Model	**Size**	**Open**	**Overall**	**Operation**	**Logic**	**Cipher**	**Puzzle**	**Counterfactual**
Gpt-4o	*	✗	12.56	12.80	27.20	0.80	12.80	9.20(81.60)
Qwen2.5-72B-Instruct	72.7B	✔	12.40	14.80	32.80	0.40	7.20	6.80(84.40)
Claude-3.5-Sonnet	*	✗	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	✔	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	*	✗	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	✔	10.00	9.60	25.60	0.40	5.20	9.20(78.00)
Qwen2-72B-Instruct	72.71B	✔	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	✔	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	✔	6.88	4.00	17.20	0.40	0.80	12.00(66.40)
Yi-1.5-9B-Chat	9B	✔	6.08	6.80	10.40	0.00	2.40	10.80(71.20)
Mistral-7B-Instruct-v0.3	7B	✔	4.48	2.40	8.00	0.00	0.80	11.20(70.80)

Table 14: **Performance of Models on KOR-Bench in Zero-Shot Setting with Only Questions (No Rules Provided)**

Model	**Size**	**Open**	**Overall**	**Operation**	**Logic**	**Cipher**	**Puzzle**	**Counterfactual**
Gpt-4o	*	✗	29.92	24.80	43.20	5.20	16.00	60.40(19.60)
Qwen2.5-72B-Instruct	72.7B	✔	25.44	32.80	47.20	4.00	8.80	34.40(54.80)
Qwen2.5-32B-Instruct	32B	✔	24.48	29.20	43.60	4.40	7.60	37.60(43.60)
Mistral-Large-Instruct-2407	123B	✔	22.48	18.00	36.80	2.80	11.60	43.20(30.80)
Qwen2-72B-Instruc	72.71B	✔	21.92	24.40	44.00	6.00	7.60	27.60(61.20)
Yi-Large	*	✗	21.12	14.40	32.80	3.20	8.40	46.80(21.60)
Meta-Llama-3.1-70B-Instruct	70B	✔	20.08	12.40	33.60	1.20	8.00	45.20(22.00)
DeepSeek-V2.5	236B	✔	19.12	16.40	41.60	2.40	8.80	26.40(53.20)
Claude-3.5-Sonnet	*	✗	18.64	13.20	22.00	3.20	15.20	39.60(28.00)
C4ai-Command-R-08-2024	32B	✔	15.36	12.00	27.60	2.40	3.20	31.60(48.40)
C4ai-Command-R-Plus-08-2024	104B	✔	14.88	10.40	26.40	3.20	6.80	27.60(54.80)
Qwen2.5-7B-Instruct	7.61B	✔	14.64	17.20	30.40	3.60	2.40	19.60(64.00)
Qwen2-7B-Instruct	7.07B	✔	14.48	14.80	30.80	2.80	3.20	20.80(66.80)
Yi-1.5-9B-Chat	9B	✔	14.08	15.20	26.40	2.80	3.60	22.40(56.80)
Mistral-7B-Instruct-v0.3	7B	✔	11.44	9.60	25.60	1.60	1.60	18.80(62.00)
Meta-Llama-3.1-8B-Instruct	8B	✔	10.88	2.80	13.60	0.80	0.00	37.20(21.20)

Table 15: **Performance of Models on KOR-Bench in Three-Shot Setting with Only Questions (No Rules Provided)**