

Supplementary Material

1 Comprehensive Multilingual Experiments Tables

We report all implemented judge backbones with complete multilingual runs. Satisfaction tables report per-requirement `SAT./UNSAT.` and are split into two side-by-side task blocks to reduce appendix length. Metrics tables report task-level means over requirements using `total_time`, `input_tokens`, and `output_tokens`.

1.1 GPT-4o-2024-08-06

1.1.1 English

Requirement-level satisfaction for English under GPT-4o-2024-08-06.

English tasks 01–10					English tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	SAT.	SAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	SAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
02	R4	SAT.	SAT.	SAT.	12	R4	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	SAT.
	R1	SAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
03	R3	SAT.	SAT.	SAT.		R2	SAT.	SAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
04	R0	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	SAT.	13	R5	SAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
05	R4	UNSAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R2	UNSAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.	14	R3	UNSAT.	UNSAT.	UNSAT.
06	R2	UNSAT.	SAT.	SAT.		R4	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	SAT.
	R4	SAT.	SAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
07	R0	UNSAT.	UNSAT.	UNSAT.	15	R0	SAT.	SAT.	SAT.
	R1	SAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
08	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R0	UNSAT.	SAT.	UNSAT.		R5	UNSAT.	UNSAT.	SAT.
09	R1	SAT.	UNSAT.	UNSAT.	16	R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	SAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
10	R5	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.	17	R4	UNSAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
11	R3	SAT.	UNSAT.	SAT.		R0	UNSAT.	UNSAT.	SAT.
	R4	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
12	R5	SAT.	SAT.	SAT.		R2	UNSAT.	SAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.	18	R3	UNSAT.	SAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	UNSAT.
13	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.	19	R0	SAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	SAT.
14	R6	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
15	R2	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.	20	R6	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
16	R6	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	UNSAT.	SAT.
17	R2	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
18	R5	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.
19	R1	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
20	R4	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.

English tasks 21–30					English tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	SAT.	SAT.	SAT.	31	R0	SAT.	SAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	UNSAT.
22	R6	UNSAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R2	UNSAT.	SAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	SAT.
	R4	UNSAT.	UNSAT.	SAT.		R5	SAT.	UNSAT.	UNSAT.
23	R5	SAT.	SAT.	SAT.	33	R6	UNSAT.	SAT.	SAT.
	R6	SAT.	SAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	SAT.		R4	SAT.	SAT.	SAT.
24	R4	UNSAT.	SAT.	SAT.	34	R5	UNSAT.	SAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	SAT.	SAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	SAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	SAT.	SAT.	SAT.
	R4	SAT.	UNSAT.	SAT.		R5	SAT.	SAT.	SAT.
	R5	UNSAT.	SAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
26	R2	SAT.	UNSAT.	UNSAT.	36	R3	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	SAT.		R5	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	UNSAT.	UNSAT.
	R0	SAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
27	R1	UNSAT.	SAT.	UNSAT.	37	R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	SAT.	SAT.		R5	UNSAT.	SAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	UNSAT.	SAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R1	UNSAT.	SAT.	SAT.
	R2	UNSAT.	SAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	UNSAT.	SAT.		R5	SAT.	SAT.	SAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	SAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	SAT.		R0	SAT.	UNSAT.	SAT.
	R2	SAT.	UNSAT.	SAT.		R1	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	UNSAT.		R3	SAT.	SAT.	SAT.
30	R5	UNSAT.	UNSAT.	SAT.	40	R4	UNSAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R3	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	UNSAT.		R3	UNSAT.	SAT.	SAT.
	R5	UNSAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
						R6	UNSAT.	UNSAT.	UNSAT.

English tasks 41–50					English tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	UNSAT.	SAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	SAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	SAT.	UNSAT.	SAT.		R1	SAT.	SAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	SAT.	UNSAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	UNSAT.	SAT.	SAT.
	R0	UNSAT.	SAT.	UNSAT.		R0	SAT.	SAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
	R3	SAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
43	R4	UNSAT.	UNSAT.	UNSAT.	54	R4	UNSAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	SAT.		R1	SAT.	SAT.	SAT.
	R1	SAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R3	SAT.	UNSAT.	SAT.
	R3	SAT.	SAT.	SAT.		R4	SAT.	UNSAT.	SAT.
44	R4	SAT.	SAT.	SAT.	55	R5	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	SAT.	UNSAT.
	R0	SAT.	SAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	SAT.	SAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
45	R3	SAT.	UNSAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	SAT.		R6	SAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	SAT.		R7	UNSAT.	SAT.	UNSAT.
46	R2	UNSAT.	UNSAT.	SAT.		R8	UNSAT.	SAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R9	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R10	SAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	UNSAT.		R11	SAT.	SAT.	SAT.
	R6	SAT.	SAT.	UNSAT.		R12	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	UNSAT.		R13	UNSAT.	UNSAT.	SAT.
47	R1	UNSAT.	UNSAT.	UNSAT.	48	R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R2	UNSAT.	SAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	UNSAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	SAT.
48	R6	UNSAT.	UNSAT.	UNSAT.	49	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	SAT.	UNSAT.		R0	UNSAT.	SAT.	UNSAT.
	R1	SAT.	SAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R2	UNSAT.	SAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	SAT.	UNSAT.
	R4	SAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
49	R5	UNSAT.	SAT.	SAT.	50	R5	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	UNSAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	SAT.	SAT.					

Table 1: GPT-4o-2024-08-06 / English: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	7.29	1062	107	6.21	1033	90	6.49	1028	104
02	4.91	907	100	6.96	1067	108	6.11	942	98
03	6.87	1110	111	6.06	917	111	5.73	897	99
04	4.57	751	99	5.92	784	99	5.34	801	87
05	5.54	1008	105	5.80	902	103	6.11	923	104
06	5.59	1005	119	6.84	939	127	5.11	940	118
07	5.36	972	102	5.65	887	97	5.67	866	104
08	5.95	920	100	5.52	916	99	5.71	1001	108
09	5.66	849	94	5.48	979	98	5.19	955	110
10	5.68	929	95	5.71	960	105	5.92	987	114
11	6.79	1157	110	5.88	1148	104	5.93	1058	101
12	5.07	1056	98	7.12	1053	111	6.68	1043	107
13	5.89	1053	108	6.13	1041	111	6.06	978	104
14	6.12	1016	101	5.41	1023	99	5.30	954	101
15	6.63	976	95	5.33	989	93	4.90	937	93
16	6.40	1099	100	5.31	1013	98	5.19	1007	96
17	6.31	980	100	5.54	949	91	5.19	1045	102
18	5.67	944	93	5.80	810	91	4.81	893	81
19	7.40	1033	110	5.92	985	96	5.74	954	97
20	5.51	945	97	5.26	1091	103	5.87	1007	101
21	5.60	1090	96	5.38	1077	107	6.07	1031	103
22	6.53	957	93	4.87	1025	92	5.19	995	89
23	5.35	1053	97	5.09	993	95	5.05	1037	89
24	6.81	1127	109	5.31	983	102	8.73	1053	98
25	5.96	1042	102	5.53	973	111	5.52	1092	106
26	5.75	1015	114	5.41	1010	100	5.82	996	110
27	7.44	1018	96	5.49	974	100	5.32	940	104
28	5.44	949	114	6.18	910	99	5.55	881	99
29	6.46	945	98	5.76	1000	109	6.58	875	99
30	5.99	1005	98	5.62	932	98	7.08	918	103
31	5.09	1213	105	5.92	1062	114	6.52	1199	112
32	6.25	1052	112	5.71	924	100	5.53	986	109
33	5.48	974	103	5.52	1029	102	6.52	953	108
34	5.52	1030	102	6.07	1077	97	5.44	989	97
35	6.03	1031	95	5.11	1048	90	4.97	967	96
36	6.04	959	110	5.13	1019	102	5.42	1028	112
37	5.40	1034	116	5.48	1071	99	5.97	1024	113
38	6.31	837	96	5.36	839	98	6.18	934	99
39	5.57	1035	95	5.54	1123	103	5.18	1156	110
40	6.19	1028	108	6.26	1023	114	4.84	988	107
41	6.16	994	105	5.41	1043	104	5.50	987	110
42	6.41	1024	111	6.32	1043	106	5.91	1040	112
43	6.44	1061	106	5.81	1005	104	5.33	997	104
44	5.94	949	104	5.93	900	96	6.04	907	112
45	5.50	958	101	5.43	944	94	5.20	946	94
46	5.96	989	101	5.41	996	94	5.08	1047	104
47	5.38	938	108	6.17	975	103	6.19	878	101
48	5.98	1032	96	5.48	986	90	5.28	1000	102
49	5.74	1104	90	6.09	995	94	5.71	1086	109
50	5.35	974	101	5.13	1067	101	5.31	933	99
51	6.84	978	117	7.40	858	103	5.87	956	121
52	5.35	878	96	4.79	882	85	5.38	941	95
53	6.10	1027	104	6.18	1012	93	4.64	988	98
54	5.83	987	133	5.97	1088	120	5.84	1040	125
55	6.25	1428	122	6.40	1390	119	6.34	1351	125

1.1.2 Arabic

Requirement-level satisfaction for Arabic under GPT-4o-2024-08-06.

Arabic tasks 01–10					Arabic tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	SAT.	SAT.	SAT.	11	R0	SAT.	SAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	SAT.		R4	UNSAT.	SAT.	UNSAT.
02	R0	UNSAT.	UNSAT.	UNSAT.	12	R0	UNSAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
03	R0	SAT.	SAT.	SAT.	13	R5	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
04	R0	UNSAT.	SAT.	SAT.	14	R3	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	UNSAT.		R0	SAT.	SAT.	SAT.
05	R0	UNSAT.	UNSAT.	UNSAT.	15	R1	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
06	R0	SAT.	SAT.	SAT.	16	R6	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R0	UNSAT.	SAT.	UNSAT.
	R2	SAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
07	R0	SAT.	SAT.	SAT.	17	R4	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R8	SAT.	SAT.	SAT.
08	R0	UNSAT.	SAT.	SAT.	18	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
09	R0	UNSAT.	UNSAT.	UNSAT.	19	R5	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
10	R0	SAT.	SAT.	SAT.	20	R3	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.

Arabic tasks 21–30					Arabic tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	SAT.	SAT.	SAT.	31	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R0	SAT.	SAT.	UNSAT.	32	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
23	R0	UNSAT.	SAT.	SAT.	33	R0	SAT.	SAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
24	R0	UNSAT.	UNSAT.	UNSAT.	34	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
25	R0	SAT.	UNSAT.	UNSAT.	35	R0	SAT.	SAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
26	R0	UNSAT.	UNSAT.	UNSAT.	36	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	SAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R2	UNSAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
27	R0	UNSAT.	UNSAT.	UNSAT.	37	R0	SAT.	SAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R2	UNSAT.	SAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R1	SAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	UNSAT.
29	R0	UNSAT.	UNSAT.	UNSAT.	39	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	UNSAT.	SAT.	UNSAT.
	R5	SAT.	UNSAT.	SAT.		R5	UNSAT.	SAT.	UNSAT.
30	R0	UNSAT.	UNSAT.	UNSAT.	40	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	UNSAT.		R1	SAT.	SAT.	UNSAT.
	R2	SAT.	UNSAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	UNSAT.
31	R0	UNSAT.	UNSAT.	UNSAT.	41	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.

