

A REVIEWER ANALYSIS

model	Full Set Acc	Hand Annotated Acc	Change	HITL Generated Acc	Change	Multi Image Acc	Change	Single Image Acc	Change
GPT-o4-mini	0.52	0.53	(0.7%)	0.60	(7.5%)	0.28	(-24.3%)	0.58	(5.8%)
GPT-5 Mini	0.52	0.50	(-2.0%)	0.59	(6.8%)	0.32	(-19.5%)	0.57	(4.7%)
GPT-5	0.49	0.52	(2.3%)	0.55	(5.5%)	0.30	(-19.7%)	0.54	(4.7%)
GPT-4.1	0.45	0.47	(2.1%)	0.52	(7.0%)	0.21	(-24.2%)	0.51	(5.8%)
GPT-4o	0.41	0.41	(0.2%)	0.48	(7.5%)	0.17	(-24.0%)	0.46	(5.7%)
GPT-5 Nano	0.37	0.35	(-2.1%)	0.44	(7.1%)	0.17	(-20.1%)	0.42	(4.8%)
Claude Opus 4	0.28	0.36	(7.9%)	0.31	(3.0%)	0.10	(-17.6%)	0.32	(4.2%)
Qwen2 2B	0.26	0.23	(-3.4%)	0.33	(6.5%)	0.09	(-17.1%)	0.30	(4.1%)
Claude Sonnet 4	0.25	0.34	(9.2%)	0.27	(2.4%)	0.08	(-17.0%)	0.29	(4.1%)
Claude 3.5 Haiku	0.22	0.32	(9.6%)	0.24	(1.3%)	0.09	(-13.6%)	0.26	(3.5%)
Llama-3 11B	0.21	0.25	(4.0%)	0.24	(2.7%)	0.09	(-12.5%)	0.24	(3.0%)
Pixtral 12B	0.16	0.24	(7.8%)	0.16	(0.5%)	0.06	(-9.8%)	0.18	(2.3%)
LLaVA-Next 72B	0.14	0.20	(6.3%)	0.16	(2.0%)	0.01	(-12.7%)	0.17	(3.0%)
LLaVA-Next 34B	0.10	0.15	(4.9%)	0.12	(1.3%)	0.01	(-9.1%)	0.13	(2.2%)
LN 7B (Mistral)	0.10	0.13	(3.0%)	0.11	(1.3%)	0.03	(-7.1%)	0.12	(1.7%)
LN 7B (Vicuna)	0.08	0.14	(5.4%)	0.09	(0.5%)	0.01	(-7.2%)	0.10	(1.7%)
LN 13B (Vicuna)	0.08	0.10	(2.6%)	0.09	(1.5%)	0.01	(-7.2%)	0.10	(1.7%)
Phi-3.5 Vision	0.05	0.04	(-0.9%)	0.06	(0.7%)	0.04	(-1.3%)	0.05	(0.3%)

Table 3: MapQA Subcategory Accuracy (Part 1). LLaVa-Next is abbreviated as LN for spacing. Human-in-the-Loop is abbreviated as HITL.

model	Military		Natural World		Urban		Aviation	
	Acc	Change	Acc	Change	Acc	Change	Acc	Change
GPT-o4-mini	0.64	(11.2%)	0.51	(-1.4%)	0.64	(11.1%)	0.30	(-22.4%)
GPT-5 Mini	0.61	(9.6%)	0.52	(0.2%)	0.57	(5.1%)	0.33	(-19.1%)
GPT-5	0.59	(10.0%)	0.45	(-4.5%)	0.61	(11.6%)	0.32	(-17.5%)
GPT-4.1	0.57	(12.0%)	0.42	(-2.7%)	0.52	(6.9%)	0.23	(-21.7%)
GPT-4o	0.52	(11.7%)	0.38	(-2.2%)	0.46	(5.7%)	0.19	(-21.2%)
GPT-5 Nano	0.45	(7.7%)	0.41	(3.9%)	0.42	(4.8%)	0.18	(-18.9%)
Claude Opus 4	0.34	(6.0%)	0.27	(-1.1%)	0.47	(19.4%)	0.13	(-14.8%)
Qwen2 2B	0.34	(8.3%)	0.27	(0.4%)	0.33	(6.7%)	0.09	(-17.4%)
Claude Sonnet 4	0.31	(6.2%)	0.23	(-1.6%)	0.44	(18.8%)	0.10	(-14.6%)
Claude 3.5 Haiku	0.28	(5.4%)	0.21	(-0.9%)	0.39	(17.1%)	0.10	(-12.1%)
Llama-3 11B	0.29	(7.6%)	0.15	(-5.7%)	0.33	(11.7%)	0.09	(-12.0%)
Pixtral 12B	0.22	(6.5%)	0.08	(-7.6%)	0.33	(16.9%)	0.06	(-9.4%)
LLaVA-Next 72B	0.19	(5.4%)	0.11	(-2.8%)	0.32	(18.0%)	0.02	(-11.7%)
LLaVA-Next 34B	0.14	(3.7%)	0.08	(-2.2%)	0.24	(13.8%)	0.02	(-8.1%)
LN 7B (Mistral)	0.13	(2.6%)	0.09	(-0.5%)	0.20	(10.5%)	0.03	(-6.9%)
LN 7B (Vicuna)	0.11	(2.8%)	0.06	(-2.7%)	0.24	(15.5%)	0.02	(-6.5%)
LN 13B (Vicuna)	0.11	(2.7%)	0.07	(-0.7%)	0.18	(10.2%)	0.01	(-6.7%)
Phi-3.5 Vision	0.05	(-0.3%)	0.08	(3.0%)	0.02	(-3.3%)	0.04	(-1.2%)

Table 4: MapQA Subcategory Accuracy (Part 2). LLaVa-Next is abbreviated as LN for spacing. Human-in-the-Loop is abbreviated as HITL.