DeepSeek Results

The tables presented below provide updated versions of Tables 2–5 from the original submission, incorporating results from the same experimental setup applied to the open-source models DeepSeek-V3 and DeepSeek-R1. Overall, the performance of DeepSeek-V3 and DeepSeek-R1 is comparable to, or slightly lower than, that of o1-mini. In these tables, "IO" refers to the direct query results reported in the original submission.

Table 1. Performance comparison of methods on 3-SAT and 4-SAT problems. Results are presented as $m + n$, where m represents correctly solved satisfiable formulas (out of **172**) and n denotes correctly identified unsatisfiable formulas (out of **27**).

Model	PEA	CoT	IO
GPT-4o	172+27	2+0	1+0
o1-mini	172+27	46+11	63+13
o1	172+27	100+24	99+24
DeepSeek-V3	172+27	23+4	15+5
DeepSeek-R1	172+27	37+8	28+3

Table 2. Performance comparison of methods on Game of 24. Results show the number of correctly assembled expressions out of **100** valid instances. ToT with o1 and o1-mini were omitted due to excessive query time (around 1 hour per instance) and poor performance (0 out of 10 in preliminary testing).

Model	PEA	CoT	IO	ToT
GPT-4o	100	2	2	4
o1-mini	100	29	44	–
o1	100	60	50	–
DeepSeek-V3	100	30	29	–
DeepSeek-R1	100	36	32	–

Table 3. Performance comparison of methods on cost-optimal Blocksworld task: Table displays the number of correctly solved instances out of **500** total tasks for various methods.

Model	PEA	CoT	IO
GPT-4o	500	115	94
o1-mini	500	185	182
o1	500	452	324
DeepSeek-V3	500	293	136
DeepSeek-R1	500	296	135

Table 4. Performance comparison on valid Logistics task: Results present the number of correctly solved instances out of **285** tasks. Note: GPT-4o PEA generates unoptimized code; 207 instances reach 30-second timeout and are terminated.

Model	**PEA**	**CoT**	**IO**
GPT-4o	78	44	28
o1-mini	285	60	65
o1	285	205	163
DeepSeek-V3	285	55	53
DeepSeek-R1	285	55	47