HumanEval

HUMANEVAL · code benchmark · 2021

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

Generated 2026-11-19T01:05:50 UTC

Summary

164 hand-written Python programming problems. Generate a function that passes provided unit tests.

At a glance

Score range

0–100 pass@1 %

Contamination risk

high

Methodology

https://arxiv.org/abs/2107.03374

Saturation

deprecated

Details

Saturated — top models ~95%. Largely superseded by SWE-bench for real-world relevance. Currency (2026-06-21): Verified-current — HumanEval remains saturated/deprecated as the article states; current pass@1 leaders (o4-mini ~97.3%, o3 ~97%, Claude Opus 4.6 ~96.3%) sit at the ceiling, consistent with the article's ">90% / ~95%" framing, and the field continues migrating to contamination-resistant successors (SWE-bench Verified/Pro, LiveCodeBench). No stale figure or claim.

How to cite this article

APA

Policy Window. (2021). HumanEval [Wiki article — Benchmark]. https://policywindow.org/wiki/humaneval

CHICAGO

Policy Window. 2021. "HumanEval." Wiki article (Benchmark). https://policywindow.org/wiki/humaneval.

HARVARD

Policy Window (2021) 'HumanEval', Wiki article — Benchmark, available at: https://policywindow.org/wiki/humaneval.

OSCOLA

Policy Window, 'HumanEval' (Wiki article — Benchmark, 2021) <https://policywindow.org/wiki/humaneval> accessed [date].

BIBTEX

```
@misc{policywindow-humaneval,
title  = {HumanEval},
author = {Policy Window},
year   = {2021},
howpublished = {HUMANEVAL (2021)},
url    = {https://policywindow.org/wiki/humaneval},
```

```
note   = {Primary source: https://arxiv.org/abs/2107.03374}
}
```