Arabic tasks 41–50					Arabic tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	SAT.	SAT.	SAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R1	SAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	UNSAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	UNSAT.
	R2	SAT.	UNSAT.	UNSAT.	54	R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
43	R5	UNSAT.	UNSAT.	UNSAT.	55	R5	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.	55	R0	UNSAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.	55	R3	UNSAT.	UNSAT.	UNSAT.
44	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.	55	R6	SAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R8	UNSAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.	55	R9	SAT.	SAT.	SAT.
45	R4	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R11	SAT.	SAT.	UNSAT.
	R0	SAT.	SAT.	SAT.	55	R12	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R13	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	SAT.	SAT.					
46	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	SAT.	SAT.	SAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
47	R3	SAT.	SAT.	SAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	SAT.	SAT.	SAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
48	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					
49	R0	SAT.	SAT.	SAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	SAT.	SAT.	SAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	SAT.	SAT.	SAT.					
50	R6	SAT.	SAT.	SAT.					
	R0	SAT.	SAT.	SAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
50	R5	UNSAT.	SAT.	UNSAT.					
	R6	SAT.	SAT.	SAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					
	R0	SAT.	SAT.	SAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					

Table 2: GPT-4o-2024-08-06 / Arabic: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	6.43	1111	90	5.66	1124	87	6.42	1143	89
02	5.03	1103	80	5.23	1156	85	5.83	1116	93
03	6.22	1114	79	5.08	1127	85	5.09	1125	84
04	5.59	808	72	5.00	838	82	5.18	819	77
05	5.16	1030	81	4.92	1029	83	6.52	1019	84
06	6.00	1125	76	4.97	1123	87	5.62	1107	91
07	5.67	1091	85	5.62	1090	87	5.35	1087	76
08	6.34	1118	85	5.28	1063	76	5.88	1081	82
09	4.81	1036	88	4.74	1031	80	4.86	1002	82
10	4.51	1038	83	5.44	1089	94	5.44	1034	76
11	6.01	1100	93	5.42	1175	76	4.92	1148	81
12	6.29	1138	92	4.72	1143	94	6.47	1153	94
13	5.97	1030	85	4.94	1057	82	4.90	1033	85
14	5.19	1148	77	4.27	1141	85	5.11	1143	75
15	5.00	1045	83	4.78	1037	84	5.23	1056	94
16	5.31	1125	83	4.49	1104	80	4.63	1115	90
17	6.13	1112	86	4.68	1118	85	5.07	1088	84
18	5.72	1020	90	4.60	1003	92	5.01	1013	94
19	5.86	1043	79	4.92	1048	90	5.16	1051	89
20	4.87	1129	82	4.95	1128	81	5.60	1103	90
21	5.90	1131	100	4.78	1144	85	5.07	1096	80
22	6.62	1079	95	4.90	1107	87	5.70	1056	92
23	6.23	1112	88	4.90	1224	84	5.94	1244	79
24	5.42	1085	81	4.60	1087	80	6.00	1107	83
25	5.17	1132	78	5.49	1152	93	6.97	1139	99
26	5.83	1100	87	4.64	1109	84	5.70	1110	90
27	6.28	1046	87	5.29	1041	89	5.94	1048	83
28	5.60	1055	80	5.33	1060	91	5.69	1073	80
29	5.01	1058	87	4.86	1062	82	5.44	1059	78
30	6.31	1012	93	4.93	1014	84	6.02	1014	94
31	5.63	1088	90	5.02	1104	87	6.17	1127	94
32	5.47	1091	82	4.67	1086	81	7.33	1116	86
33	5.07	1070	89	4.48	1087	87	5.34	1065	91
34	5.61	1081	81	4.84	1139	85	4.91	1065	90
35	6.22	1090	79	4.88	1075	81	6.94	1055	78
36	5.86	1039	81	4.72	1084	87	6.02	1069	81
37	5.73	1086	94	4.44	1080	86	6.29	1069	84
38	6.04	1028	88	4.61	1037	83	5.46	1021	88
39	6.11	1173	85	5.10	1171	92	4.74	1179	85
40	4.82	1062	87	4.49	1070	84	5.92	1061	88
41	5.50	1057	89	4.74	1069	89	5.12	1046	76
42	5.42	1043	80	4.71	1058	86	6.35	1053	88
43	5.12	1115	91	4.54	1112	82	6.47	1098	85
44	5.17	998	86	4.88	1003	97	5.25	1004	91
45	5.30	1011	80	4.77	1034	79	5.67	1020	89
46	6.03	1089	91	4.96	1062	84	5.10	1069	80
47	6.27	1016	85	4.90	1063	97	5.57	1038	88
48	6.69	1119	82	4.72	1130	80	5.44	1156	86
49	7.90	1160	90	4.63	1109	80	5.19	1072	82
50	16.32	1095	86	4.03	1102	68	5.12	1110	79
51	21.77	949	87	4.81	977	93	5.49	950	98
52	20.99	884	85	4.51	902	80	5.23	906	97
53	21.37	1018	96	5.66	1030	98	5.02	978	97
54	21.24	1214	97	5.28	1229	100	5.57	1223	94
55	21.33	1381	93	5.55	1397	94	5.12	1336	92

1.1.3 Turkish

Requirement-level satisfaction for Turkish under GPT-4o-2024-08-06.

Turkish tasks 01–10					Turkish tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	SAT.	UNSAT.	SAT.	11	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
02	R0	UNSAT.	UNSAT.	UNSAT.	12	R0	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R2	SAT.	UNSAT.	SAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	UNSAT.	SAT.
03	R0	SAT.	SAT.	SAT.	13	R5	SAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	SAT.	UNSAT.
	R3	SAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	SAT.		R2	SAT.	SAT.	SAT.
04	R0	SAT.	SAT.	SAT.	14	R3	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	SAT.		R0	SAT.	SAT.	SAT.
05	R0	UNSAT.	UNSAT.	UNSAT.	15	R1	UNSAT.	UNSAT.	SAT.
	R1	SAT.	UNSAT.	UNSAT.		R2	SAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R4	UNSAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	SAT.
06	R0	UNSAT.	UNSAT.	SAT.	16	R6	SAT.	SAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	UNSAT.		R2	SAT.	SAT.	UNSAT.
	R4	UNSAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	SAT.
07	R0	UNSAT.	SAT.	SAT.	17	R4	SAT.	SAT.	SAT.
	R1	SAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	SAT.		R7	SAT.	UNSAT.	SAT.
	R4	UNSAT.	SAT.	UNSAT.		R8	SAT.	SAT.	SAT.
08	R0	SAT.	SAT.	SAT.	18	R0	UNSAT.	SAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
09	R0	UNSAT.	UNSAT.	UNSAT.	19	R5	UNSAT.	SAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R6	UNSAT.	SAT.	SAT.
	R2	UNSAT.	SAT.	UNSAT.		R0	UNSAT.	SAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
10	R0	UNSAT.	UNSAT.	SAT.	20	R3	UNSAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	SAT.		R6	UNSAT.	UNSAT.	SAT.
	R4	UNSAT.	UNSAT.	SAT.					

Turkish tasks 21–30					Turkish tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	SAT.	SAT.	SAT.	31	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	UNSAT.	SAT.	UNSAT.
	R5	UNSAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	SAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	SAT.	SAT.
	R0	SAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	SAT.
	R4	UNSAT.	SAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
23	R5	SAT.	SAT.	SAT.	33	R6	UNSAT.	UNSAT.	UNSAT.
	R6	SAT.	SAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
24	R4	SAT.	SAT.	UNSAT.	34	R5	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	SAT.	SAT.
	R0	SAT.	SAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R2	SAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	UNSAT.		R3	UNSAT.	SAT.	SAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	UNSAT.	UNSAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	SAT.	UNSAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R0	UNSAT.	SAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
26	R2	SAT.	SAT.	SAT.	36	R3	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
	R4	SAT.	UNSAT.	UNSAT.		R5	SAT.	UNSAT.	SAT.
	R5	SAT.	UNSAT.	SAT.		R6	UNSAT.	UNSAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
27	R1	UNSAT.	UNSAT.	UNSAT.	37	R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	SAT.	UNSAT.		R6	UNSAT.	SAT.	UNSAT.
	R6	SAT.	UNSAT.	SAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	UNSAT.
	R2	UNSAT.	SAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	UNSAT.	SAT.	UNSAT.
30	R5	SAT.	SAT.	SAT.	40	R5	SAT.	SAT.	SAT.
	R6	UNSAT.	SAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R0	UNSAT.	SAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	SAT.
	R3	UNSAT.	SAT.	SAT.		R3	UNSAT.	SAT.	UNSAT.
31	R4	UNSAT.	SAT.	SAT.	41	R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	SAT.	SAT.
	R6	SAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R0	UNSAT.	SAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	SAT.

Turkish tasks 41–50					Turkish tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	SAT.	SAT.	SAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	SAT.	SAT.	52	R0	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R1	SAT.	UNSAT.	SAT.
	R5	UNSAT.	SAT.	UNSAT.		R2	SAT.	SAT.	SAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R0	UNSAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	SAT.	SAT.
	R4	UNSAT.	SAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
43	R5	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	SAT.	UNSAT.
	R0	SAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	SAT.
	R3	UNSAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R5	SAT.	UNSAT.	UNSAT.	55	R6	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	UNSAT.	SAT.		R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	SAT.	UNSAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	UNSAT.	SAT.	UNSAT.
45	R5	SAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	SAT.		R6	SAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R7	SAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R8	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	SAT.		R9	SAT.	UNSAT.	SAT.
	R4	UNSAT.	SAT.	UNSAT.		R10	UNSAT.	SAT.	SAT.
46	R5	UNSAT.	UNSAT.	UNSAT.		R11	SAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R12	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R13	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	UNSAT.	SAT.					
	R4	SAT.	UNSAT.	UNSAT.					
47	R5	SAT.	SAT.	SAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	SAT.	SAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	SAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	SAT.	SAT.					
48	R4	UNSAT.	UNSAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	SAT.	SAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
49	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	SAT.	UNSAT.					
	R4	UNSAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	SAT.	SAT.	SAT.					
	R7	SAT.	SAT.	SAT.					
50	R0	SAT.	SAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	SAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	SAT.					
	R5	SAT.	UNSAT.	SAT.					
	R6	UNSAT.	UNSAT.	SAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					

