

Induct-Learn

Induct-Learn introduces a method to create Induct-Learn Prompts from LLMs, significantly enhancing smaller models' efficiency and performance with fewer resources.

Dataset

BBHI Dataset

The BBHI dataset utilizes all sub-tasks from the BBH dataset, with the only difference being the removal of common prefixes and suffixes within each sub-task. For detailed descriptions of the sub-tasks, please refer to the task descriptions in BBH(<https://github.com/suzgunmirac/BIG-Bench-Hard/tree/main>).

We provide a processed dataset, the directory location is as follows. /BBH_preprocess/BBHI

BBHI Directory Structure Explanation

The BBHI directory is organized with each sub-task contained within its own folder. Below is an outline of the directory structure:

BBHI

-boolean_expressions

- boolean_expressions.3-example.txt : Original examples from the BBH paper.

- boolean_expressions.3-human-cot-example.txt : Examples of Chain-of-Thought (COT) written by humans.

- boolean_expressions.ori_instruction.txt : Original Task instructions(TI) written by humans.

- boolean_expressions.json : Full dataset file.

- boolean_expressions.test.jsonl : Split test set of the dataset.

- boolean_expressions.train.jsonl : Split training set of the dataset.

- boolean_expressions.3-shot_example.txt : first 3 data in training set

- boolean_expressions.9-shot_example.txt : first 9 data in training set

-causal_judgement ...

If you want. You can download and process on your own.

If you wish to download and process the BIG-Bench-Hard dataset on your own (<https://github.com/suzgunmirac/BIG-Bench-Hard/tree/main>): please place the bbh folder under source_dir="BIG-Bench-Hard-main/bbh" and execute the following ipynb file.

BBH_preprocess_BBHI.ipynb

Read each subtask file from the source_dir location, and read the rule file to clean the common prefixes and suffixes in each subtask, finally writing to the target_dir.

source_dir = "BIG-Bench-Hard-main/bbh" target_dir = "BBH_preprocess/BBHI" rules_file = "BBH_preprocess/remove_prefix_suffix.txt"

BBH_extract_COT_example.ipynb

Extract human-written COT examples from BBH, and convert the Q: A: format into [Question]... [Answer]... format.

BBHI_split_preprocess_dataset.ipynb

Split the training set and test set. And generate xxx.3-shot_example.txt and xxx.9-shot_example.txt files.

Evals-Induct Dataset

We present a compilation of 25 sub-task datasets carefully selected from the OpenAI Evals(<https://github.com/openai/evals>) platform. Each dataset is accompanied by a task description, as provided by its respective author. Each Evals-Induct tasks has around 91–200 examples, and the total number of instances is 3,883.

We provide a processed dataset, the directory location is as follows. /Evals-Induct_preprocess/Evals-Induct

Evals-Induct

-boolean_expressions

- samples.ori_instruction.txt : Original Task instructions(TI) written by humans.

- samples.json : Full dataset file.

- samples.test.jsonl : Split test set of the dataset.

- samples.test.200.jsonl : 200 sample test data

- `samples.train.jsonl` : Split training set of the dataset.
 - `samples.3-shot_example.txt` : first 3 data in training set
 - `samples.9-shot_example.txt` : first 9 data in training set
- causal_judgement ...

Please note that each subtask folder can also contain its own subtasks. They are distinguished by different prefixes. For example, the `word_association` subtask has 4 subtasks.

Evals-Induct

- `word_association`
- `related_words_2`
- `related_words_3`
- `related_words_4` -`related_words_5`

License

Our dataset and code are licensed under the MIT License.