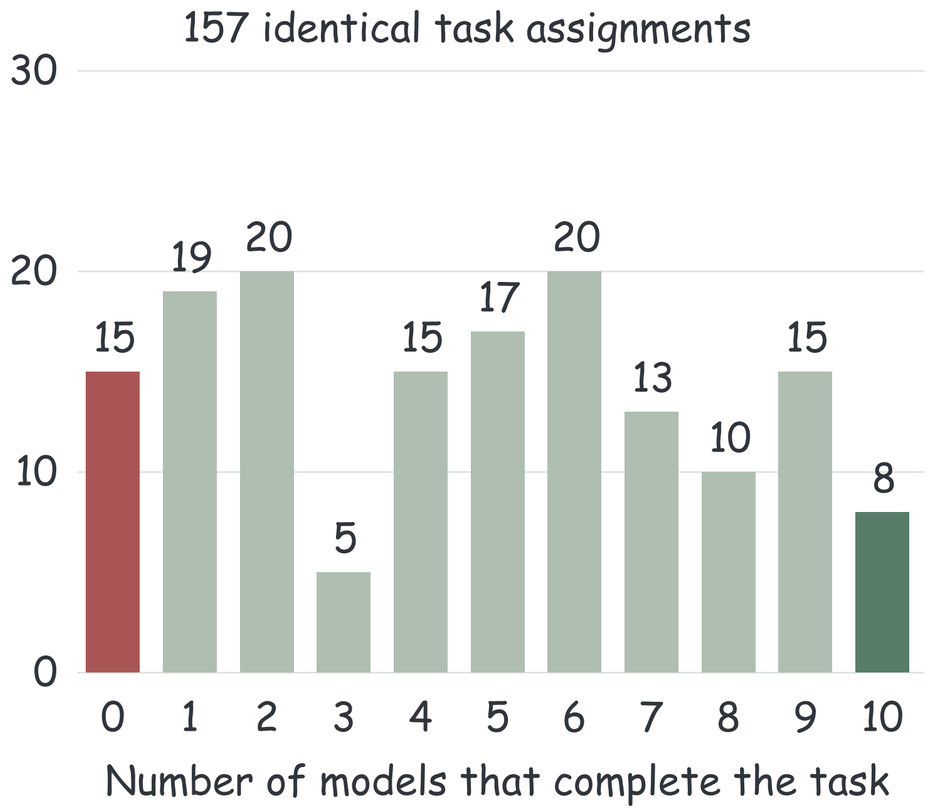
(a) Shared Planning outcomes



(b) Completion and token allocation

Model	Tasks completed (%)	Failed-episode tokens (%)
Astra	80.3	55.5
Sol	44.6	76.0
Fable 5.1	63.7	64.9
Gemini 3.8	48.4	77.1
Qwen Max	36.9	81.2
Qwen 397B	19.1	91.9
Qwen 27B	19.1	94.3
Opus 5.5	48.4	74.9
DeepSeek-1017	31.2	83.8
Gemini 4 Argon	70.1	64.0

Tokens: input + output, among episodes with recorded usage