Table 3: GPT-4o-2024-08-06 / Turkish: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	6.59	1140	100	8.46	1078	100	6.37	1059	116
02	7.26	1100	127	6.15	1002	98	6.67	1006	100
03	6.03	1105	117	6.74	1160	118	5.83	1011	104
04	4.99	779	89	8.09	834	87	5.93	807	87
05	6.56	1103	96	6.94	1029	97	5.59	1031	99
06	6.41	954	89	6.14	943	103	5.43	1016	93
07	6.89	1049	124	5.79	955	112	6.32	956	112
08	6.25	972	99	5.91	896	87	5.46	1017	85
09	6.31	996	96	5.55	1053	97	6.41	922	96
10	6.41	964	104	5.21	1048	87	6.56	970	98
11	5.99	1125	105	5.56	1055	102	6.71	1080	97
12	5.08	1109	97	6.13	1064	104	5.43	1096	101
13	7.38	964	107	6.59	964	104	5.92	956	99
14	5.84	991	97	5.31	1008	86	6.21	1116	95
15	5.72	1072	91	5.49	971	95	5.36	933	98
16	5.59	931	98	5.80	1017	89	5.35	932	93
17	5.62	930	90	5.77	1052	108	5.79	1021	109
18	5.76	986	101	5.35	854	106	5.67	1040	112
19	5.80	957	98	6.36	999	104	5.66	965	118
20	5.98	1070	105	6.10	988	94	5.22	922	103
21	6.28	964	94	5.41	1005	96	6.16	1132	100
22	6.49	958	96	4.78	847	101	6.08	1025	103
23	5.90	1112	104	6.37	1317	100	6.15	1329	88
24	7.26	1050	97	6.49	1166	101	6.45	1070	116
25	5.74	1027	104	6.25	1105	90	6.02	988	95
26	6.32	983	92	5.93	1139	98	5.53	1042	98
27	6.36	1062	97	5.79	980	100	6.24	1031	102
28	5.33	965	95	5.55	1137	91	5.78	1030	97
29	6.46	964	116	6.07	978	119	6.23	1000	91
30	6.43	991	115	5.43	893	112	5.49	900	96
31	5.30	1182	92	5.62	1007	97	5.77	1033	101
32	5.72	1053	110	5.67	1016	97	6.29	920	104
33	6.24	1045	105	6.17	1059	116	5.43	1052	106
34	6.69	996	116	6.50	1004	99	5.81	1066	97
35	6.10	903	98	5.92	1025	101	5.91	982	101
36	5.80	941	100	6.17	982	119	5.74	1028	97
37	6.04	1043	109	6.12	1080	95	5.91	1036	103
38	6.20	1013	95	7.05	1050	94	5.23	928	96
39	6.02	1051	104	6.34	1089	97	6.33	997	112
40	5.96	1019	106	6.85	950	113	5.39	947	107
41	6.34	925	97	6.11	953	94	5.42	921	106
42	5.71	1011	104	7.14	960	103	5.52	1008	108
43	6.28	1079	121	7.26	1087	121	6.17	1064	103
44	6.53	998	102	6.33	964	104	5.36	884	92
45	6.59	1003	96	5.70	924	96	4.77	854	100
46	5.63	971	100	6.05	1013	96	5.02	1056	106
47	6.23	1012	101	6.07	976	99	5.29	881	109
48	6.17	1083	102	5.80	1102	95	5.36	1134	107
49	6.42	1103	116	6.64	1087	99	5.50	1089	104
50	5.67	1112	94	6.29	1108	98	4.61	1033	87
51	5.95	827	110	6.03	932	98	6.50	947	93
52	6.31	895	97	6.55	929	110	6.24	896	109
53	6.81	931	97	10.55	920	103	6.09	1012	121
54	6.71	1037	115	7.32	1179	116	5.22	1145	101
55	7.54	1339	114	22.03	1221	119	6.10	1175	112

1.1.4 Chinese

Requirement-level satisfaction for Chinese under GPT-4o-2024-08-06.

Chinese tasks 01–10					Chinese tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	SAT.	SAT.	SAT.	11	R0	UNSAT.	SAT.	SAT.
	R1	SAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	UNSAT.	SAT.
02	R0	UNSAT.	UNSAT.	UNSAT.	12	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	UNSAT.	SAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
03	R0	UNSAT.	UNSAT.	SAT.	13	R5	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R0	SAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R2	UNSAT.	SAT.	UNSAT.
04	R0	SAT.	SAT.	SAT.	14	R3	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R4	SAT.	SAT.	UNSAT.
	R2	SAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R6	SAT.	SAT.	UNSAT.
	R4	UNSAT.	SAT.	SAT.		R0	SAT.	SAT.	SAT.
05	R0	UNSAT.	UNSAT.	UNSAT.	15	R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R2	UNSAT.	SAT.	UNSAT.
	R2	SAT.	SAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R5	SAT.	UNSAT.	UNSAT.
06	R0	UNSAT.	UNSAT.	SAT.	16	R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R2	UNSAT.	SAT.	SAT.		R1	SAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	SAT.	SAT.
07	R0	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R6	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	UNSAT.		R7	SAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R8	SAT.	SAT.	SAT.
08	R0	UNSAT.	UNSAT.	UNSAT.	18	R0	UNSAT.	UNSAT.	SAT.
	R1	UNSAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	SAT.	SAT.
	R3	SAT.	UNSAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	SAT.	UNSAT.	UNSAT.
09	R5	SAT.	SAT.	SAT.	19	R5	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	UNSAT.	UNSAT.		R6	UNSAT.	SAT.	UNSAT.
	R1	UNSAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
10	R4	UNSAT.	UNSAT.	SAT.	20	R3	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.

Chinese tasks 21–30					Chinese tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	SAT.	SAT.	SAT.	31	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	UNSAT.		R3	UNSAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	SAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	UNSAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	SAT.		R2	SAT.	SAT.	SAT.
	R2	UNSAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
	R4	UNSAT.	UNSAT.	SAT.		R5	SAT.	UNSAT.	UNSAT.
23	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.	33	R0	UNSAT.	UNSAT.	SAT.
	R0	SAT.	UNSAT.	SAT.		R1	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	UNSAT.
24	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	SAT.
	R6	SAT.	SAT.	SAT.	34	R0	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R1	UNSAT.	SAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
25	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	UNSAT.	UNSAT.
	R5	SAT.	UNSAT.	UNSAT.	35	R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	SAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R1	UNSAT.	SAT.	UNSAT.		R2	UNSAT.	SAT.	UNSAT.
	R2	UNSAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
26	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
	R4	UNSAT.	UNSAT.	SAT.	36	R5	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	SAT.	UNSAT.		R6	UNSAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R1	UNSAT.	SAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	UNSAT.	SAT.
27	R2	SAT.	SAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.	37	R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	SAT.	UNSAT.	UNSAT.		R7	UNSAT.	SAT.	UNSAT.
	R0	SAT.	UNSAT.	UNSAT.		R0	SAT.	UNSAT.	SAT.
28	R1	UNSAT.	UNSAT.	SAT.	38	R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	SAT.		R3	UNSAT.	UNSAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	SAT.	UNSAT.
29	R0	UNSAT.	UNSAT.	UNSAT.	39	R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	SAT.	SAT.
	R4	UNSAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
30	R6	SAT.	UNSAT.	UNSAT.	40	R5	SAT.	SAT.	UNSAT.
	R0	UNSAT.	SAT.	SAT.		R6	SAT.	SAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R0	SAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R3	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.
31	R5	UNSAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	SAT.
	R6	UNSAT.	UNSAT.	SAT.		R5	SAT.	SAT.	UNSAT.
	R0	UNSAT.	UNSAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.					
	R2	UNSAT.	UNSAT.	SAT.					
	R3	UNSAT.	UNSAT.	SAT.					

Chinese tasks 41–50					Chinese tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	SAT.	SAT.	SAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
42	R6	UNSAT.	SAT.	SAT.	53	R3	UNSAT.	SAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R0	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	SAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
43	R5	UNSAT.	UNSAT.	SAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R3	SAT.	SAT.	SAT.		R4	SAT.	UNSAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R5	SAT.	UNSAT.	SAT.	55	R6	UNSAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R0	SAT.	SAT.	SAT.
	R1	UNSAT.	SAT.	UNSAT.		R1	UNSAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	SAT.	SAT.
	R3	SAT.	UNSAT.	SAT.		R3	UNSAT.	UNSAT.	SAT.
	R4	SAT.	UNSAT.	SAT.		R4	UNSAT.	SAT.	UNSAT.
45	R5	SAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R7	SAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R9	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	SAT.		R10	SAT.	SAT.	UNSAT.
46	R5	SAT.	UNSAT.	SAT.		R11	SAT.	SAT.	SAT.
	R0	UNSAT.	SAT.	UNSAT.		R12	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	SAT.		R13	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	SAT.					
	R3	SAT.	SAT.	SAT.					
	R4	SAT.	SAT.	SAT.					
47	R5	SAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	SAT.					
	R1	SAT.	SAT.	SAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	SAT.	SAT.					
48	R4	UNSAT.	UNSAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	SAT.	UNSAT.	UNSAT.					
	R7	SAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	SAT.					
49	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	SAT.	SAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	SAT.					
	R6	SAT.	SAT.	SAT.					
	R0	SAT.	UNSAT.	UNSAT.					
50	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	SAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	SAT.	SAT.	SAT.					
	R6	UNSAT.	SAT.	SAT.					
	R7	SAT.	UNSAT.	SAT.					

Table 4: GPT-4o-2024-08-06 / Chinese: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	6.44	1025	79	13.01	1046	92	6.16	1031	125
02	7.36	1015	124	6.08	923	102	6.17	998	138
03	7.09	961	113	5.61	800	106	5.33	876	90
04	5.41	705	93	5.80	714	79	5.33	706	75
05	6.36	899	126	5.66	863	95	5.63	929	97
06	6.18	1005	116	6.06	925	109	5.90	1000	151
07	6.00	985	130	6.22	947	115	5.99	975	117
08	7.35	961	123	5.64	857	89	5.50	957	98
09	6.19	958	91	6.72	949	130	6.00	807	94
10	6.75	971	124	5.55	936	98	5.49	944	112
11	6.06	1048	94	5.82	1100	122	6.07	1009	90
12	5.28	1016	92	6.09	1037	110	6.34	1003	102
13	6.71	904	100	6.23	859	93	5.81	891	106
14	6.17	939	100	6.38	1066	118	7.02	1042	102
15	6.56	932	109	5.29	903	93	5.68	892	97
16	6.20	961	103	5.82	976	120	5.56	1011	95
17	6.26	982	94	5.70	990	95	6.01	946	118
18	5.64	889	93	6.24	922	127	5.86	934	123
19	6.77	925	102	5.43	930	102	5.97	1017	106
20	5.60	986	111	6.05	1034	100	5.79	1019	117
21	6.63	936	107	6.52	1078	125	5.96	1042	109
22	6.38	981	116	5.88	949	89	5.88	871	93
23	5.73	1099	97	5.92	1086	97	5.79	1060	101
24	6.39	1007	95	6.29	1046	103	5.95	969	120
25	6.07	1009	119	5.44	1004	111	6.26	1033	123
26	5.21	1015	79	5.52	1010	100	5.40	936	95
27	6.27	956	114	6.38	947	114	6.10	973	103
28	5.85	993	97	5.07	869	82	6.02	972	96
29	6.20	956	122	6.24	990	115	6.11	878	118
30	6.74	867	108	5.86	893	96	5.84	873	115
31	7.23	1163	105	6.41	1147	94	6.94	1110	121
32	6.23	969	109	5.72	1019	88	5.34	970	83
33	5.40	808	93	6.30	1003	122	5.61	948	115
34	6.88	996	106	6.14	1060	109	6.50	1043	117
35	6.98	942	92	5.36	954	78	5.86	981	94
36	6.63	1017	113	5.98	1013	105	5.43	993	97
37	5.89	1075	104	5.67	991	114	5.42	1020	98
38	5.97	909	102	7.09	905	114	5.41	912	96
39	5.63	1015	104	5.96	1040	117	5.77	1048	105
40	6.06	921	106	6.20	953	115	5.31	916	92
41	5.87	952	110	6.23	982	127	5.72	976	104
42	5.16	906	83	5.93	931	101	6.36	903	134
43	6.66	976	126	5.20	962	105	5.78	1004	111
44	6.16	861	106	5.62	925	83	5.58	865	94
45	6.19	935	88	6.47	925	103	5.88	901	93
46	6.87	871	97	5.38	956	91	5.75	933	122
47	6.18	934	118	6.04	924	118	5.92	912	104
48	6.23	1075	96	5.72	1032	103	5.95	1058	103
49	6.51	1007	101	6.19	1016	114	5.88	997	101
50	5.49	1015	140	5.48	982	111	5.65	1020	103
51	5.72	921	83	6.36	888	109	6.05	905	121
52	5.21	784	83	6.76	867	115	7.05	856	91
53	5.37	870	78	5.90	879	79	5.40	873	101
54	6.57	1082	127	6.00	1111	116	6.47	1106	158
55	7.58	1349	169	6.61	1341	127	6.05	1286	119

1.1.5 Hindi

Requirement-level satisfaction for Hindi under GPT-4o-2024-08-06.

