Model	Size	Open	Overall	Operation	Logic	Cipher	Puzzle	Counterfactua
Gpt-4o	*	Х	12.56	12.80	27.20	0.80	12.80	9.20(81.60)
Qwen2.5-72B-Instruct	72.7B	V	12.40	14.80	32.80	0.40	7.20	6.80(84.40)
Claude-3.5-Sonnet	*	X	11.04	10.40	27.20	0.00	8.40	9.20(80.40)
Meta-Llama-3.1-70B-Instruct	70B	✓	10.80	12.00	26.80	0.40	3.60	11.20(76.00)
Qwen2.5-32B-Instruct	32B	V	10.72	11.60	27.20	0.80	6.00	8.00(82.00)
DeepSeek-V2.5	236B	✓	10.48	12.00	24.40	0.80	4.40	10.80(77.60)
Yi-Large	*	X	10.32	10.80	28.40	0.40	5.60	6.40(81.60)
Mistral-Large-Instruct-2407	123B	✓	10.24	8.00	25.20	0.80	8.40	8.80(80.40)
Qwen2.5-7B-Instruct	7.61B	V	10.00	9.60	25.60	0.40	5.20	9.20(78.00)
Qwen2-72B-Instruct	72.71B	V	8.96	8.80	23.60	0.00	4.40	8.00(81.60)
Qwen2-7B-Instruct	7.07B	✓	8.16	7.20	20.80	0.40	2.40	10.00(73.60)
Meta-Llama-3.1-8B-Instruct	8B	V	7.60	5.60	19.20	0.00	1.60	11.60(72.00)
C4ai-Command-R-08-2024	32B	✓	7.28	5.20	15.60	0.40	2.00	13.20(70.80)
C4ai-Command-R-Plus-08-2024	104B	V	6.88	4.00	17.20	0.40	0.80	12.00(66.40)
Yi-1.5-9B-Chat	9B	V	6.08	6.80	10.40	0.00	2.40	10.80(71.20)
Mistral-7B-Instruct-v0.3	7B	V	4.48	2.40	8.00	0.00	0.80	11.20(70.80)
Table 14: Performance of No Rules Provided)	Model	s on K	OR-Ben	ich in Zer	o-Shot	Setting	with C	Only Question
No Rules Provided)								
No Rules Provided) Model	Size	Open	Overall	Operation	Logic	Cipher	Puzzle	Counterfactua
No Rules Provided) Model Gpt-40	Size *	Open X	Overall 29.92	Operation 24.80	Logic 43.20	Cipher 5.20	Puzzle 16.00	Counterfactua 60.40(19.60)
Model Gpt-40 Qwen2.5-72B-Instruct	Size * 72.7B	Open X	Overall 29.92 25.44	Operation 24.80 32.80	Logic 43.20 47.20	Cipher 5.20 4.00	Puzzle 16.00 8.80	Counterfactua 60.40(19.60) 34.40(54.80)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct	Size * 72.7B 32B	Open X	Overall 29.92 25.44 24.48	Operation 24.80 32.80 29.20	Logic 43.20 47.20 43.60	Cipher 5.20 4.00 4.40	Puzzle 16.00 8.80 7.60	Counterfactua 60.40(19.60) 34.40(54.80) 37.60(43.60)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct Mistral-Large-Instruct-2407	* 72.7B 32B 123B	Open X	Overall 29.92 25.44 24.48 22.48	Operation 24.80 32.80 29.20 18.00	43.20 47.20 43.60 36.80	5.20 4.00 4.40 2.80	Puzzle 16.00 8.80 7.60 11.60	Counterfactua 60.40(19.60) 34.40(54.80) 37.60(43.60) 43.20(30.80)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct Mistral-Large-Instruct-2407 Qwen2-72B-Instruc	* 72.7B 32B 123B 72.71B	Open X	Overall 29.92 25.44 24.48 22.48 21.92	Operation 24.80 32.80 29.20 18.00 24.40	