

Table 28 | C-EVAL evaluates a model's breadth and depth of knowledge across 52 diverse academic disciplines, spanning humanities, social sciences, STEM (Science, Technology, Engineering, and Mathematics), and other professional fields (e.g., medicine, law). All question in C-Eval are Chinese.

---

**PROMPT**

以下是中国关于逻辑学考试的单项选择题，请选出其中的正确答案。

1991年6月15日，菲律宾吕宋岛上的皮纳图博火山突然大喷发，2000万吨二氧化硫气体冲入平流层，形成的霾像毯子一样盖在地球上空，把部分要照射到地球的阳光反射回太空几年之后，气象学家发现这层霾使得当时地球表面的温度累计下降了 $0.5^{\circ}\text{C}$ ，而皮纳图博火山喷发前的一个世纪，因人类活动而造成的温室效应已经使地球表面温度升高 $1^{\circ}\text{C}$ 。某位持“人工气候改造论”的科学家据此认为，可以用火箭弹等方式将二氧化硫充入大气层，阻挡部分阳光，达到地球表面降温的目的。以下哪项如果为真，最能对该科学家的提议构成质疑？

- A. 如果利用火箭弹将二氧化硫充入大气层，会导致航空乘客呼吸不适。
- B. 火山喷发形成的降温效应只是暂时的，经过一段时间温度将再次回升。
- C. 可以把大气层中的碳取出来存储在地下，减少大气层的碳含量。
- D. 不论何种方式，“人工气候改造”都将破坏地区的大气层结构。

答案：B

...

新疆的哈萨克人用经过训练的金雕在草原上长途追击野狼。某研究小组为研究金雕的飞行方向和判断野狼群的活动范围，将无线电传导器放置在一只金雕身上进行追踪。野狼为了觅食，其活动范围通常很广。因此，金雕追击野狼的飞行范围通常也很大。然而两周以来，无线电传导器不断传回的信号显示，金雕仅在放飞地3公里的范围内飞行。以下哪项如果为真，最有助于解释上述金雕的行为？

- A. 金雕放飞地周边重峦叠嶂，险峻异常。
- B. 金雕的放飞地2公里范围内有一牧羊草场，成为狼群袭击的目标。
- C. 由于受训金雕的捕杀，放飞地广阔草原的野狼几乎灭绝了。
- D. 无线电传导信号仅能在有限的范围内传导。

---

**Evaluation**

Parse the last line in response to judge if the choice equals to ground truth.

---

Table 29 | GPQA (Graduate-Level Google-Proof QA Benchmark) is a rigorous evaluation framework designed to measure an LLM's ability to tackle complex, graduate-level multiple-choice problems in STEM domains—specifically biology, physics, and chemistry.

---

**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.

Two quantum states with energies E1 and E2 have a lifetime of  $10^{-9}$  sec and  $10^{-8}$  sec, respectively. We want to clearly distinguish these two energy levels. Which one of the following options could be their energy difference so that they can be clearly resolved?

- A)  $10^{-9}$  eV
- B)  $10^{-8}$  eV
- C)  $10^{-4}$  eV
- D)  $10^{-11}$  eV

---

**Evaluation**

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

---

Table 30 | SimpleQA is a factuality evaluation benchmark that measures a model's ability to answer short, fact-seeking questions with precise, verifiable correctness.