Hindi tasks 01–10					Hindi tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	SAT.	SAT.	SAT.	11	R0	SAT.	UNSAT.	SAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	SAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	UNSAT.	SAT.	UNSAT.
02	R0	UNSAT.	UNSAT.	UNSAT.	12	R5	SAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R2	UNSAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R3	SAT.	SAT.	SAT.		R2	SAT.	UNSAT.	SAT.
	R4	UNSAT.	SAT.	UNSAT.		R3	UNSAT.	SAT.	SAT.
03	R0	SAT.	SAT.	SAT.	13	R4	UNSAT.	UNSAT.	SAT.
	R1	SAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	SAT.		R0	UNSAT.	UNSAT.	SAT.
	R4	UNSAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	SAT.
04	R0	SAT.	SAT.	SAT.	14	R2	UNSAT.	SAT.	SAT.
	R1	UNSAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	UNSAT.
	R4	UNSAT.	SAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
05	R0	UNSAT.	UNSAT.	UNSAT.	15	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	SAT.	UNSAT.		R1	UNSAT.	UNSAT.	SAT.
	R2	SAT.	SAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	SAT.	SAT.	SAT.		R3	SAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	UNSAT.	SAT.	UNSAT.
06	R0	SAT.	SAT.	SAT.	16	R5	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	SAT.
07	R0	SAT.	SAT.	SAT.	17	R4	SAT.	SAT.	SAT.
	R1	SAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	SAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	SAT.		R8	SAT.	SAT.	SAT.
08	R0	UNSAT.	UNSAT.	UNSAT.	18	R0	SAT.	SAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R1	UNSAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	SAT.	UNSAT.	UNSAT.
09	R0	UNSAT.	SAT.	UNSAT.	19	R5	UNSAT.	SAT.	UNSAT.
	R1	SAT.	SAT.	UNSAT.		R0	SAT.	SAT.	UNSAT.
	R2	UNSAT.	SAT.	UNSAT.		R1	UNSAT.	UNSAT.	SAT.
	R3	UNSAT.	SAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	SAT.
10	R0	SAT.	SAT.	SAT.	20	R4	SAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R5	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.

Hindi tasks 21–30					Hindi tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	SAT.	SAT.	SAT.	31	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	SAT.	UNSAT.
	R3	SAT.	SAT.	UNSAT.		R3	SAT.	UNSAT.	SAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
22	R0	UNSAT.	SAT.	UNSAT.	32	R0	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	SAT.	SAT.
	R2	SAT.	UNSAT.	SAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	SAT.
	R4	SAT.	SAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
23	R0	SAT.	UNSAT.	SAT.	33	R0	SAT.	SAT.	UNSAT.
	R1	SAT.	UNSAT.	SAT.		R1	UNSAT.	SAT.	UNSAT.
	R2	UNSAT.	SAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	UNSAT.		R3	UNSAT.	UNSAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
24	R0	SAT.	SAT.	SAT.	34	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	SAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	SAT.
25	R0	UNSAT.	UNSAT.	UNSAT.	35	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	SAT.
	R2	SAT.	UNSAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	SAT.	UNSAT.		R4	SAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
26	R0	SAT.	SAT.	SAT.	36	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	SAT.		R3	SAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	SAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	SAT.		R5	SAT.	SAT.	SAT.
27	R0	SAT.	SAT.	SAT.	37	R0	SAT.	SAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	UNSAT.	UNSAT.		R2	SAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	SAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	SAT.
	R4	UNSAT.	SAT.	SAT.		R4	SAT.	UNSAT.	SAT.
	R5	SAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
29	R0	UNSAT.	UNSAT.	UNSAT.	39	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	SAT.		R3	SAT.	UNSAT.	SAT.
	R4	UNSAT.	SAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
30	R0	SAT.	SAT.	UNSAT.	40	R0	SAT.	SAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	SAT.
	R3	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	SAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	SAT.	UNSAT.
31	R0	UNSAT.	UNSAT.	UNSAT.	41	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.

Hindi tasks 41–50					Hindi tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	SAT.	SAT.	SAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	UNSAT.		R1	SAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R1	SAT.	SAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	SAT.	UNSAT.	SAT.
	R0	UNSAT.	SAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	SAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	SAT.
43	R5	UNSAT.	UNSAT.	SAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	SAT.
	R1	SAT.	SAT.	UNSAT.		R2	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	SAT.
	R3	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R5	UNSAT.	UNSAT.	UNSAT.	55	R6	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R0	SAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	SAT.
	R3	SAT.	UNSAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	SAT.	UNSAT.	UNSAT.
45	R5	UNSAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	SAT.	SAT.		R6	UNSAT.	SAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R9	SAT.	SAT.	SAT.
	R4	SAT.	UNSAT.	UNSAT.		R10	SAT.	UNSAT.	UNSAT.
46	R5	SAT.	SAT.	UNSAT.		R11	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	UNSAT.		R12	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R13	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	SAT.	UNSAT.					
	R3	SAT.	SAT.	SAT.					
	R4	UNSAT.	SAT.	UNSAT.					
47	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	SAT.					
	R0	SAT.	SAT.	SAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	SAT.	UNSAT.					
	R3	SAT.	SAT.	SAT.					
48	R4	UNSAT.	UNSAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	SAT.	UNSAT.	UNSAT.					
	R7	SAT.	SAT.	SAT.					
	R0	UNSAT.	SAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
49	R2	UNSAT.	SAT.	SAT.					
	R3	SAT.	UNSAT.	SAT.					
	R4	UNSAT.	UNSAT.	SAT.					
	R5	SAT.	SAT.	SAT.					
	R6	SAT.	SAT.	SAT.					
	R0	SAT.	UNSAT.	SAT.					
50	R1	SAT.	SAT.	UNSAT.					
	R2	SAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	SAT.	UNSAT.	SAT.					
	R6	UNSAT.	SAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					

Table 5: GPT-4o-2024-08-06 / Hindi: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	4.81	1028	67	5.37	930	57	5.34	935	64
02	5.04	1011	69	4.04	907	64	4.62	844	71
03	4.52	811	73	4.77	837	65	4.57	778	59
04	4.42	630	67	4.39	706	63	4.21	644	57
05	4.66	865	65	4.26	841	62	4.76	881	73
06	4.59	864	56	4.97	996	80	5.47	971	73
07	4.70	837	82	4.45	800	74	4.66	877	75
08	5.37	953	77	4.49	864	67	4.66	878	76
09	5.36	811	69	4.06	783	59	4.45	862	73
10	4.64	903	82	4.68	864	75	5.35	900	70
11	4.33	956	78	4.37	942	68	5.66	926	68
12	4.27	810	58	4.14	836	72	4.93	833	56
13	4.91	890	70	4.69	849	68	4.89	795	62
14	4.47	924	58	4.15	887	67	4.58	918	63
15	4.25	880	61	4.32	898	70	4.60	829	65
16	4.40	931	68	4.32	831	70	5.12	851	67
17	4.53	967	62	4.49	888	74	5.29	873	69
18	3.89	772	71	3.75	835	63	4.63	788	68
19	4.43	911	62	3.99	852	70	5.06	766	60
20	4.45	868	67	4.27	939	69	4.10	903	73
21	5.27	910	68	4.68	950	56	4.64	907	67
22	5.07	896	62	4.75	902	64	5.07	939	77
23	5.00	924	70	5.00	854	67	5.06	1033	72
24	4.73	1026	79	4.73	1010	71	4.89	910	63
25	4.47	930	72	4.71	901	70	5.72	962	66
26	4.49	989	64	4.03	933	66	4.16	830	56
27	4.32	934	60	5.68	877	71	5.16	906	75
28	4.72	808	70	4.62	870	69	4.90	853	67
29	4.79	899	77	4.30	807	58	4.89	763	62
30	4.55	792	69	4.58	762	73	5.00	794	59
31	5.01	1043	81	5.07	1059	80	5.42	938	71
32	4.64	911	66	4.52	876	64	4.97	784	66
33	4.47	897	71	5.50	933	67	4.97	843	70
34	4.97	844	65	4.66	915	73	4.97	839	79
35	4.65	822	68	4.32	836	61	4.44	818	62
36	4.35	863	58	4.41	960	75	5.25	871	73
37	4.37	831	69	4.18	825	67	4.41	773	71
38	4.40	818	64	4.40	808	80	4.65	798	71
39	4.41	819	64	4.48	934	75	5.17	793	65
40	4.22	809	59	4.37	776	74	5.43	879	71
41	4.62	852	68	4.60	865	79	5.12	866	66
42	4.11	813	63	4.50	914	65	5.40	860	73
43	4.37	854	65	4.60	844	76	5.11	874	67
44	4.69	806	75	5.35	863	79	5.36	793	70
45	4.06	718	61	4.75	765	78	4.71	850	54
46	4.39	844	69	3.92	768	61	4.93	867	79
47	4.61	901	76	3.99	820	62	5.09	775	83
48	4.97	1010	73	4.36	928	66	5.03	885	70
49	4.48	957	66	4.41	913	61	4.94	945	62
50	4.71	830	68	4.54	827	70	5.47	903	70
51	4.49	694	72	6.78	750	73	5.01	701	57
52	4.95	727	62	4.72	802	66	5.95	762	65
53	4.75	834	73	4.85	850	74	4.64	854	60
54	4.75	902	77	4.98	822	83	5.84	836	69
55	5.63	1344	79	5.74	1271	87	6.21	1287	87

1.2 GPT-5.4

1.2.1 English

Requirement-level satisfaction for English under GPT-5.4.

