MMLU

MMLU · general reasoning benchmark · 2020

Source: https://policywindow.org/wiki/mmlu

Generated 2026-07-09T20:33:54 UTC

Summary

Massive Multitask Language Understanding — 57-subject multiple-choice covering humanities, STEM, social sciences, professional/legal.

At a glance

Score range

0–100 % accuracy

Contamination risk

high

Methodology

https://arxiv.org/abs/2009.03300

Saturation

saturated

Details

Saturating — top models ~92%. Test-set leakage to training corpora is widely documented. MMLU-Pro is the harder successor. Currency (2026-06-21): Verified current. MMLU still saturated with top scores around 90 to 92 percent (GLM 5 about 91.7), matching the article 92 percent band and the saturated and high classifications. Gema et al label-error figures and the MMLU-Pro and MMLU-CF successor framing are confirmed. Only minor non-material additions exist (2026 contamination dose-response work, multilingual MMLU-ProX and IndicMMLU-Pro variants).

How to cite this article

APA

Policy Window. (2020). MMLU [Wiki article — Benchmark]. https://policywindow.org/wiki/mmlu

CHICAGO

Policy Window. 2020. "MMLU." Wiki article (Benchmark). https://policywindow.org/wiki/mmlu.

HARVARD

Policy Window (2020) 'MMLU', Wiki article — Benchmark, available at: https://policywindow.org/wiki/mmlu.

OSCOLA

Policy Window, 'MMLU' (Wiki article — Benchmark, 2020) <https://policywindow.org/wiki/mmlu> accessed [date].

BIBTEX

```
@misc{policywindow-mmlu,
title  = {MMLU},
author = {Policy Window},
year   = {2020},
howpublished = {MMLU (2020)},
url    = {https://policywindow.org/wiki/mmlu},
```

```
note   = {Primary source: https://arxiv.org/abs/2009.03300}
}
```