HumanEval Dataset HP-Score:

Refer to dataset at: https://huggingface.co/datasets/openai/openai_humaneval

S.No	Sample Id	Basic Recall and Reproduction	Understanding and Interpretation	Analysis and Reasoning	Application of knowledge and execution	HP-score
1	HumanEval/0	4	4	4	4	4
2	HumanEval/1	5	5	5	5	5
3	HumanEval/2	5	5	5	5	5
4	HumanEval/3	5	5	5	5	5
5	HumanEval/4	5	5	5	5	5
6	HumanEval/5	4	4	4	4	4
7	HumanEval/6	4	5	4	5	4.5
8	HumanEval/7	5	5	5	5	5
	Fianl Scores	4.625	4.75	4.625	4.75	4.68

Policy:

- Each rule of the taxonomy is scored by a human on a scale from 1 to 5, as the hierarchical prompt framework has 5 levels.
- Each sample is scored based on its own complexity.
- A representative set of dataset is randomly sampled for scoring the dataset. The size of the set is nearly 5% of the original dataset.