English tasks 01–10					English tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
02	R4	UNSAT.	UNSAT.	UNSAT.	12	R4	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
03	R3	SAT.	SAT.	SAT.	13	R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
04	R2	UNSAT.	UNSAT.	UNSAT.	14	R6	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
05	R1	UNSAT.	UNSAT.	UNSAT.	15	R3	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
07	R4	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
08	R3	UNSAT.	UNSAT.	UNSAT.	18	R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
09	R2	UNSAT.	UNSAT.	UNSAT.	19	R6	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
10	R0	UNSAT.	UNSAT.	UNSAT.	20	R3	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.

English tasks 21–30					English tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	UNSAT.	UNSAT.	UNSAT.	31	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	UNSAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
23	R5	SAT.	SAT.	SAT.	33	R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
24	R4	UNSAT.	UNSAT.	UNSAT.	34	R5	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	UNSAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
26	R2	UNSAT.	UNSAT.	UNSAT.	36	R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	UNSAT.
	R4	UNSAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
27	R1	UNSAT.	UNSAT.	UNSAT.	37	R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	SAT.		R3	SAT.	SAT.	SAT.
30	R5	UNSAT.	UNSAT.	UNSAT.	40	R4	UNSAT.	SAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.

English tasks 41–50					English tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	UNSAT.	UNSAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
42	R0	UNSAT.	UNSAT.	UNSAT.	53	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
43	R0	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R0	UNSAT.	UNSAT.	UNSAT.	55	R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
45	R0	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
46	R0	UNSAT.	UNSAT.	UNSAT.		R9	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R11	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R12	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R13	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.					
47	R0	UNSAT.	UNSAT.	UNSAT.	48	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
48	R0	UNSAT.	UNSAT.	UNSAT.	49	R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
49	R0	UNSAT.	UNSAT.	UNSAT.	50	R5	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					

Table 6: GPT-5.4 / English: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	5.59	1260	89	5.86	1267	85	5.23	1260	78
02	7.35	1152	96	7.58	1188	97	6.74	1142	95
03	7.23	1158	107	6.82	1192	108	6.20	1134	108
04	6.78	729	79	5.59	745	82	6.09	728	78
05	6.37	1079	89	5.97	1087	96	6.22	1106	89
06	6.41	1059	103	5.51	1049	96	5.98	1069	100
07	5.91	1044	88	6.08	1070	97	6.33	1050	89
08	7.98	1091	92	6.91	1059	99	6.52	1090	99
09	5.58	1038	86	6.36	1070	84	5.69	1069	84
10	5.87	1026	95	5.89	1114	91	5.70	1100	90
11	6.44	1144	97	6.04	1150	102	5.80	1155	101
12	6.33	1143	97	5.44	1126	94	5.96	1119	94
13	5.91	1026	90	6.08	1049	97	5.72	1021	87
14	6.64	1121	101	5.75	1101	98	5.77	1098	102
15	5.66	1082	82	5.77	1104	86	5.67	1092	94
16	6.13	1071	91	6.23	1092	95	5.80	1074	92
17	5.95	1180	101	5.55	1163	95	6.16	1176	98
18	5.97	1054	86	5.46	1037	90	5.82	1047	92
19	6.06	1092	96	5.81	1104	101	5.64	1092	98
20	5.86	1084	96	6.24	1092	93	6.03	1077	95
21	5.97	1075	93	6.21	1095	101	5.99	1100	99
22	5.96	1148	95	6.19	1130	94	5.78	1155	92
23	6.02	1181	100	5.48	1196	88	5.88	1189	96
24	6.31	1300	96	5.81	1288	93	6.04	1292	96
25	6.08	1124	97	5.68	1110	101	5.59	1080	95
26	6.44	1085	92	5.87	1081	91	5.39	1063	98
27	6.31	1316	99	6.59	1321	95	6.13	1282	97
28	5.81	1018	96	5.43	1023	88	5.78	1017	90
29	5.31	1006	91	6.13	1053	91	5.49	1013	90
30	6.40	1086	87	6.52	1123	92	6.58	1084	83
31	6.55	1254	100	6.06	1221	95	5.90	1222	101
32	5.85	1055	91	5.86	1056	92	5.38	1045	91
33	5.93	1099	95	5.68	1088	92	9.27	1090	92
34	5.86	1111	100	5.90	1055	98	5.70	1104	96
35	7.12	1057	92	6.05	1045	89	5.95	1060	95
36	6.35	1105	100	5.97	1118	106	5.87	1083	98
37	6.06	1133	94	5.97	1108	91	5.91	1094	93
38	5.88	1204	90	5.56	1210	90	6.00	1223	90
39	6.04	1277	101	6.09	1286	93	7.60	1430	106
40	6.28	1022	93	5.85	1053	99	5.86	1032	101
41	7.04	1021	89	5.33	996	94	5.78	976	95
42	6.67	1111	93	6.39	1094	91	6.54	1090	93
43	5.88	1230	99	6.08	1241	98	6.07	1246	99
44	6.14	1023	89	5.79	1041	91	6.09	1043	98
45	6.12	984	84	5.86	996	82	5.38	980	84
46	5.79	1121	81	5.96	1154	87	6.10	1152	85
47	6.33	1160	106	5.56	1073	95	5.92	1154	104
48	5.87	1106	94	5.72	1107	91	6.05	1094	95
49	5.97	1060	96	7.81	1065	90	6.21	1053	92
50	5.64	1068	92	5.77	1086	88	5.92	1104	89
51	7.21	866	101	7.03	841	106	6.63	835	108
52	7.08	834	98	6.27	806	103	6.49	801	95
53	6.25	1067	106	5.93	1063	103	7.02	1113	102
54	6.24	1173	104	6.07	1172	103	6.15	1180	99
55	6.51	1199	111	6.86	1239	109	6.60	1222	110

1.2.2 Arabic

Requirement-level satisfaction for Arabic under GPT-5.4.

Arabic tasks 01–10					Arabic tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
02	R0	UNSAT.	UNSAT.	UNSAT.	12	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	SAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
03	R0	UNSAT.	UNSAT.	UNSAT.	13	R5	UNSAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
04	R0	UNSAT.	UNSAT.	UNSAT.	14	R3	UNSAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R5	UNSAT.	SAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
05	R0	UNSAT.	UNSAT.	UNSAT.	15	R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R6	SAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	SAT.	SAT.
07	R0	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
08	R0	UNSAT.	UNSAT.	UNSAT.	18	R3	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
09	R5	SAT.	UNSAT.	SAT.	19	R1	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
10	R4	UNSAT.	UNSAT.	UNSAT.	20	R6	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					

Arabic tasks 21–30					Arabic tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	UNSAT.	UNSAT.	UNSAT.	31	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	SAT.
	R5	UNSAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R0	UNSAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
23	R0	UNSAT.	UNSAT.	UNSAT.	33	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	SAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
24	R0	UNSAT.	UNSAT.	UNSAT.	34	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
25	R0	UNSAT.	UNSAT.	UNSAT.	35	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	UNSAT.	UNSAT.
26	R0	UNSAT.	UNSAT.	UNSAT.	36	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
27	R0	UNSAT.	UNSAT.	UNSAT.	37	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	SAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	UNSAT.
29	R0	UNSAT.	UNSAT.	UNSAT.	39	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	SAT.	SAT.		R4	SAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
30	R0	UNSAT.	UNSAT.	UNSAT.	40	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
31	R0	UNSAT.	UNSAT.	UNSAT.	41	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.

Arabic tasks 41–50					Arabic tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	UNSAT.	UNSAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	UNSAT.		R2	UNSAT.	SAT.	UNSAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	SAT.	UNSAT.		R0	UNSAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R2	UNSAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	UNSAT.	UNSAT.
43	R5	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R5	UNSAT.	SAT.	UNSAT.	55	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	SAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	SAT.	UNSAT.
	R4	SAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	SAT.
45	R5	UNSAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	SAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R9	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	UNSAT.
46	R5	UNSAT.	UNSAT.	UNSAT.		R11	UNSAT.	SAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R12	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R13	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	SAT.					
	R4	SAT.	SAT.	SAT.					
47	R5	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	SAT.					
	R4	UNSAT.	SAT.	SAT.					
48	R5	UNSAT.	UNSAT.	SAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	UNSAT.	SAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
49	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	SAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
50	R4	UNSAT.	SAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	SAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					

Table 7: GPT-5.4 / Arabic: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	7.00	1401	99	8.46	1428	97	7.67	1415	109
02	8.30	1308	111	9.88	1317	112	10.56	1335	113
03	8.20	1319	125	8.90	1361	125	9.01	1286	121
04	7.22	890	110	8.21	904	122	8.66	917	101
05	9.22	1296	112	10.55	1341	114	8.74	1325	104
06	7.64	1212	123	8.91	1232	118	8.35	1212	131
07	8.12	1231	127	9.60	1211	114	8.46	1212	115
08	8.39	1230	106	8.93	1215	112	9.24	1233	104
09	8.27	1256	107	9.54	1251	112	8.36	1228	103
10	8.22	1206	115	9.78	1303	111	8.49	1255	112
11	7.15	1283	113	8.75	1334	106	9.63	1320	120
12	7.39	1253	119	8.92	1301	116	9.75	1287	117
13	7.40	1205	104	8.41	1188	109	9.55	1229	97
14	7.61	1292	115	8.73	1283	117	9.31	1288	113
15	7.84	1265	109	8.99	1315	103	9.27	1312	100
16	7.23	1219	102	9.13	1263	105	9.37	1266	112
17	8.14	1369	116	9.03	1331	115	9.13	1265	105
18	7.38	1206	106	8.15	1223	101	8.71	1188	99
19	7.69	1210	106	9.47	1295	119	9.85	1315	109
20	7.69	1236	106	8.66	1260	104	8.74	1263	113
21	7.65	1228	113	8.15	1209	117	8.19	1257	107
22	6.65	1166	101	8.10	1242	112	9.62	1268	124
23	7.95	1197	110	7.94	1246	113	7.98	1219	111
24	7.68	1389	105	9.02	1366	110	7.33	1417	112
25	7.99	1261	110	9.42	1283	120	8.76	1260	118
26	7.51	1199	110	8.20	1258	113	8.64	1237	115
27	8.09	1309	116	9.10	1508	119	8.88	1343	111
28	6.94	1148	112	8.12	1176	107	8.33	1167	111
29	8.38	1164	117	7.71	1179	106	7.94	1185	110
30	8.53	1210	110	9.07	1308	111	10.11	1281	109
31	8.49	1324	113	9.60	1438	115	8.75	1329	117
32	7.87	1228	114	8.09	1215	111	8.15	1219	123
33	7.63	1161	113	8.57	1201	113	8.40	1203	109
34	7.25	1176	113	8.25	1298	108	8.11	1271	123
35	8.65	1234	118	8.76	1288	113	8.26	1234	119
36	7.89	1255	123	9.42	1239	118	9.08	1286	113
37	7.86	1265	114	8.22	1278	108	8.30	1274	105
38	7.40	1396	104	8.07	1393	111	7.72	1272	106
39	8.99	1520	104	8.83	1526	107	8.32	1386	118
40	7.96	1187	115	7.73	1161	114	7.98	1192	120
41	7.67	1241	111	8.45	1278	116	8.60	1226	118
42	9.45	1310	120	8.92	1240	107	9.34	1290	109
43	8.36	1328	114	8.45	1294	121	8.18	1286	110
44	8.00	1112	108	8.66	1241	108	8.89	1131	114
45	8.06	1115	106	8.04	1189	108	8.13	1152	108
46	7.25	1284	109	7.29	1274	101	8.05	1282	98
47	7.88	1236	115	9.11	1269	113	9.35	1380	125
48	7.64	1303	104	9.23	1329	110	9.02	1322	110
49	8.66	1216	111	8.74	1203	100	9.06	1257	120
50	9.11	1288	107	8.07	1260	103	9.23	1299	111
51	8.57	1004	129	8.58	989	117	9.94	1077	121
52	9.71	992	118	9.01	993	132	9.44	1031	119
53	9.52	1221	134	9.73	1288	104	9.10	1218	110
54	9.00	1348	124	9.31	1374	126	9.50	1403	121
55	9.94	1574	135	9.42	1494	127	10.55	1531	132

1.2.3 Turkish

Requirement-level satisfaction for Turkish under GPT-5.4.

