A Prompt Templates

In the LlamaDuo pipeline, we employ two prompt templates that serve different purposes: one for the generation of synthetic datasets and another for the evaluation of the outputs produced by the fine-tuned LLMs.

Figure 5 illustrates the prompt template used to assess the precision and similarity of the response `$lm_response` generated by fine-tuned small-scale LLMs, based on the prompt `$instruction` and response `$human_response` from the test subset of the coverage dataset. It is important to note that the `$` symbol indicates a placeholder, designed to be substituted with actual data during the runtime.

```
Given an instruction and two responses—one generated by a human and the other by a language model—I'm seeking to evaluate how closely the language model's response mirrors the human-generated one. Additionally, I want to assess the accuracy and relevance of the language model's response to the original instruction.

Instruction:
```
$instruction
```

Human Response:
```
$human_response
```

Language Model Response:
```
$lm_response
```
You are quality assessor who analyzes the similarity between the Human Response and the Language Model Response on a scale of 1 to 100, where 1 indicates no similarity and 100 indicates identical responses.

Also you analyze the Language Model Response how it accurately answers the given Instruction on a scale of 1 to 100. Analysis MUST be rigorous and thorough. Provide the assessment in the following JSON format:

{
  "similarity_assessment": {
    "score": [Insert similarity score here]
  },
  "precision_assessment": {
    "score": [Insert precision score here]
  }
}
```

Figure 5: Prompt template to evaluate the fine-tuned model's response.

Figure 6 shows the prompt template designed for the generation of synthetic data tailored to the summarization task while Figure 7 shows the prompt template for other tasks. Specifically, we use a prompt `$instruction` alongside its corresponding response `$response`, both sourced from the train subset of the coverage dataset, serving as an example pair. This example pair is utilized to instruct service LLMs to generate analogous data samples. In addition, our template is designed to generate multiple synthetic data samples through a singular request, thereby enhancing the efficiency of API utilization. Due to the unique features of different downstream tasks, there is no optimal prompt template that universally applies. The actual content of the prompt template is adjusted to align with the specific requirements of the task for which the synthetic dataset is being generated.

```
Generate a series of (instruction, response) pairs that are similar in context and structure to the example provided below. Each pair should consist of a concise instruction followed by an appropriate, detailed response. The instruction should pose a clear task or question, while the response should provide a comprehensive answer or solution that could be understood by someone with a basic understanding of the subject.

Example pair:
Instruction: $instruction
Response: $response

Your task is to generate more pairs that maintain this level of clarity and detail. The topic is $topic. Write a long text of instruction by yourself, then summarize the given instruction in a response. Ensure that the responses are informative and accurate, suitable for an educational context.

Store the generated pairs in JSON format, with each pair as an object within an array. Each object should have two key-value pairs: "instruction" and "response". For instance:

{
  "contents":
    [
      {"instruction": "text", "response": "text"},
      {"instruction": "text", "response": "text"},
      ...
    ]
}

Remember to maintain consistency in the format and ensure the generated pairs are diverse and cover a broad range of subjects. You must return the response in the asked format and you must not add any additional text in your response.
```

Figure 6: Prompt template of data synthesis for summarization tasks.

```
Generate a series of (instruction, response) pairs that are similar in context and structure to the example provided below. Each pair should consist of a concise instruction followed by an appropriate, detailed response. The instruction should pose a clear task or question, while the response should provide a comprehensive answer or solution that could be understood by someone with a basic understanding of the subject.

Example pair:
Instruction: $instruction
Response: $response

Your task is to generate more pairs that maintain this level of clarity and detail. The topic is $topic. Ensure that the responses are informative and accurate, suitable for an educational context.

Store the generated pairs in JSON format, with each pair as an object within an array. Each object should have two key-value pairs: "instruction" and "response". For instance:

{
  "contents":
    [
      {"instruction": "text", "response": "text"},
      {"instruction": "text", "response": "text"},
      ...
    ]
}

Remember to maintain consistency in the format and ensure the generated pairs are diverse and cover a broad range of subjects. You must return the response in the asked format and you must not add any additional text in your response.
```

Figure 7: Prompt template of data synthesis for classification, coding, and closed QA tasks.

B Implementation Configuration

This section delineates the statistical information of the coverage dataset and synthetic dataset generated by service LLMs. In addition, we present the details of the training configurations of our experiments. The detailed pipeline implementation of LlamaDuo is available at https://github.com/deep-diver/llamaduo.

B.1 Coverage Datasets

The foundational coverage dataset employed in our study is the "No Robots" dataset (Rajani et al., 2023). We leverage four subsets of the coverage dataset, namely summarization, classification, coding, and closed QA, for synthetic data generation. Table 3 illustrates the initial composition of the