Reported general capability benchmarks for popular foundation models

Source: AI Index, 2025 | Table: 2025 AI Index report

Capability benchmark	o1	GPT-4.5	DeepSeek-R1	Gemini 2.5	Grok-2	Claude 3.7 Sonnet	Llama 3.3
MMLU, MMLU-Pro or MMMLU	✓	✓	✓	✓	✓	✓	✓
GPQA or GPQA-Diamond	✓	✓	✓	✓	✓	✓	✓
MATH-500	✓		✓		✓	✓	✓
AIME 2024	✓	✓	✓	✓		✓	
SWE-bench verified	✓	✓	✓	✓		✓	
MMMU	✓	✓		✓	✓	✓	