Turkish tasks 01–10					Turkish tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
02	R0	UNSAT.	UNSAT.	SAT.	12	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	UNSAT.	UNSAT.		R2	SAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
03	R0	UNSAT.	UNSAT.	UNSAT.	13	R5	SAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	SAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	SAT.	UNSAT.
04	R0	UNSAT.	SAT.	SAT.	14	R3	UNSAT.	UNSAT.	SAT.
	R1	SAT.	SAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
	R2	UNSAT.	SAT.	SAT.		R5	SAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	SAT.
	R4	UNSAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
05	R0	UNSAT.	UNSAT.	UNSAT.	15	R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	UNSAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R6	SAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.
07	R0	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	SAT.		R7	SAT.	SAT.	UNSAT.
	R4	SAT.	UNSAT.	UNSAT.		R8	SAT.	UNSAT.	UNSAT.
08	R0	UNSAT.	UNSAT.	SAT.	18	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
09	R0	UNSAT.	UNSAT.	SAT.	19	R5	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.		R6	UNSAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.	20	R0	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
10	R0	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	SAT.					

Turkish tasks 21–30					Turkish tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	UNSAT.	SAT.	UNSAT.	31	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R0	UNSAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
23	R0	UNSAT.	UNSAT.	UNSAT.	33	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	SAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
24	R0	UNSAT.	UNSAT.	UNSAT.	34	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	UNSAT.		R5	SAT.	SAT.	UNSAT.
25	R0	UNSAT.	UNSAT.	UNSAT.	35	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	SAT.		R5	UNSAT.	SAT.	UNSAT.
26	R0	UNSAT.	UNSAT.	UNSAT.	36	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
27	R0	UNSAT.	UNSAT.	UNSAT.	37	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	SAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	SAT.
29	R0	UNSAT.	UNSAT.	UNSAT.	39	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
30	R0	UNSAT.	UNSAT.	UNSAT.	40	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
31	R0	UNSAT.	UNSAT.	UNSAT.	41	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.

Turkish tasks 41–50					Turkish tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	UNSAT.	SAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	UNSAT.
	R2	SAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
42	R0	UNSAT.	UNSAT.	UNSAT.	53	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	UNSAT.	UNSAT.
43	R0	SAT.	SAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
	R5	SAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R0	UNSAT.	UNSAT.	UNSAT.	55	R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	UNSAT.		R2	UNSAT.	UNSAT.	SAT.
	R4	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
	R5	SAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
45	R0	UNSAT.	SAT.	UNSAT.	55	R5	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	SAT.		R8	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R9	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	UNSAT.
46	R0	UNSAT.	UNSAT.	UNSAT.	55	R11	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R12	UNSAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R13	UNSAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	SAT.	UNSAT.					
	R5	SAT.	SAT.	SAT.					
47	R0	UNSAT.	UNSAT.	UNSAT.	55				
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	UNSAT.	SAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	SAT.	UNSAT.	UNSAT.					
48	R0	UNSAT.	UNSAT.	UNSAT.	55				
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	SAT.	UNSAT.	SAT.					
49	R0	UNSAT.	UNSAT.	UNSAT.	55				
	R1	SAT.	SAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
50	R0	UNSAT.	UNSAT.	UNSAT.	55				
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	SAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
50	R0	UNSAT.	UNSAT.	UNSAT.	55				
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	SAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
50	R0	UNSAT.	UNSAT.	UNSAT.	55				
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	SAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
50	R0	UNSAT.	UNSAT.	UNSAT.	55				
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	SAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					

Table 8: GPT-5.4 / Turkish: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	8.04	1400	95	6.93	1357	105	7.06	1338	94
02	9.27	1240	115	8.80	1242	119	7.41	1210	110
03	9.38	1238	115	7.76	1233	120	7.21	1218	123
04	8.40	862	99	7.66	845	104	6.37	805	110
05	9.02	1224	109	7.98	1220	126	7.61	1185	111
06	9.24	1226	136	7.50	1161	113	7.32	1151	117
07	9.21	1140	117	7.87	1163	107	6.88	1081	113
08	9.38	1184	118	7.95	1155	114	8.57	1155	105
09	8.76	1170	103	7.96	1155	109	6.98	1102	104
10	8.76	1184	120	8.07	1149	117	6.71	1135	118
11	10.33	1268	120	7.46	1210	110	7.50	1236	116
12	8.84	1218	120	7.09	1199	123	7.54	1220	121
13	9.60	1130	107	7.49	1120	113	7.11	1116	99
14	9.17	1245	119	7.56	1206	119	7.29	1210	116
15	9.18	1221	104	8.72	1176	117	8.46	1191	103
16	9.60	1220	109	7.71	1184	105	7.24	1144	114
17	9.94	1301	116	7.41	1246	110	7.36	1246	115
18	8.93	1165	110	7.58	1102	116	8.37	1132	104
19	8.95	1208	112	9.73	1179	107	9.40	1201	123
20	8.40	1197	102	7.83	1163	102	8.41	1176	118
21	9.21	1217	117	7.75	1177	110	8.39	1164	111
22	9.38	1242	113	7.18	1188	118	8.80	1212	124
23	8.82	1309	114	7.27	1240	119	8.42	1264	118
24	8.84	1396	111	8.84	1369	121	8.92	1364	123
25	8.88	1214	114	8.34	1178	116	8.41	1180	116
26	9.63	1245	117	8.18	1149	118	9.39	1169	116
27	9.64	1299	114	9.44	1431	123	9.02	1383	119
28	10.11	1169	112	7.89	1098	104	8.05	1099	111
29	9.79	1178	112	7.99	1131	114	8.33	1122	118
30	9.06	1223	114	8.56	1176	110	8.43	1176	113
31	7.72	1305	113	7.94	1263	116	7.57	1289	123
32	9.13	1207	116	7.85	1152	112	7.35	1158	116
33	9.04	1283	115	8.59	1198	122	7.90	1191	118
34	8.52	1251	113	8.62	1217	111	7.73	1181	110
35	8.82	1168	114	7.30	1105	101	8.08	1131	100
36	8.20	1212	117	8.19	1167	112	8.25	1216	114
37	7.92	1220	103	8.52	1176	107	6.91	1162	108
38	7.58	1347	102	8.05	1304	107	7.12	1306	101
39	8.08	1414	107	8.05	1370	115	8.68	1545	120
40	8.01	1167	127	9.07	1139	117	7.08	1144	115
41	8.34	1257	110	8.44	1197	103	7.72	1212	116
42	8.58	1228	113	9.70	1168	114	7.26	1140	109
43	8.11	1368	123	9.21	1312	117	7.10	1324	118
44	7.94	1136	109	9.46	1116	115	7.41	1127	117
45	8.30	1082	103	8.27	1080	107	6.59	1050	107
46	7.74	1267	110	8.25	1239	109	7.09	1202	111
47	8.55	1293	128	8.07	1213	122	8.17	1265	114
48	8.27	1235	106	8.62	1214	108	7.76	1203	108
49	8.23	1166	118	7.56	1118	112	7.43	1118	112
50	8.26	1217	106	7.42	1180	104	7.96	1163	105
51	10.42	995	138	8.41	949	124	7.51	901	118
52	9.26	958	134	7.39	871	126	8.04	910	131
53	9.31	1193	112	8.05	1131	119	7.68	1146	125
54	9.49	1339	124	9.16	1322	140	8.43	1249	126
55	10.46	1395	139	8.35	1361	130	9.45	1360	133

1.2.4 Chinese

Requirement-level satisfaction for Chinese under GPT-5.4.

Chinese tasks 01–10					Chinese tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
02	R0	UNSAT.	UNSAT.	UNSAT.	12	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
03	R0	UNSAT.	UNSAT.	UNSAT.	13	R5	UNSAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
04	R0	UNSAT.	UNSAT.	UNSAT.	14	R3	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
05	R0	UNSAT.	UNSAT.	UNSAT.	15	R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
07	R0	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
08	R0	UNSAT.	UNSAT.	UNSAT.	18	R3	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
09	R5	UNSAT.	SAT.	SAT.	19	R1	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
10	R4	UNSAT.	UNSAT.	UNSAT.	20	R6	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.

Chinese tasks 21–30					Chinese tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	UNSAT.	UNSAT.	UNSAT.	31	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	UNSAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
23	R5	SAT.	SAT.	SAT.	33	R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
24	R4	UNSAT.	UNSAT.	UNSAT.	34	R5	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
26	R2	UNSAT.	UNSAT.	UNSAT.	36	R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
27	R1	UNSAT.	UNSAT.	UNSAT.	37	R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	SAT.	SAT.		R4	UNSAT.	SAT.	UNSAT.
30	R5	UNSAT.	UNSAT.	UNSAT.	40	R5	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
31	R4	UNSAT.	UNSAT.	UNSAT.	40	R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					

Chinese tasks 41–50					Chinese tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	UNSAT.	UNSAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
43	R5	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R5	UNSAT.	UNSAT.	UNSAT.	55	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	SAT.		R3	UNSAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
45	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R9	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	UNSAT.
46	R5	UNSAT.	UNSAT.	UNSAT.		R11	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R12	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R13	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
47	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
48	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
49	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					
50	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					