43.20 47.20 43.60 36.80 44.00	Cipher 5.20 4.00 4.40 2.80 6.00	Puzzle 16.00 8.80 7.60 11.60 7.60	Counterfactua 60.40(19.60) 34.40(54.80) 37.60(43.60) 43.20(30.80) 27.60(61.20)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct Mistral-Large-Instruct-2407 Qwen2-72B-Instruct Yi-Large	* 72.7B 32B 123B 72.71B *	Open X V V X	Overall 29.92 25.44 24.48 22.48 21.92 21.12	Operation 24.80 32.80 29.20 18.00 24.40 14.40	43.20 47.20 43.60 36.80 44.00 32.80	5.20 4.00 4.40 2.80 6.00 3.20	Puzzle 16.00 8.80 7.60 11.60 7.60 8.40	Counterfactua 60.40(19.60) 34.40(54.80) 37.60(43.60) 43.20(30.80) 27.60(61.20) 46.80(21.60)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct Mistral-Large-Instruct-2407 Qwen2-72B-Instruc Yi-Large Meta-Llama-3.1-70B-Instruct	* 72.7B 32B 123B 72.71B * 70B	Open X V X	Overall 29.92 25.44 24.48 22.48 21.92 21.12 20.08	Operation 24.80 32.80 29.20 18.00 24.40 14.40 12.40	Logic 43.20 47.20 43.60 36.80 44.00 32.80 33.60	5.20 4.00 4.40 2.80 6.00 3.20 1.20	Puzzle 16.00 8.80 7.60 11.60 7.60 8.40 8.00	Counterfactua 60.40(19.60) 34.40(54.80) 37.60(43.60) 43.20(30.80) 27.60(61.20) 46.80(21.60) 45.20(22.00)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct Mistral-Large-Instruct-2407 Qwen2-72B-Instruc Yi-Large Meta-Llama-3.1-70B-Instruct DeepSeek-V2.5	* 72.7B 32B 123B 72.71B * 70B 236B	Open X V X X V	Overall 29.92 25.44 24.48 22.48 21.92 21.12 20.08 19.12	Operation 24.80 32.80 29.20 18.00 24.40 14.40 12.40 16.40	43.20 47.20 43.60 36.80 44.00 32.80 33.60 41.60	5.20 4.00 4.40 2.80 6.00 3.20 1.20 2.40	Puzzle 16.00 8.80 7.60 11.60 7.60 8.40 8.00 8.80	Counterfactu: 60.40(19.60) 34.40(54.80) 37.60(43.60) 43.20(30.80) 27.60(61.20) 46.80(21.60) 45.20(22.00) 26.40(53.20)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct Mistral-Large-Instruct-2407 Qwen2-72B-Instruc Yi-Large Meta-Llama-3.1-70B-Instruct DeepSeek-V2.5 Claude-3.5-Sonnet	* 72.7B 32B 123B 72.71B * 70B 236B *	Open X X X X X X	Overall 29.92 25.44 24.48 22.48 21.92 21.12 20.08 19.12 18.64	Operation 24.80 32.80 29.20 18.00 24.40 14.40 12.40 16.40 13.20	43.20 47.20 43.60 36.80 44.00 32.80 33.60 41.60 22.00	5.20 4.00 4.40 2.80 6.00 3.20 1.20 2.40 3.20	Puzzle 16.00 8.80 7.60 11.60 8.40 8.00 8.80 15.20	Counterfactua 60.40(19.60) 34.40(54.80) 37.60(43.60) 43.20(30.80) 27.60(61.20) 46.80(21.60) 45.20(22.00) 26.40(53.20) 39.60(28.00)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct Mistral-Large-Instruct-2407 Qwen2-72B-Instruct Yi-Large