CORE-Bench: LLM Reasoning Performance Comparison

Model	Overall Accuracy (%)
Claude Haiku 4.5	94.12%
Qwen 3 Next 80B	96.08%
DeepSeek V3.2	96.08%
Gemini 3 Flash	96.08%
Gemini 3 Pro	96.08%

95% threshold

Legend:
- ≥96% (Top Tier)
- <96%

Overall Accuracy (%)