [CORRESPONDING SQL QUERY] [Instruction]: get saturday fares from washington to montreal [Answer]: [CORRESPONDING SQL QUERY] [Instruction]: all flights and fares from atlanta to dallas round trip after 12pm less than 1100 dollars [Answer]: [CORRESPONDING SQL QUERY] Convert the following natural language instruction into its equivalent SQL instruction. Ensure your response is a syntactically valid SQL instruction. Only return the output and nothing else. [Instruction]: [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 selected based on how well representative their task was to the rest of the training data. The first example provided above I felt was most representative of the training dataset in terms of both the NL instruction and the corresponding SQL elements. The second example provided was second most representative and so on. For k-shot prompting, the first k examples were selected from the above list.

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 succintly 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
Everything in this table serves as a placeholder to show an example of prompt and formatting. It is just for reference, and you need to remove all the text here and put your own prompt. You can design and experiment with different prompts to find what works better. In the prompt, it could be helpful to provide clear instructions, provide the relevant context, and you can also provide examples. You need to experiment with different numbers of examples. In this prompt, two examples are provided below ($k = 2$, and this is 2-shot prompting): Instruction: [instruction-sentence] Response: [response] Instruction: [instruction-sentence] Response: [response] You can choose to provide additional instruction or not before providing the instruction to the model as well. Below is the task that you are asking the model to perform: Instruction: [instruction-sentence] Response:

Table 6: The best prompt. Remove this text and place a short description of your best prompt. If it is the same as the prompt in Table 5, clearly mention the value of k. If it is not the same, state explicitly how you built it.

Color	**Description**
ForestGreen	This is the description for the ablation experiment where the part in ForestGreen is ablated from the best prompt.
Blue	This is the description for the ablation experiment where the part in Blue is ablated from the best prompt. In the ablation experiment, we experimented with swapping the order of both sentences, because...

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		
Gemma-2B	12.02	12.42
ICL, $k = 0$	11.80	11.80
ICL, $k = 1$	12.02	12.42
ICL, $k = 3$	11.37	12.00
Variant3 (e.g. ablating the explanation sentence in Table 6)	XX.XX	XX.XX
Variant4 (e.g. ablating by doing ... in Table 6)	XX.XX	XX.XX
Variant	XX.XX	XX.XX
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.

Remove this text and place your plot here!

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.