Meta-Llama-3.1-70B-Instruct DeepSeek-V2.5 Claude-3.5-Sonnet C4ai-Command-R-08-2024	Size * 72.7B 32B 123B 72.71B * 70B 236B * 32B	Open X X X X X X X	29.92 25.44 24.48 22.48 21.92 21.12 20.08 19.12 18.64 15.36	Operation 24.80 32.80 29.20 18.00 24.40 14.40 16.40 13.20 12.00	43.20 47.20 43.60 36.80 44.00 32.80 33.60 41.60 22.00 27.60	5.20 4.00 4.40 2.80 6.00 3.20 1.20 2.40 3.20 2.40	Puzzle 16.00 8.80 7.60 11.60 7.60 8.40 8.00 8.80 15.20 3.20	Counterfactua 60.40(19.60) 34.40(54.80) 37.60(43.60) 43.20(30.80) 27.60(61.20) 46.80(21.60) 45.20(22.00) 26.40(53.20) 39.60(28.00) 31.60(48.40)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct Mistral-Large-Instruct-2407 Qwen2-72B-Instruc Yi-Large Meta-Llama-3.1-70B-Instruct DeepSeek-V2.5 Claude-3.5-Sonnet C4ai-Command-R-08-2024 C4ai-Command-R-Plus-08-2024	Size * 72.7B 32B 123B 72.71B * 70B 236B * 32B 104B	Open X X X X X X	Overall 29.92 25.44 24.48 22.48 21.92 21.12 20.08 19.12 18.64 15.36 14.88	24.80 32.80 29.20 18.00 24.40 14.40 12.40 16.40 13.20 12.00 10.40	43.20 47.20 43.60 36.80 44.00 32.80 33.60 41.60 22.00 27.60 26.40	5.20 4.00 4.40 2.80 6.00 3.20 1.20 2.40 3.20 2.40 3.20	Puzzle 16.00 8.80 7.60 11.60 7.60 8.40 8.00 15.20 3.20 6.80	Counterfactua 60.40(19.60) 34.40(54.80) 37.60(43.60) 43.20(30.80) 27.60(61.20) 46.80(21.60) 45.20(22.00) 26.40(53.20) 39.60(28.00) 31.60(48.40) 27.60(54.80)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct Mistral-Large-Instruct-2407 Qwen2-72B-Instruc Yi-Large Meta-Llama-3.1-70B-Instruct DeepSeek-V2.5 Claude-3.5-Sonnet C4ai-Command-R-08-2024 C4ai-Command-R-Plus-08-2024 Qwen2.5-7B-Instruct	* 72.7B 32B 123B 72.71B * 70B 236B * 32B 104B 7.61B	Open X X X X X X X X X X X X X	Overall 29.92 25.44 24.48 22.48 21.92 21.12 20.08 19.12 18.64 15.36 14.88 14.64	Operation 24.80 32.80 29.20 18.00 24.40 14.40 12.40 16.40 13.20 12.00 10.40 17.20	43.20 47.20 43.60 36.80 44.00 32.80 33.60 41.60 22.00 27.60 26.40 30.40	5.20 4.00 4.40 2.80 6.00 3.20 1.20 2.40 3.20 2.40 3.20 3.60	Puzzle 16.00 8.80 7.60 11.60 7.60 8.40 8.00 8.80 15.20 3.20 6.80 2.40	Counterfactua 60.40(19.60) 34.40(54.80) 37.60(43.60) 43.20(30.80) 27.60(61.20) 46.80(21.60) 45.20(22.00) 39.60(28.00) 31.60(48.40) 27.60(54.80) 19.60(64.00)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct Mistral-Large-Instruct-2407 Qwen2-72B-Instruc Yi-Large Meta-Llama-3.1-70B-Instruct DeepSeek-V2.5 Claude-3.5-Sonnet C4ai-Command-R-08-2024 C4ai-Command-R-Plus-08-2024 Qwen2-5-7B-Instruct Qwen2-7B-Instruct	* 72.7B 