Here is an example of running the inference code to interact with DeepSeek-R1:

```
# Download the model weights from Hugging Face
huggingface-cli download deepseek-ai/DeepSeek-R1 --local-dir
/path/to/DeepSeek-R1

# Clone DeepSeek-V3 GitHub repository
git clone https://github.com/deepseek-ai/DeepSeek-V3.git

# Install necessary dependencies
cd DeepSeek-R1/inference
pip install -r requirements.txt

# Convert Hugging Face model weights to a specific format (for running
the model on 16 H800 GPUs)
python convert.py --hf-ckpt-path /path/to/DeepSeek-R1 --save-path
/path/to/DeepSeek-R1-Demo --n-experts 256 --model-parallel 16

# Run the model and interact with it
torchrun --nnodes 2 --nproc-per-node 8 --node-rank $RANK --master-addr
$MASTER_ADDR generate.py --ckpt-path /path/to/DeepSeek-R1-Demo --config
configs/config_671B.json --interactive --temperature 0.7
--max-new-tokens 8192
```

We also release SFT and RL data to the public at xxx. In the review process, we upload the data as an attachment.

J. Evaluation Prompts and Settings

Table 18 | MMLU assesses a model's factual and conceptual understanding across 57 tasks spanning STEM (science, technology, engineering, mathematics), humanities, social sciences, and professional fields (e.g., law, medicine). The benchmark is commonly used to evaluate a model's ability to perform general knowledge reasoning and multitask proficiency across a diverse range of subjects and tasks. Here is an example of MMLU.

PROMPT
Answer the following multiple choice question. The last line of your response should be of the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of ABCD. Think step by step before answering. Which tool technology is associated with Neandertals? A. Aurignacian B. Acheulean C. Mousterian D. both b and c
Evaluation
Parse the last line in response to judge if the choice equals to ground truth.

Table 19 | MMLU-Redux is a subset of 5,700 manually re-annotated questions across all 57 MMLU subjects. MMLU-Redux focuses on improving the quality, clarity, and robustness of the benchmark by reducing noise, ambiguities, and potential biases in the MMLU, while potentially adjusting the scope or difficulty of tasks to better align with modern evaluation needs. Here is an example of MMLU-Redux.

PROMPT

Question:
Sauna use, sometimes referred to as "sauna bathing," is characterized by short-term passive exposure to extreme heat . . . In fact, sauna use has been proposed as an alternative to exercise for people who are unable to engage in physical activity due to chronic disease or physical limitations.[13]
According to the article, which of the following is NOT a benefit of sauna use?
Choices:
- (A) Decreased risk of heart attacks.
- (B) Increase in stroke volume.
- (C) Improved mental health.
- (D) Decreased rate of erectile dysfunction.
Instruction
Please answer this question by first reasoning and then selecting the correct choice.
Present your reasoning and solution in the following json format.
Please show your choice in the ‘answer‘ field with only the choice letter, e.g.,‘"answer": "C"‘.
{
"reasoning": "___",
"answer": "___"
}

Evaluation

Parse the json output in response to judge if the answer equals to ground truth.

Table 20 | LiveCodeBench aims to evaluate model performance on the algorithm competition task, which collects new problems over time from contests across three competition platforms, namely LeetCode, AtCoder, and CodeForces.

PROMPT

Question: There is a stack of N cards, and the ith card from the top has an integer A_i written on it. You take K cards from the bottom of the stack and place them on top of the stack, maintaining their order.
Print the integers written on the cards from top to bottom after the operation.
Input
The input is given from Standard Input in the following format:
N K
$A_1 A_2 \ldots A_N$
Output
Let B_i be the integer written on the ith card from the top of the stack after the operation. Print $B_1, B_2, \ldots, B_N$ in this order, separated by spaces.
Constraints
$-1 \leqslant K < N \leqslant 100$
$-1 \leqslant A_i \leqslant 100$
All input values are integers.
Sample Input 1
5 3
1 2 3 4 5
Sample Output 1
3 4 5 1 2
Initially, the integers written on the cards are 1,2,3,4,5 from top to bottom. After taking three cards from the bottom of the stack and placing them on top, the integers written on the cards become 3,4,5,1,2 from top to bottom.

Sample Input 2

6 2
1 2 1 2 1 2

Sample Output 2

1 2 1 2 1 2

The integers written on the cards are not necessarily distinct.

Please write a python code to solve the above problem. Your code must read the inputs from stdin and output the results to stdout.