Table 9: GPT-5.4 / Chinese: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	7.45	1160	93	6.94	1187	93	10.82	1183	91
02	7.09	1055	98	9.90	1049	104	7.22	1034	96
03	8.25	1038	104	7.90	1044	95	7.01	1028	96
04	6.01	643	84	7.41	651	84	6.44	646	84
05	7.03	988	95	6.95	1019	90	6.49	987	91
06	6.37	982	100	6.88	970	98	6.39	929	96
07	6.43	969	92	7.47	973	97	6.57	958	85
08	6.81	986	91	7.48	1001	96	7.61	978	99
09	5.93	896	85	6.77	890	83	6.23	889	84
10	6.37	996	94	8.01	1007	98	7.27	1004	96
11	6.52	1052	98	6.75	1020	96	6.57	1018	102
12	6.55	1016	94	7.28	1018	107	6.59	1025	91
13	7.86	953	90	7.51	960	88	7.87	939	90
14	6.78	1032	93	6.54	1010	99	6.66	1021	98
15	7.23	1019	88	7.37	1008	87	7.55	1004	85
16	6.05	988	86	6.37	988	85	6.98	999	89
17	6.35	1072	99	6.81	1095	93	6.55	1070	94
18	7.01	955	93	7.55	952	97	6.69	932	92
19	7.09	989	100	7.10	1002	98	6.70	987	92
20	6.79	1002	98	6.67	996	87	7.38	996	89
21	6.56	993	92	7.12	988	97	6.69	974	93
22	6.34	1040	95	6.79	1032	89	6.53	1010	95
23	6.23	1080	95	7.06	1084	99	6.78	1070	99
24	6.21	1258	79	7.19	1190	91	6.95	1153	91
25	7.43	984	94	7.84	1000	91	7.02	1000	90
26	7.40	979	96	6.57	976	94	5.65	970	94
27	7.69	1197	93	7.74	1210	103	6.62	1201	95
28	6.17	912	82	6.77	924	85	6.05	917	89
29	6.45	933	92	6.40	949	93	6.26	926	98
30	7.65	995	103	8.15	994	103	7.23	1026	101
31	6.39	1092	91	7.76	1117	98	6.61	1098	102
32	6.87	960	91	6.61	971	85	7.25	965	94
33	7.35	994	98	6.47	993	95	6.47	1002	96
34	6.42	1000	93	7.98	1008	95	6.81	996	100
35	6.09	966	97	6.58	960	100	6.35	962	95
36	7.00	1013	101	7.05	1012	94	5.90	988	95
37	6.60	994	93	7.40	1012	94	5.86	990	85
38	6.34	1110	86	6.63	1120	93	6.18	1124	87
39	7.83	1177	110	7.25	1202	113	5.78	1183	112
40	7.26	948	92	6.61	924	92	5.94	945	88
41	6.44	985	93	6.38	977	87	6.10	990	85
42	7.32	1022	89	7.81	1023	93	6.16	1090	81
43	6.41	1137	101	7.24	1158	97	6.59	1140	110
44	7.18	934	90	7.84	951	98	6.51	945	97
45	6.34	892	95	7.18	906	86	6.82	898	88
46	6.17	1047	90	6.25	1044	86	5.56	1040	90
47	7.02	1042	107	6.90	1031	104	6.53	1058	107
48	6.40	1022	85	6.42	1008	88	5.88	986	84
49	6.99	951	95	6.93	963	100	6.64	945	91
50	6.45	983	87	7.16	1000	85	5.89	972	83
51	7.64	780	97	7.40	772	114	6.67	777	100
52	6.71	698	109	7.58	778	94	6.18	709	105
53	7.98	990	106	7.30	979	110	6.26	968	108
54	7.02	1055	100	7.18	1081	101	7.23	1063	110
55	7.34	1121	116	7.81	1135	113	6.26	1119	107

1.2.5 Hindi

Requirement-level satisfaction for Hindi under GPT-5.4.

Hindi tasks 01–10					Hindi tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	SAT.	UNSAT.
	R4	SAT.	SAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
					R5	SAT.	SAT.	UNSAT.	
02	R0	UNSAT.	UNSAT.	UNSAT.	12	R0	SAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	SAT.	UNSAT.
	R2	UNSAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
					R5	SAT.	UNSAT.	SAT.	
03	R0	UNSAT.	UNSAT.	UNSAT.	13	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	SAT.		R1	UNSAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
					R5	SAT.	UNSAT.	UNSAT.	
04	R0	SAT.	UNSAT.	SAT.	14	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	SAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
					R5	UNSAT.	UNSAT.	UNSAT.	
05	R0	UNSAT.	UNSAT.	UNSAT.	15	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
					R5	UNSAT.	UNSAT.	UNSAT.	
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R0	SAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
					R5	UNSAT.	SAT.	UNSAT.	
07	R0	UNSAT.	UNSAT.	SAT.	17	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	SAT.		R4	SAT.	SAT.	UNSAT.
					R5	UNSAT.	UNSAT.	UNSAT.	
08	R0	UNSAT.	UNSAT.	UNSAT.	18	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	UNSAT.	UNSAT.
					R5	UNSAT.	UNSAT.	UNSAT.	
09	R0	UNSAT.	SAT.	UNSAT.	19	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
					R5	UNSAT.	UNSAT.	UNSAT.	
10	R0	UNSAT.	UNSAT.	UNSAT.	20	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	UNSAT.	UNSAT.
					R5	UNSAT.	SAT.	UNSAT.	
					R6	UNSAT.	UNSAT.	UNSAT.	

Hindi tasks 21–30					Hindi tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	SAT.	UNSAT.	SAT.	31	R0	UNSAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	UNSAT.	UNSAT.
	R5	SAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	UNSAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
23	R5	SAT.	SAT.	SAT.	33	R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
24	R4	UNSAT.	UNSAT.	UNSAT.	34	R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	SAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	UNSAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
26	R2	UNSAT.	UNSAT.	UNSAT.	36	R3	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
27	R1	UNSAT.	UNSAT.	UNSAT.	37	R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	UNSAT.		R6	SAT.	UNSAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	SAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	SAT.		R4	UNSAT.	SAT.	SAT.
	R5	UNSAT.	SAT.	UNSAT.		R5	SAT.	SAT.	SAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
30	R5	UNSAT.	UNSAT.	UNSAT.	40	R4	SAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
31	R4	SAT.	SAT.	UNSAT.	41	R3	UNSAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R4	UNSAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					

Hindi tasks 41–50					Hindi tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	UNSAT.	UNSAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
42	R6	UNSAT.	SAT.	UNSAT.	53	R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	UNSAT.		R2	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
43	R5	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R5	SAT.	UNSAT.	SAT.	55	R6	UNSAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	UNSAT.	SAT.
45	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	SAT.		R9	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	UNSAT.
46	R5	UNSAT.	UNSAT.	UNSAT.		R11	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R12	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R13	SAT.	UNSAT.	SAT.
	R2	UNSAT.	SAT.	UNSAT.					
	R3	SAT.	SAT.	SAT.					
	R4	SAT.	SAT.	SAT.					
47	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	SAT.	SAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	SAT.	SAT.					
48	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	SAT.	UNSAT.	SAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
49	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	SAT.	SAT.	SAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
50	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					

Table 10: GPT-5.4 / Hindi: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	7.63	1273	66	7.86	1279	71	7.58	1230	75
02	8.58	1172	74	8.86	1187	77	8.63	1098	73
03	8.25	1135	90	9.89	1247	94	8.36	1099	86
04	7.70	768	69	7.84	757	64	6.78	693	65
05	9.06	1182	71	8.60	1125	76	7.83	1072	69
06	7.77	1127	76	9.34	1168	84	8.17	1063	71
07	7.69	1072	79	8.46	1076	75	8.95	1044	70
08	8.55	1112	75	10.09	1125	81	8.54	1073	71
09	8.16	1108	72	8.96	1115	76	7.59	998	74
10	7.93	1100	78	8.92	1102	82	8.90	1115	87
11	7.85	1165	73	8.19	1141	70	7.86	1107	70
12	8.46	1292	72	8.68	1187	77	8.78	1091	80
13	7.90	1071	67	9.50	1118	80	7.77	1008	78
14	7.84	1154	70	9.60	1131	75	7.68	1097	74
15	7.90	1142	75	9.33	1184	66	8.91	1068	73
16	7.43	1107	69	8.03	1102	64	7.93	1081	70
17	7.88	1211	80	7.92	1191	77	8.74	1160	82
18	7.53	1065	73	8.11	1037	70	7.78	1023	68
19	7.83	1114	71	8.73	1129	73	8.42	1096	87
20	7.99	1103	74	8.34	1110	74	6.95	1030	70
21	7.80	1118	75	8.51	1137	85	6.00	1061	72
22	7.93	1167	76	8.76	1189	71	6.46	1115	67
23	7.74	1230	77	8.14	1265	79	6.68	1205	67
24	8.15	1328	71	8.08	1312	71	6.77	1280	81
25	7.49	1125	79	8.96	1197	77	6.64	1087	71
26	8.01	1102	80	9.61	1160	76	6.72	1066	75
27	7.52	1293	83	8.36	1344	71	7.57	1122	77
28	6.85	1025	68	8.45	1049	73	6.28	969	70
29	7.64	1050	77	8.52	1068	79	7.12	1061	70
30	8.04	1141	74	9.13	1158	73	7.21	1128	76
31	7.82	1240	76	9.36	1290	75	6.32	1193	71
32	7.31	1048	71	8.51	1090	74	7.11	1070	75
33	8.04	1074	73	7.63	1095	72	7.50	1051	67
34	7.36	1133	78	7.79	1136	78	8.23	1158	73
35	7.88	1076	69	9.01	1144	71	7.41	1065	75
36	8.04	1109	78	7.29	1101	68	8.20	1160	73
37	7.42	1116	75	8.14	1137	77	7.95	1153	78
38	7.78	1235	74	8.54	1266	73	10.20	1252	68
39	7.70	1291	74	8.27	1300	81	7.86	1317	72
40	8.40	1074	77	8.85	1099	80	7.62	1070	70
41	7.84	1112	74	8.43	1139	74	8.95	1167	74
42	8.58	1122	70	9.39	1196	66	9.25	1135	72
43	8.20	1274	76	9.26	1430	72	7.53	1218	68
44	8.88	1089	77	9.46	1082	76	8.30	1088	70
45	8.39	1018	79	8.21	1002	69	8.05	1016	80
46	7.06	1139	73	10.57	1190	75	7.42	1148	69
47	8.57	1154	77	8.45	1163	77	8.36	1092	83
48	7.81	1116	77	9.09	1174	78	7.48	1117	70
49	7.75	1068	75	8.02	1015	64	8.38	1136	73
50	7.56	1091	74	8.43	1075	73	7.45	1116	74
51	8.71	898	74	9.13	861	87	8.17	910	79
52	11.96	1379	75	10.31	900	67	7.78	942	78
53	7.49	1060	73	8.40	1067	80	7.16	1069	78
54	8.44	1203	72	8.48	1156	71	8.42	1242	70
55	9.37	1284	81	8.48	1221	78	8.41	1286	82

1.3 Claude Sonnet 4.6

1.3.1 English

Requirement-level satisfaction for English under Claude Sonnet 4.6.

English tasks 01–10					English tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
02	R4	UNSAT.	UNSAT.	UNSAT.	12	R4	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
03	R3	SAT.	SAT.	SAT.	13	R2	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
04	R2	UNSAT.	UNSAT.	UNSAT.	14	R6	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
05	R1	UNSAT.	UNSAT.	UNSAT.	15	R3	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
07	R4	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
08	R3	UNSAT.	UNSAT.	UNSAT.	18	R2	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
09	R1	UNSAT.	UNSAT.	UNSAT.	19	R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
10	R5	UNSAT.	UNSAT.	UNSAT.	20	R3	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.

