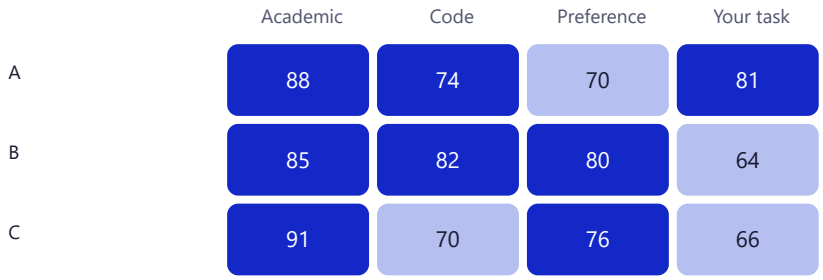
CHAPTER 10 / BENCHMARK

Different test. Different winner.

A score only answers the question its test asked.

A score only answers the question its test asked.



Illustrative scores · the table gives the exact values

Invented Models A, B and C; equal-weight averages only illustrate a ranking method, not a real leaderboard.

Conditions

Prompting, sampling and scoring conditions belong beside a benchmark result.

Coverage

Exams, code tests and preference votes measure different things.

Your task

Test representative work against criteria set before comparing candidates.

↗ **Prompting, sampling and scoring conditions belong beside a benchmark result.**

PUT IT TO WORK

Academic-only favors C; code-only favors B; the task column favors A. Cleaning changes only C's academic cell.

CLOSE THE PAGE. EXPLAIN IT.

What conditions, contamination risks and task match would you check before believing a score gap?