The smartest model running on a single GPU

Model	AIME24	AIME25	AIME26	GPQA Diamond	MATH-500	HLE
Giotto 1	86.7	83.3	93.3	85.4	99.6	18.4
Gemma 4	–	–	89.2	84.3	–	19.5
GPT-OSS-120B	80.4	80.0	–	73.1	97	8.6
NVIDIA Nemotron 3	–	89.1	–	75.7	98	10.6
DeepSeek R1 32B	72.6	63	–	62.1	94.3	–
Ministral 3	–	30	–	57.2	–	4.6

Gemma 4 : Gemma 4 31B
GPT-OSS-120B : GPT OSS 120B, medium reasoning
NVIDIA Nemotron 3 : NVIDIA Nemotron 3 Nano 30B-A3B
DeepSeek R1 32B : DeepSeek-R1-Distill-Qwen-32B
Ministral 3 : Ministral 3 14B Reasoning

Sources: official model cards reported by providers, or other sources from https://huggingface.co or artificialanalysis.ai.