		GSM8K	**SVAMP**	**HotpotQA**	**Sports**	**LLC**	**Average**
GPT-3.5-0613	Standard Prompt	74.9	82.2	51.0	75.6	68.0	70.3
	+ Critical Prompt (Huang et al., 2023)	74.1	80.0	47.0	53.6	76.0	66.1
	+ IoE Prompt (Ours)	77.1	81.9	**55.0**	**77.1**	74.0	73.0
	+ IoE Prompt + Decision (Ours)	**78.5**	**83.3**	53.0	76.5	**77.3**	**73.7**
GPT-3.5-1106	Standard Prompt	80.1	82.9	61.0	74.1	41.3	67.9
	+ Critical Prompt (Huang et al., 2023)	77.3	81.5	54.0	68.4	40.7	64.4
	+ IoE Prompt (Ours)	80.9	83.2	62.0	**75.7**	38.7	68.1
	+ IoE Prompt + Decision (Ours)	**82.3**	**84.2**	**63.0**	74.7	**44.7**	**69.8**
GPT-4	Standard Prompt	92.5	92.8	68.0	80.7	91.3	85.1
	+ Critical Prompt (Huang et al., 2023)	88.4	89.5	62.0	82.9	89.9	82.5
	+ IoE Prompt (Ours)	93.4	**93.2**	**70.0**	83.1	93.3	86.6
	+ IoE Prompt + Decision (Ours)	**93.6**	93.1	**70.0**	**83.3**	**94.7**	**86.9**
Mistral-Medium	Standard Prompt	84.8	85.7	67.0	75.6	60.7	74.8
	+ Critical Prompt (Huang et al., 2023)	62.5	74.5	65.0	51.0	35.4	57.7
	+ IoE Prompt (Ours)	85.4	85.7	**68.0**	75.6	**61.3**	75.2
	+ IoE Prompt + Decision (Ours)	**85.6**	**85.8**	**68.0**	**75.9**	**61.3**	**75.3**