Please kindly review the answers for the AI task.

Firstly, I have to explain that OpenAI is blocked in the China mainland, so I choose the Alibaba models which are compatible with the OpenAI API for the task from the beginning.

**1. Run the script, fix it until it is fully-functional**

**[Elaine]:** Done, please refer to the [GitHub link](https://github.com/YanZ-Github/ai-assignment).

**2. Implement the TODOs**

**[Elaine]:** Done, please refer to the [GitHub link](https://github.com/YanZ-Github/ai-assignment).

**3. Experiment with different prompts, find the best one.**

**\* Feel free to change the code as you see fit and use an interactive notebook environment.**

**[Elaine]:** Done, please refer to the [GitHub link](https://github.com/YanZ-Github/ai-assignment).

Model: ' qwen2-7b-instruct '

For the best prompt: please refer to the file ‘bestprompt.txt’ via [GitHub link](https://github.com/YanZ-Github/ai-assignment/blob/main/bestprompt.txt).

**\* What happens if you change the model to gpt-4o-mini?**

**[Elaine]:** As ‘gpt-4o-mini’ is inaccessible and I am using Alibaba model, I change the Alibaba model from 'qwen2-7b-instruct' to 'qwen-turbo-0624'.

* The accuracy is not significantly improved.
* Prompt (including assistant) is still necessary to help model provide a correct answer.

**4. What happens when you run the same experiment several times? How the results can be reported and accounted for?**

**Bonus: (optional) change the script to do that**

**[Elaine]:** Done, please refer to the [GitHub link](https://github.com/YanZ-Github/ai-assignment).

1) To report and account for the result:

Generate JSON files as result reports with the accuracy and response after each turn of running.

2) What happens after running the same experiment several times:

* The result is random. The answer can be correct or incorrect for the same question in different turns of execution.
* The accuracy is unstable.
* Other comments:

- The larger the dataset(limit) is the lower the average accuracy will be.

- Prompt have affection on the accuracy of LLM.

- The higher the model's ability is, the less the prompt's affection will be.

The followings are results summary for two models,

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **qwen2-7b-instruct:** | **Limit** | **Accuracy** | | | | | | | | | | | | **AVG** |
| aprompt | 5 | 0.8 | 0.8 | 0.4 | 1 | 0.6 | 0.8 | 0.6 | 0.6 | 0.4 | 0.8 | 0.6 | 0.6 | **0.667** |
| 50 | 0.46 | 0.4 | 0.42 | 0.42 | 0.44 | 0.4 | 0.4 | 0.42 | 0.46 | 0.44 |  |  | **0.426** |
| bestprompt | 5 | 0.8 | 0.6 | 0.6 | 0.6 | 0.8 | 0.8 | 0.4 | 0.8 | 0.6 | 0.6 | 0.8 | 0.8 | **0.683** |
| 50 | 0.62 | 0.52 | 0.6 | 0.6 | 0.54 | 0.6 | 0.52 | 0.62 | 0.58 | 0.52 |  |  | **0.572** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **qwen-turbo-0624:** | **Limit** | **Accuracy** | | | | | | **AVG** |
| aprompt | 5 | 0.6 | 0.6 | 0.4 | 0.8 | 0.8 | 0.6 | **0.63** |
| 50 | 0.44 | 0.4 | 0.4 |  |  |  | **0.41** |
| bestprompt | 5 | 0.6 | 0.4 | 0.6 | 0.8 | 0.6 | 0.4 | **0.57** |
| 50 | 0.52 | 0.64 | 0.56 |  |  |  | **0.57** |

*Note:*

*If no prompt, accuracy for the two models is both 0.2~0.4; if no prompt and no assistant, accuracy for the two models is 0.0.(limit= 5)*

*Please refer to the JSON files for all the* [*result reports*](https://github.com/YanZ-Github/ai-assignment/tree/main/reports)*.*

**5. Bonus: (optional) change the script to support Alibaba models and the report their results, if time permits**

**[Elaine]:** Done, please refer to the [GitHub link](https://github.com/YanZ-Github/ai-assignment).

**6. Commit a self-contained and reproducible notebook to GitHub or any other code hosting service.**

**It should show several best prompts you found for each model and be runnable in a new Python environment.**

**Send us a link to the notebook as a solution.**

**[Elaine]:** Done, please refer to the [GitHub link](https://github.com/YanZ-Github/ai-assignment).

Mainly use model 'qwen2-7b-instruct' and ‘qwen-turbo-0624’.

For result comparation, please refer to the two tables above.

For prompts, please refer to the file ‘bestprompt.txt’ via GitHub link. (use the same best prompt, and compared the results)