Evaluation

Extract the code wrapped by "‘python"‘ in response to judge if the answer passes the test cases.

Table 21 | Compared to MMLU, MMLU-Pro features a curated subset of tasks, but with significantly increased difficulty. Questions in MMLU-Pro are designed to require deeper reasoning, multi-step problem-solving, and advanced domain-specific knowledge. For example, STEM tasks may involve complex mathematical derivations or nuanced scientific concepts, while humanities tasks may demand intricate contextual analysis.

PROMPT

The following are multiple choice questions (with answers) about business. Think step by step and then output the answer in the format of "The answer is (X)" at the end.

...

Question: Typical advertising regulatory bodies suggest, for example that adverts must not: encourage ___, cause unnecessary ___ or ___, and must not cause ___ offence.

Options: A. Safe practices, Fear, Jealousy, Trivial

B. Unsafe practices, Distress, Joy, Trivial

C. Safe practices, Wants, Jealousy, Trivial

D. Safe practices, Distress, Fear, Trivial

E. Unsafe practices, Wants, Jealousy, Serious

F. Safe practices, Distress, Jealousy, Serious

G. Safe practices, Wants, Fear, Serious

H. Unsafe practices, Wants, Fear, Trivial

I. Unsafe practices, Distress, Fear, Serious

Answer: Let's think step by step.

Evaluation

Parse the capital letter following "Answer: " in response to judge if the answer equals to ground truth.

Table 22 | DROP assesses a model's ability to understand and extract relevant information from extended textual passages. Unlike simpler question-answering benchmarks that focus on factual recall, DROP requires models to process and interpret context-rich paragraphs.

PROMPT

You will be asked to read a passage and answer a question. Some examples of passages and Q&A are provided below.

Examples — Passage: Looking to avoid back-to-back divisional losses, the Patriots traveled to Miami to face the 6-4 Dolphins at Dolphin Stadium . . . Cassel's 415 passing yards made him the second quarterback in Patriots history to throw for at least 400 yards in two or more games; Drew Bledsoe had four 400+ yard passing games in his Patriots career.

Question: How many points did the Dolphins lose by? Answer: 20.

— Passage: In week 2, the Seahawks took on their division rivals, the San Francisco 49ers. Prior to the season, NFL analysts rated this rivalry as the top upcoming rivalry, as well as the top rivalry of the decade . . . Seattle was now 2-0, and still unbeaten at home.

Question: How many field goals of at least 30 yards did Hauschka make? Answer: 2.

— Passage: at Raymond James Stadium, Tampa, Florida TV Time: CBS 1:00pm eastern The Ravens opened the regular season on the road against the Tampa Bay Buccaneers on September 10. . . . With the win, the Ravens were 1-0 and 1-0 against NFC Opponents.

Question: how many yards did lewis get Answer: 4. # Your Task

— Passage: The Chargers (1-0) won their season opener 22-14 against the Oakland Raiders after five field goals by Nate Kaeding and three botched punts by the Raiders. The Raiders Pro Bowl long snapper Jon Condo suffered a head injury in the second quarter. He was replaced by linebacker Travis Goethel, who had not snapped since high school. Goethel rolled two snaps to punter Shane Lechler, each giving the Chargers the ball in Raiders territory, and Lechler had another punt blocked by Dante Rosario. The Chargers scored their only touchdown in the second quarter after a 13-play, 90-yard drive resulted in a 6-yard touchdown pass from Philip Rivers to wide receiver Malcom Floyd. The Chargers failed to score four out of five times in the red zone. San Diego led at halftime 10-6, and the Raiders did not scored a touchdown until 54 seconds remained in the game. Undrafted rookie Mike Harris made his first NFL start, filing in for left tackle for an injured Jared Gaither. San Diego protected Harris by having Rivers throw short passes; sixteen of Rivers' 24 completions were to running backs and tight ends, and he threw for 231 yards while only being sacked once. He did not have an interception after throwing 20 in 2011. The win was the Chargers' eighth in their previous nine games at Oakland. It improved Norv Turner's record to 4-2 in Chargers' season openers. Running back Ryan Mathews and receiver Vincent Brown missed the game with injuries.

Question: How many yards did Rivers pass? Answer:

Think step by step, then write a line of the form "Answer: $ANSWER" at the end of your response.

Evaluation

Parse the capital letter following "Answer: " in response to judge if the answer equals to ground truth.