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.

Table 23 | Instruction-Following Evaluation (IFEval) is a benchmark designed to assess a model’s ability to comply with explicit, verifiable instructions embedded within prompts. It targets a core competency of large language models (LLMs): producing outputs that meet multiple, clearly defined constraints specified by the user.

PROMPT

Kindly summarize the text below in XML format. Make sure the summary contains less than 4 sentences.

Quantum entanglement is the phenomenon that occurs when a group of particles are generated, interact, or share spatial proximity in such a way that the quantum state of each particle of the group cannot be described independently of the state of the others, including when the particles are separated by a large distance. The topic of quantum entanglement is at the heart of the disparity between classical and quantum physics: entanglement is a primary feature of quantum mechanics not present in classical mechanics.

Measurements of physical properties such as position, momentum, spin, and polarization performed on entangled particles can, in some cases, be found to be perfectly correlated. For example, if a pair of entangled particles is generated such that their total spin is known to be zero, and one particle is found to have clockwise spin on a first axis, then the spin of the other particle, measured on the same axis, is found to be anticlockwise. However, this behavior gives rise to seemingly paradoxical effects: any measurement of a particle’s properties results in an apparent and irreversible wave function collapse of that particle and changes the original quantum state. With entangled particles, such measurements affect the entangled system as a whole.

Such phenomena were the subject of a 1935 paper by Albert Einstein, Boris Podolsky, and Nathan Rosen, and several papers by Erwin Schrödinger shortly thereafter, describing what came to be known as the EPR paradox. Einstein and others considered such behavior impossible, as it violated the local realism view of causality (Einstein referring to it as "spooky action at a distance") and argued that the accepted formulation of quantum mechanics must therefore be incomplete.

Evaluation

Call official functions to check if the answer is consistent with the instructions.

Table 24 | FRAMES (Factuality, Retrieval, And reasoning MEasurement Set) is a comprehensive benchmark designed to evaluate core components of retrieval-augmented generation (RAG) systems. Our evaluation employs the benchmark's official "Oracle Prompt" configuration. In this setting, each test prompt includes the question along with all the ground truth Wikipedia articles, thus eliminating the need for an external retrieval component (e.g., BM25). This setting allows us to specifically measure a model's ability to reason over and synthesize information from provided sources to generate correct and verifiable facts.

PROMPT

Here are the relevant Wikipedia articles:
url: https:en.wikipedia.orgwikiPresident_of_the_United_States
url content: The president of the United States (POTUS) is the head of state and head of government of the United States of America. The president directs the executive branch of the federal government and is the commander-in-chief of the United States Armed Forces. . . .
Based on all the information, answer the query.
Query: If my future wife has the same first name as the 15th first lady of the United States' mother and her surname is the same as the second assassinated president's mother's maiden name, what is my future wife's name?

Evaluation

===Task===
I need your help in evaluating an answer provided by an LLM against a ground truth answer. Your task is to determine if the ground truth answer is present in the LLM's response. Please analyze the provided data and make a decision.
===Instructions===
1. Carefully compare the "Predicted Answer" with the "Ground Truth Answer".
2. Consider the substance of the answers - look for equivalent information or correct answers.
Do not focus on exact wording unless the exact wording is crucial to the meaning.
3. Your final decision should be based on whether the meaning and the vital facts of the "Ground Truth Answer" are present in the "Predicted Answer:"
===Input Data=== - Question: If my future wife has the same first name as the 15th first lady of the United States' mother and her surname is the same as the second assassinated president's mother's maiden name, what is my future wife's name?
- Predicted Answer: . . .
- Ground Truth Answer: Jane Ballou
===Output Format===
Provide your final evaluation in the following format:
Explanation: xxx
Decision: "TRUE" or "FALSE"
Please proceed with the evaluation.