Points to Note:

1. Most of the files in the main folder use default temperature settings, which may lead to deviations from the results specified in the “Kaggle Project Progress” PDF. Additionally, some files have been overwritten due to repeated experiments with different temperature settings and GPT models. For example, the file “2shotsphysics1each” contains results from various experiments with the GPT-3.5-turbo model, but the Python code might reference the GPT-4o-mini model.
2. I observed that few-shot learning improved performance for GPT-3.5-turbo but not for GPT-4 models. Specifically, for GPT-4 models, a well-optimized zero-shot prompt produced results comparable to or better than those from few-shot prompts with weaker system messages. In some cases, few-shot learning actually reduced performance for these models.
3. The domains present in the folders were generated using GPT-4o-mini after conducting few-shot experiments to assess whether they differed from those generated by GPT-3.5-turbo. The few-shot inference results mentioned in the “Kaggle Project Progress” PDF are based on domains generated by GPT-3.5-turbo.
4. In the Updated folder, results were obtained using zero-shot learning with a temperature setting of 0 and Prompt 5 for all three models. With a temperature setting of 0.8743, GPT-4-turbo achieved an accuracy of 86%, which is currently the best result. I also attempted an ensemble method with majority voting across the three models (GPT-3.5-turbo, GPT-4o-mini, GPT-4-turbo). However, due to significant performance variance (70.5%, 80.5%, and 85.5%, respectively), the ensemble approach did not exceed 84% accuracy and proved to be less effective.