32B 123B 72.71B * 70B 236B * 32B 104B 7.61B 7.07B	Open X X X X X X X X X X X X X	Overall 29.92 25.44 24.48 22.48 21.92 21.12 20.08 19.12 18.64 15.36 14.88 14.64 14.48	24.80 32.80 29.20 18.00 24.40 14.40 16.40 13.20 12.00 10.40 17.20 14.80	43.20 47.20 43.60 36.80 44.00 32.80 33.60 41.60 22.00 27.60 26.40 30.40 30.80	5.20 4.00 4.40 2.80 6.00 3.20 1.20 2.40 3.20 2.40 3.20 3.60 2.80	Puzzle 16.00 8.80 7.60 11.60 7.60 8.40 8.00 8.80 15.20 3.20 6.80 2.40 3.20	Counterfactu: 60.40(19.60) 34.40(54.80) 37.60(43.60) 43.20(30.80) 27.60(61.20) 46.80(21.60) 45.20(22.00) 26.40(53.20) 39.60(28.00) 27.60(54.80) 19.60(64.00) 20.80(66.80)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct Qwen2-5-32B-Instruct Mistral-Large-Instruct-2407 Qwen2-72B-Instruc Yi-Large Meta-Llama-3.1-70B-Instruct DeepSeek-V2.5 Claude-3.5-Sonnet C4ai-Command-R-Plus-08-2024 C4ai-Command-R-Plus-08-2024 Qwen2.5-7B-Instruct Qwen2.7B-Instruct Yi-1.5-9B-Chat	Size * 72.7B 32B 123B 72.71B * 70B 236B * 32B 104B 7.61B 7.07B 9B	Open X V X X V X V V V V V V V	Overall 29.92 25.44 24.48 22.48 21.92 21.12 20.08 19.12 18.64 15.36 14.88 14.64 14.48 14.08	Operation 24.80 32.80 29.20 18.00 24.40 14.40 12.40 16.40 13.20 12.00 10.40 17.20 14.80 15.20	43.20 47.20 43.60 36.80 44.00 32.80 27.60 22.00 27.60 30.40 30.80 26.40	5.20 4.00 4.40 2.80 6.00 3.20 1.20 2.40 3.20 2.40 3.20 2.80 2.80	Puzzle 16.00 8.80 7.60 11.60 7.60 8.40 8.00 8.80 15.20 3.20 6.80 2.40 3.20 3.60	Counterfactu: 60.40(19.60) 34.40(54.80) 37.60(43.60) 43.20(30.80) 27.60(61.20) 46.80(21.60) 45.20(22.00) 26.40(53.20) 39.60(28.00) 31.60(48.40) 27.60(54.80) 19.60(64.00) 20.80(66.80) 22.40(56.80)
Model Gpt-40 Qwen2.5-72B-Instruct Qwen2.5-32B-Instruct Mistral-Large-Instruct-2407 Qwen2-72B-Instruc Yi-Large Meta-Llama-3.1-70B-Instruct DeepSeek-V2.5 Claude-3.5-Sonnet C4ai-Command-R-08-2024 C4ai-Command-R-Plus-08-2024 Qwen2-5-7B-Instruct Qwen2-7B-Instruct	* 72.7B 32B 123B 72.71B * 70B 236B * 32B 104B 7.61B 7.07B	Open X X X X X X X X X X X X X	Overall 29.92 25.44 24.48 22.48 21.92 21.12 20.08 19.12 18.64 15.36 14.88 14.64 14.48	24.80 32.80 29.20 18.00 24.40 14.40 16.40 13.20 12.00 10.40 17.20 14.80	43.20 47.20 43.60 36.80 44.00 32.80 33.60 41.60 22.00 27.60 26.40 30.40 30.80	5.20 4.00 4.40 2.80 6.00 3.20 1.20 2.40 3.20 2.40 3.20 3.60 2.80	Puzzle 16.00 8.80 7.60 11.60 7.60 8.40 8.00 8.80 15.20 3.20 6.80 2.40 3.20	Counterfactua 60.40(19.60) 34.40(54.80) 37.60(43.60) 43.20(30.80) 27.60(61.20) 46.80(21.60) 45.20(22.00)