---

<b>PROMPT</b>
Who received the IEEE Frank Rosenblatt Award in 2010?
<b>Evaluation</b>
Your job is to look at a question, a gold target, and a predicted answer, and then assign a grade of either ["CORRECT", "INCORRECT", "NOT_ATTEMPTED"]. First, I will give examples of each grade, and then you will grade a new example.
The following are examples of CORRECT predicted answers.
Question: What are the names of Barack Obama's children?
Gold target: Malia Obama and Sasha Obama
Predicted answer 1: sasha and malia obama
Predicted answer 2: most people would say Malia and Sasha, but I'm not sure and would have to double check
...
The following are examples of INCORRECT predicted answers.
Question: What are the names of Barack Obama's children?
Gold target: Malia and Sasha
Predicted answer 1: Malia.
Predicted answer 2: Malia, Sasha, and Susan.
...
The following are examples of NOT_ATTEMPTED predicted answers.
Question: What are the names of Barack Obama's children?
Gold target: Malia and Sasha
Predicted answer 1: I don't know.
Predicted answer 2: I need more context about which Obama you are ...
Also note the following things:
...
Here is a new example. Simply reply with either CORRECT, INCORRECT, NOT_ATTEMPTED. Don't apologize or correct yourself if there was a mistake; we are just trying to grade the answer.
Question: Who received the IEEE Frank Rosenblatt Award in 2010?
Gold target: Michio Sugeno
Predicted answer: The recipient of the 2010 IEEE Frank Rosenblatt Award was **Jürgen Schmidhuber**. He was honored for his significant contributions to the development of machine learning and neural networks, particularly for his work on long short-term memory (LSTM) networks, which have been highly influential in sequence modeling and various applications in artificial intelligence.
Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT
C: NOT_ATTEMPTED
Just return the letters "A", "B", or "C", with no text around it.

---

Table 31 | An example of C-SimpleQA. It measures a model's ability to answer short, fact-seeking questions in Chinese with precise, verifiable correctness.

---

**PROMPT**

显脉香茶菜可以用来治疗急性的什么类型的黄疸型肝炎?

---

**Evaluation**

请根据给定问题、标准答案和模型预测的答案来评估模型的回答是否正确。您的任务是将结果评定为：【正确】、【错误】或【未尝试】。

首先，我们将列出每个评定类别的示例，然后请您对新问题的预测答案进行评定。

以下是【正确】的答复示例：

问题：贝拉克·奥巴马的孩子叫什么名字？

标准答案：玛丽亚·奥巴马和萨莎·奥巴马

模型预测1：Malia Obama and Sasha Obama

模型预测2：玛丽亚和萨沙

...

以下是【错误】的答复示例：

问题：贝拉克·奥巴马的孩子叫什么名字？

标准答案：玛丽亚·奥巴马和萨莎·奥巴马

模型预测1：玛丽亚

模型预测2：玛丽亚、萨莎和苏珊

...

以下是【未尝试】的答复示例：

问题：贝拉克·奥巴马的孩子叫什么名字？

标准答案：玛丽亚·奥巴马和萨莎·奥巴马

模型预测1：我不知道。

模型预测2：我需要更多关于您所指奥巴马的上下文。

...

下面是一个新的问题示例。请只回复A、B、C之一，不要道歉或纠正自己的错误，只需要评估该回答。

问题：显脉香茶菜可以用来治疗急性的什么类型的黄疸型肝炎？

正确答案：黄疸型肝炎

预测答案：...

将此新问题的预测答案评定为以下之一：

A: 【正确】

B: 【错误】

C: 【未尝试】

只返回字母"A"、"B"或"C"，无须添加其他文本。

---

Table 32 | An example of math evaluation, which applies to AIME, MATH, and CNMO. These benchmarks evaluate model performance on mathematical tasks.

---

**PROMPT**

Let  $b \geq 2$  be an integer. Call a positive integer  $n$  *b-eautiful* if it has exactly two digits when expressed in base  $b$ , and these two digits sum to  $\sqrt{n}$ . For example, 81 is 13-beautiful because  $81 = \underline{6} \underline{3}_{13}$  and  $6 + 3 = \sqrt{81}$ . Find the least integer  $b \geq 2$  for which there are more than ten *b-eautiful* integers. Please reason step by step, and put your final answer within \boxed{}.

**Evaluation**

Parse the final answer within \boxed{} and use a rule-based grader to determine if it equals the ground truth. Round numerical values as needed, and use 'SymPy'<sup>1</sup> to parse expressions.

---