Title: MMLU

URL: https://en.wikipedia.org/wiki/MMLU

PageID: 76696530

Categories: Category:Benchmarks (computing), Category:Large language models

-----

Measuring Massive Multitask Language Understanding (MMLU) is a popular benchmark for evaluating the capabilities of large language models . It inspired several other versions and spin-offs, such as MMLU-Pro, MMMLU and MMLU-Redux.

Overview

MMLU consists of 15,908 multiple-choice questions, with 1,540 of them being used to select and assess optimal settings for models – temperature, batch size and learning rate . The questions span across 57 subjects, from highly complex STEM fields and international law, to nutrition and religion. It was one of the most commonly used benchmarks for comparing the capabilities of large language models , with over 100 million downloads as of July 2024. [1] [2]

The benchmark was released by Dan Hendrycks and a team of researchers on 7 September 2020. It was purpose-made to be more challenging than existing benchmarks at the time, such as General Language Understanding Evaluation (GLUE), as models began outperforming humans in easier tests. When MMLU was released, most existing language models scored near the level of random chance (25%). The best performing model, GPT-3 175B, achieved 43.9% accuracy. The creators of the MMLU estimated that human domain-experts achieve around 89.8% accuracy. [1] By mid-2024, the majority of powerful language models such as Claude 3.5 Sonnet , GPT-4o and Llama 3.1 405B consistently achieved 88%. [3] [4] [5] As of 2025, MMLU has been partially phased out in favor of more difficult alternatives.

Limitations

On 5 June 2024, experts released a paper detailing their manual analysis of 5,700 questions in the benchmark, which revealed that it contained a very significant amount of ground-truth errors. For example, 57% of questions in the " Virology " subset were marked as harboring errors, such as multiple correct answers (4%), unclear questions (14%), or completely incorrect answers (33%). Overall, they estimated that 6.5% of questions in MMLU contained an error, suggesting the maximum attainable score was significantly below 100%. [6] Data contamination also posed a significant threat for this benchmark's validity; companies could easily include questions and answers into their models' training data, effectively rendering it ineffective. [7]

Examples

The following examples are sourced from the " Abstract Algebra ", " International Law " and "Professional Medicine " tasks, respectively. [1] The correct answers are marked in boldface:

Question 1:

Find all c {\displaystyle c} in Z 3 {\displaystyle \mathbb {Z} _{3}} such that Z 3 [x] / (x 2 + c) {\displaystyle \mathbb {Z} _{3}[x]/(x^{2}+c)} is a field.

(A) 0 ■ (B) 1 ■ (C) 2 ■ (D) 3

Question 2:

Would a reservation to the definition of torture in the International Covenant on Civil and Political Rights (ICCPR) be acceptable in contemporary practice?

(A) This is an acceptable reservation if the reserving country’s legislation employs a different definition. (B) This is an unacceptable reservation because it contravenes the object and purpose of the ICCPR. (C) This is an unacceptable reservation because the definition of torture in the ICCPR is consistent with customary international law. (D) This is an acceptable reservation because under general international law States have the right to enter reservations to treaties.

Question 3:

A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?

(A) Branch of the costocervical trunk. (B) Branch of the external carotid artery. (C) Branch of the thyrocervical trunk. (D) Tributary of the internal jugular vein.

References