Model	Vision	CMSC				CMCC				PCC				OC				MC			
		P	B	I	PBI	P	B	I	PBI	P	B	I	PBI	P	B	I	PBI	P	B	I	PBI
Human	✓	98.3	98.3	85.0	85.0	98.3	88.3	85.0	78.3	96.7	98.3	86.7	83.3	98.3	98.3	81.7	81.7	100.0	96.7	73.3	73.3
Llama-3.2-11B-Vision-Instruct		82.5	22.5	13.5	2.2	64.2	0.0	17.2	0.0	55.5	60.5	59.5	20.2	55.5	60.5	33.5	10.2	65.8	35.5	22.0	5.2
Llama-3.2-11B-Vision-Instruct	✓	87.8	25.2	5.2	1.0	65.2	1.0	10.8	0.0	73.8	56.2	67.8	27.3	73.8	56.2	21.5	8.0	61.5	40.2	20.2	5.0
Llama-3.2-1B-Instruct		50.2	23.8	30.2	3.0	48.2	24.5	28.5	4.5	51.0	36.2	51.0	11.2	51.0	36.2	27.0	6.2	53.8	22.2	23.2	3.0
Llama-3.2-3B-Instruct		63.7	18.2	20.8	2.5	65.2	0.0	18.0	0.0	63.5	63.0	80.2	31.8	63.5	63.0	20.0	8.8	46.2	30.0	22.8	3.5
Mistral-7B-Instruct-v0.3		72.8	24.5	16.0	4.0	46.2	0.0	15.8	0.0	60.5	61.3	66.8	22.2	60.5	61.3	26.0	8.2	25.0	46.5	22.2	2.2
Molmo-7B-D-0924		56.0	23.8	38.8	5.2	59.2	4.8	34.2	1.0	44.5	44.0	37.5	7.2	44.5	44.0	17.8	2.8	51.0	36.8	10.2	1.2
Molmo-7B-D-0924	✓	43.5	29.8	33.0	5.0	51.0	4.0	28.0	0.2	41.5	44.0	41.0	6.8	41.5	44.0	16.0	3.8	51.0	39.0	14.8	2.2
Qwen2-VL-7B-Instruct		35.8	17.0	1.8	0.0	12.8	0.0	4.2	0.0	32.2	66.8	67.2	14.5	32.2	66.8	39.2	8.2	0.0	42.2	10.2	0.0
Qwen2-VL-7B-Instruct	✓	84.5	16.0	0.8	0.0	55.0	0.0	5.5	0.0	85.0	71.0	65.8	41.5	85.0	71.0	38.8	23.0	32.2	39.0	10.2	1.2
claude-3.5-haiku-20241022		68.8	84.2	26.0	16.5	69.2	3.2	76.2	2.5	11.2	92.2	76.5	7.8	11.2	92.2	23.5	2.0	72.8	100.0	4.2	3.5
claude-3.5-haiku-20241022	✓	5.2	75.8	23.2	0.5	12.0	3.0	70.0	0.0	4.8	91.8	81.2	4.2	4.8	91.8	17.5	0.2	46.5	100.0	3.0	1.5
claude-3.5-sonnet		85.2	100.0	49.0	42.5	75.2	21.2	75.5	13.2	64.5	99.8	95.5	61.8	64.5	99.8	95.2	61.3	84.5	100.0	85.2	71.8
claude-3.5-sonnet	✓	16.5	91.0	20.5	6.2	18.2	13.0	75.2	3.5	18.8	99.8	97.8	18.2	18.8	99.8	97.8	18.8	24.8	98.8	49.8	11.8
gemini-2.5-flash		8.2	7.2	0.0	0.0	10.0	0.2	2.0	0.0	14.8	4.0	0.2	0.0	14.8	4.0	0.2	0.0	13.0	2.0	0.2	0.0
gemini-2.5-flash	✓	6.0	10.0	1.8	0.2	6.8	2.8	6.2	0.0	20.8	24.0	14.5	1.2	20.8	24.0	4.2	0.2	12.2	12.8	6.5	0.0
gemini-flash-1.5		45.0	100.0	16.5	7.5	15.2	1.8	57.5	0.0	15.5	78.8	35.8	4.5	15.5	78.8	29.5	4.2	41.5	99.8	6.2	2.8
gemini-flash-1.5	✓	21.0	93.2	26.0	5.5	12.2	1.8	46.8	0.0	30.0	84.2	43.2	8.8	30.0	84.2	39.5	9.8	29.2	100.0	10.0	3.5
gemini-pro-1.5		71.8	100.0	48.5	34.5	44.2	75.8	70.0	24.8	12.8	92.5	86.5	10.2	12.8	92.5	76.2	9.0	54.5	100.0	30.8	17.8
gemini-pro-1.5	✓	30.2	99.2	29.2	11.5	19.5	21.2	55.5	2.8	21.0	90.2	88.5	15.5	21.0	90.2	78.5	14.8	35.8	100.0	18.0	6.8
gemma-2-9b-it		99.0	81.2	1.5	1.2	38.8	32.5	24.5	3.2	69.2	74.8	42.0	22.0	69.2	74.8	39.0	18.5	99.2	96.5	24.2	22.5
gpt-4o-2024-11-20		98.0	98.8	41.2	39.8	88.0	19.8	54.8	11.2	60.0	99.8	97.2	58.5	60.0	99.8	74.2	45.8	93.2	99.8	55.5	51.7
gpt-4o-2024-11-20	✓	14.2	35.5	26.2	9.5	16.8	10.5	32.0	1.0	27.5	99.2	90.2	25.8	27.5	99.2	86.2	22.2	0.0	100.0	0.0	0.0
gpt-4o-mini-2024-07-18		65.5	100.0	35.8	23.8	58.8	25.8	47.5	7.8	16.0	86.2	57.2	9.8	16.0	86.2	10.0	2.2	60.0	99.8	6.8	4.2
gpt-4o-mini-2024-07-18	✓	23.0	98.8	34.0	7.8	33.8	9.0	59.8	1.8	16.8	86.5	64.0	10.8	16.8	86.5	11.2	2.0	25.2	100.0	6.8	2.0
llama-3.2-90b-vision-instruct		30.2	75.8	26.2	6.2	6.5	0.0	43.2	0.0	8.2	91.2	66.0	4.2	8.2	91.2	16.5	2.2	52.2	100.0	3.2	1.0
llama-3.2-90b-vision-instruct	✓	9.2	55.8	23.0	0.8	1.8	4.8	78.8	0.0	14.0	72.5	82.8	8.5	14.0	72.5	42.5	3.5	25.8	66.8	17.8	3.2
o4-mini-2025-04-16	✓	14.8	69.0	25.2	8.8	15.5	17.0	34.2	4.0	33.5	100.0	95.5	32.8	33.5	100.0	86.8	25.5	33.5	100.0	74.2	25.5
qwen-2-vl-72b-instruct		59.2	91.0	3.2	2.5	7.8	0.2	21.2	0.0	50.0	76.2	52.2	17.8	50.0	76.2	50.5	18.8	36.2	99.5	8.5	2.2
qwen-2-vl-72b-instruct	✓	21.2	89.5	7.8	1.8	0.8	6.5	26.8	0.0	49.5	78.5	50.7	17.0	49.5	78.5	51.7	21.8	13.2	99.8	7.5	1.0

Table 1: Models' accuracy across the three question types (P: Perception, B: Belief, I: Intent) for each task in .