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.