Table 25 | Arena-Hard is an open-ended evaluation benchmark specifically designed to assess the capabilities of LLMs. It presents models with challenging, novel, and diverse prompts curated from Chatbot Arena, a continuously evolving, crowd-sourced platform. It focuses on measuring model performance in open-ended tasks, with particular emphasis on coding and mathematics-related prompts. Given the inherently subjective nature of open-ended tasks, where multiple valid responses may exist, the benchmark necessitates the use of an evaluation model to approximate human judgment effectively. Higher evaluation scores suggest that the model is more likely to be favored by human users in real-world scenarios.

PROMPT

To write an SQL query that selects the top 10 rows in a database and joins to 3 different tables based on a field called "code", you would need to know the names of the tables and the specific fields you want to select. Assuming the tables are named "table1", "table2", and "table3", and you want to select all fields from each table, the query would look like this:

“‘sql SELECT * FROM (SELECT * FROM table1 LIMIT 10) AS t1 JOIN table2 ON t1.code = table2.code JOIN table3 ON t1.code = table3.code; “‘

This query first selects the top 10 rows from "table1" and then joins "table2" and "table3" based on the "code" field. If you want to select specific fields from each table, you can replace the asterisks (*) with the field names separated by commas.

Evaluation

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user prompt displayed below. You will be given assistant A’s answer and assistant B’s answer. Your job is to evaluate which assistant’s answer is better.

Begin your evaluation by generating your own answer to the prompt. You must provide your answers before judging any answers.

When evaluating the assistants’ answers, compare both assistants’ answers with your answer. You must identify and correct any mistakes or inaccurate information.

Then consider if the assistant’s answers are helpful, relevant, and concise. Helpful means the answer correctly responds to the prompt or follows the instructions. Note when user prompt has any ambiguity or more than one interpretation, it is more helpful and appropriate to ask for clarifications or more information from the user than providing an answer based on assumptions. Relevant means all parts of the response closely connect or are appropriate to what is being asked. Concise means the response is clear and not verbose or excessive.

Then consider the creativity and novelty of the assistant’s answers when needed. Finally, identify any missing important information in the assistants’ answers that would be beneficial to include when responding to the user prompt.

After providing your explanation, you must output only one of the following choices as your final verdict with a label:

1. Assistant A is significantly better: [[A>>B]]
2. Assistant A is slightly better: [[A >B]]
3. Tie, relatively the same: [[A=B]]
4. Assistant B is slightly better: [[B>A]]
5. Assistant B is significantly better: [[B>>A]]

Example output: "My final verdict is tie: [[A=B]]".

Table 26 | AlpacaEval 2.0 is an open-ended evaluation dataset, similar in nature to ArenaHard, and leverages an LLM to assess model performance on subjective tasks. However, in contrast to ArenaHard, the prompts in AlpacaEval 2.0 are generally less challenging and only a small subset necessitates the deployment of reasoning capabilities by the evaluated models.

PROMPT

What are the names of some famous actors that started their careers on Broadway?

Evaluation

< $|im_start|$ >system
You are a highly efficient assistant, who evaluates and selects the best large language model (LLMs) based on the quality of their responses to a given instruction. This process will be used to create a leaderboard reflecting the most accurate and human-preferred answers.
< $|im_end|$ >
< $|im_start|$ >user
I require a leaderboard for various large language models. I'll provide you with prompts given to these models and their corresponding outputs. Your task is to assess these responses, and select the model that produces the best output from a human perspective.
Instruction
{
"instruction": """{instruction}""",
}
Model Outputs
Here are the unordered outputs from the models. Each output is associated with a specific model, identified by a unique model identifier.
{
{
"model_identifier": "m",
"output": """{output_1}"""
},
{
"model_identifier": "M",
"output": """{output_2}"""
}
}
Task
Evaluate the models based on the quality and relevance of their outputs, and select the model that generated the best output. Answer by providing the model identifier of the best model. We will use your output as the name of the best model, so make sure your output only contains one of the following model identifiers and nothing else (no quotes, no spaces, no new lines, ...): m or M.
Best Model Identifier
< $|im_end|$ >

Table 27 | The CLUEWSC (Chinese Language Understanding Evaluation Benchmark - Winograd Schema Challenge) is a specialized task within the CLUE benchmark suite designed to evaluate a model's commonsense reasoning and contextual understanding capabilities in Chinese.

PROMPT

请参考示例的格式，完成最后的测试题。
下面是一些示例：他伯父还有许多女弟子，大半是富商财主的外室；这些财翁白天忙着赚钱，怕小公馆里的情妇长日无聊，要不安分，常常叫她们学点玩艺儿消遣。
上面的句子中的"她们"指的是
情妇
耶律克定说到雁北义军时，提起韦大哥，就连声说不可挡、不可挡，似有谈虎色变之味。后来又听说粘罕在云中，特派人厚币卑词，要与'韦义士修好'。韦大哥斩钉截铁地回绝了，大义凛然，端的是条好汉。如今张孝纯也想结识他，几次三番派儿子张浚上门来厮缠，定要俺引他上雁门山去见韦大哥。
上面的句子中的"他"指的是
张浚
"你何必把这事放在心上？何况你的还不过是手稿，并没有发表出来。"龙点睛越发坦率："如果发表出来了，那倒也就算了。不过既然没发表出来，我何必还让它飘在外头呢？你给我找一找吧，我要收回。"
上面的句子中的"它"指的是
手稿
这个身材高大，曾经被孙光平拿着菜刀追赶得到处乱窜的年轻人，那天早晨穿上了全新的卡其布中山服，像一个城里来的干部似的脸色红润，准备过河去迎接他的新娘。
上面的句子中的"他"指的是
年轻人
负责接待我们的是两位漂亮的朝鲜女导游，身材高挑，露出比例完美的小腿。一个姓韩，一个姓金，自称小韩和小金。她们的中文说得毫无口音，言谈举止也相当亲切。
上面的句子中的"她们"指的是
两位漂亮的朝鲜女导游
下面是测试题，请在思考结束后（</think>后）用一句话输出答案，不要额外的解释。
崩龙珍夫妻康健和美；鞠琴十年前丧偶，两年前重结良缘，现在的老伴是一位以前未曾有过婚史的高级工程师；崩龙珍和鞠琴都尽量避免谈及自己的爱人，也尽量回避提及蒋盈波的亡夫屈晋勇——尽管她们对他都很熟悉；当然也绝不会愚蠢地提出蒋盈波今后是一个人过到底还是再找个老伴的问题来加以讨论，那无论如何还为时过早。
上面的句子中的"他"指的是

Evaluation

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