English tasks 21–30					English tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	UNSAT.	UNSAT.	UNSAT.	31	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	UNSAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
23	R5	SAT.	SAT.	SAT.	33	R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
24	R4	UNSAT.	UNSAT.	UNSAT.	34	R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
26	R2	UNSAT.	UNSAT.	UNSAT.	36	R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
27	R1	UNSAT.	UNSAT.	UNSAT.	37	R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
30	R5	UNSAT.	UNSAT.	UNSAT.	40	R4	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
31	R4	SAT.	SAT.	SAT.	41	R3	SAT.	SAT.	SAT.
	R5	SAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					

English tasks 41–50					English tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	UNSAT.	UNSAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
42	R0	UNSAT.	UNSAT.	UNSAT.	53	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
43	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.	55	R0	UNSAT.	UNSAT.	UNSAT.
44	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
45	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R9	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R11	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R12	UNSAT.	UNSAT.	UNSAT.
46	R0	UNSAT.	UNSAT.	UNSAT.		R13	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.	47	R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
47	R0	UNSAT.	UNSAT.	UNSAT.	48	R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.	49	R5	SAT.	SAT.	SAT.
48	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.	50	R5	UNSAT.	UNSAT.	UNSAT.
49	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
50	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					

Table 11: Claude Sonnet 4.6 / English: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot				OpenHands			
	Time	Input	Output Tokens	Time	Input	Output Tokens		Time	Input	Output Tokens	
01	14.28	1708	156	13.06	1680	176		12.53	1675	170	
02	15.67	1413	196	14.11	1405	202		12.24	1420	132	
03	15.20	1469	179	14.02	1485	152		13.49	1462	160	
04	12.88	989	167	12.71	998	169		13.59	993	180	
05	13.94	1456	137	12.84	1448	149		12.94	1447	130	
06	12.42	1353	171	12.22	1341	164		13.14	1383	181	
07	14.60	1399	147	14.77	1396	164		13.17	1370	154	
08	12.89	1360	131	12.46	1357	133		12.37	1359	134	
09	12.66	1386	134	12.11	1391	134		12.37	1370	129	
10	13.65	1412	156	12.97	1394	158		13.32	1400	160	
11	12.28	1456	152	13.89	1462	143		11.97	1447	147	
12	13.66	1479	161	13.23	1475	164		13.26	1462	159	
13	13.81	1350	143	13.73	1362	143		13.82	1336	144	
14	12.17	1454	148	14.09	1456	165		12.45	1446	160	
15	12.31	1452	137	12.79	1455	143		12.25	1425	137	
16	12.60	1395	132	12.36	1378	140		11.95	1387	138	
17	13.29	1530	143	13.88	1521	176		12.31	1516	140	
18	12.39	1334	142	12.92	1344	138		12.50	1332	142	
19	16.28	1458	128	13.14	1413	145		13.09	1436	133	
20	11.95	1389	137	11.99	1391	136		11.87	1378	129	
21	13.73	1429	151	13.08	1422	136		12.75	1422	150	
22	13.79	1479	141	12.72	1470	151		12.78	1459	135	
23	13.12	1551	146	12.74	1546	146		12.25	1538	144	
24	13.06	1658	136	12.20	1665	148		11.13	1638	141	
25	13.19	1421	148	12.57	1420	160		13.71	1434	164	
26	14.34	1412	145	12.49	1407	140		12.34	1380	150	
27	12.81	1667	164	13.00	1677	154		13.42	1689	159	
28	13.06	1312	147	12.35	1304	148		12.44	1308	140	
29	12.16	1313	141	14.05	1347	150		12.22	1337	156	
30	12.85	1411	130	13.14	1431	153		12.49	1420	148	
31	13.20	1561	139	13.50	1549	186		12.83	1532	151	
32	14.81	1393	197	13.93	1389	199		13.69	1383	180	
33	12.93	1431	178	12.40	1437	171		12.17	1408	164	
34	15.16	1439	204	14.07	1416	192		15.99	1427	184	
35	13.09	1414	144	13.36	1432	143		13.20	1423	141	
36	13.24	1394	165	14.78	1413	160		12.26	1399	143	
37	12.88	1450	193	13.53	1458	142		12.22	1455	137	
38	12.64	1623	143	12.42	1585	133		11.85	1592	146	
39	12.67	1681	166	13.48	1655	166		14.13	1784	149	
40	14.05	1393	196	13.75	1359	199		12.80	1383	154	
41	13.66	1441	152	12.96	1466	153		12.54	1443	148	
42	12.65	1394	148	12.41	1370	147		12.41	1380	161	
43	11.98	1599	142	13.28	1616	171		12.98	1615	142	
44	12.55	1358	153	13.52	1371	150		13.13	1342	184	
45	12.34	1312	142	14.91	1320	144		12.41	1304	142	
46	13.10	1527	134	12.23	1519	130		12.87	1504	138	
47	12.91	1504	137	13.03	1515	146		15.00	1496	135	
48	13.23	1451	119	12.69	1467	132		14.33	1462	125	
49	15.11	1372	138	12.99	1376	140		13.42	1376	135	
50	12.91	1427	132	12.72	1406	136		11.58	1402	127	
51	16.54	1058	351	16.96	1062	351		19.52	1085	328	
52	14.51	1060	218	14.48	1032	206		13.55	1012	215	
53	13.24	1321	162	13.14	1330	165		12.70	1328	153	
54	13.56	1470	194	12.97	1461	184		12.78	1428	214	
55	14.91	1529	288	14.75	1540	293		15.17	1525	262	

1.3.2 Arabic

Requirement-level satisfaction for Arabic under Claude Sonnet 4.6.

Arabic tasks 01–10					Arabic tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
02	R4	UNSAT.	UNSAT.	UNSAT.	12	R4	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
03	R3	SAT.	SAT.	SAT.	13	R2	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	UNSAT.		R5	SAT.	SAT.	SAT.
04	R2	UNSAT.	UNSAT.	UNSAT.	14	R6	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
05	R1	UNSAT.	UNSAT.	UNSAT.	15	R3	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	SAT.	UNSAT.
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
07	R4	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
08	R3	UNSAT.	UNSAT.	UNSAT.	18	R8	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
09	R1	UNSAT.	UNSAT.	UNSAT.	19	R3	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
10	R5	UNSAT.	UNSAT.	UNSAT.	20	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.

Arabic tasks 21–30					Arabic tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	UNSAT.	UNSAT.	UNSAT.	31	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	UNSAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
23	R5	SAT.	SAT.	SAT.	33	R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
24	R4	UNSAT.	UNSAT.	UNSAT.	34	R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
26	R2	UNSAT.	UNSAT.	UNSAT.	36	R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
27	R1	UNSAT.	UNSAT.	UNSAT.	37	R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
30	R5	UNSAT.	UNSAT.	UNSAT.	40	R4	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
31	R4	SAT.	SAT.	SAT.	41	R3	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.

Arabic tasks 41–50					Arabic tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	UNSAT.	UNSAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
43	R5	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R5	SAT.	SAT.	SAT.	55	R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
45	R4	UNSAT.	SAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R9	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	UNSAT.
46	R4	UNSAT.	UNSAT.	UNSAT.		R11	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R12	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R13	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.	47	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
47	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.	48	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
48	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R7	UNSAT.	UNSAT.	UNSAT.	49	R6	UNSAT.	UNSAT.	UNSAT.
49	R0	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.	50	R5	UNSAT.	UNSAT.	UNSAT.
50	R6	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R7	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.

Table 12: Claude Sonnet 4.6 / Arabic: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot				OpenHands			
	Time	Input	Output Tokens	Time	Input	Output Tokens		Time	Input	Output Tokens	
01	12.29	2003	214	11.20	2019	208		14.49	2024		235
02	16.39	1844	214	13.40	1848	198		17.30	1854		217
03	13.60	1836	207	13.39	1884	213		15.95	1870		224
04	12.83	1346	201	12.46	1369	193		14.78	1347		190
05	12.84	1788	190	14.01	1809	209		14.80	1799		193
06	15.14	1714	202	12.90	1765	216		14.80	1760		228
07	15.00	1783	226	14.57	1794	246		16.42	1772		227
08	14.95	1701	210	13.76	1784	214		17.62	1768		234
09	13.08	1714	226	12.32	1709	208		14.55	1705		208
10	14.09	1810	209	14.36	1844	223		15.85	1819		230
11	13.39	1856	212	13.13	1840	212		13.91	1837		206
12	13.53	1879	230	13.60	1878	204		15.10	1856		213
13	13.93	1696	232	14.77	1662	217		17.36	1672		216
14	12.91	1879	209	13.67	1888	218		14.84	1871		222
15	13.37	1829	220	12.66	1795	212		14.81	1815		219
16	13.53	1839	206	12.85	1809	206		15.28	1787		196
17	12.62	1915	209	13.16	1899	220		15.29	1863		205
18	13.71	1778	230	12.87	1784	228		13.94	1748		209
19	13.75	1802	218	12.81	1798	222		14.49	1787		210
20	12.92	1848	191	13.44	1856	201		15.18	1809		193
21	13.02	1819	219	12.06	1791	196		15.89	1775		203
22	13.30	1838	215	13.09	1862	221		15.19	1856		211
23	12.79	1951	200	12.41	1882	209		15.87	1932		202
24	14.58	2110	210	13.80	2081	223		16.83	2115		219
25	12.20	1764	211	12.11	1766	203		15.24	1761		216
26	13.42	1757	230	13.94	1768	243		15.95	1759		231
27	14.14	2042	235	12.90	2051	231		15.51	2027		221
28	12.33	1708	214	12.57	1728	212		14.86	1734		208
29	12.35	1712	230	14.11	1735	233		15.53	1723		214
30	12.69	1794	218	13.36	1822	231		14.54	1797		233
31	14.03	1992	206	14.21	2010	211		15.85	1947		212
32	12.91	1762	223	13.51	1759	218		15.25	1755		219
33	12.26	1813	229	13.56	1821	216		15.65	1823		236
34	14.82	1825	240	13.45	1802	223		15.62	1790		211
35	13.75	1797	218	14.46	1841	209		15.27	1780		219
36	12.50	1744	229	14.83	1791	227		14.38	1756		213
37	12.97	1830	219	15.33	1836	220		14.91	1840		218
38	13.93	2008	219	15.59	1999	245		14.96	2011		217
39	14.19	2099	226	16.37	2120	218		15.71	2113		229
40	14.05	1793	216	16.39	1790	225		15.61	1744		206
41	14.29	1788	224	15.39	1796	212		15.33	1769		217
42	12.99	1746	236	14.16	1761	208		15.67	1766		236
43	12.71	1955	237	15.86	1978	229		15.50	1956		234
44	13.36	1722	228	14.84	1731	214		15.57	1721		226
45	13.01	1706	235	14.56	1683	231		15.31	1693		238
46	13.37	1866	233	14.87	1866	234		14.78	1850		219
47	14.42	1870	226	15.63	1847	232		15.47	1857		235
48	13.68	1837	217	15.21	1854	216		14.54	1836		204
49	12.50	1757	210	16.42	1783	227		15.64	1753		210
50	13.83	1827	209	14.80	1838	200		14.93	1816		198
51	12.58	1390	211	13.31	1390	207		13.67	1387		214
52	14.06	1432	238	15.08	1456	262		14.79	1461		268
53	13.42	1762	260	14.15	1755	265		16.06	1763		243
54	13.03	1885	231	15.07	1893	226		14.76	1884		232
55	13.08	1930	234	14.58	1910	238		15.59	1920		248

1.3.3 Turkish

Requirement-level satisfaction for Turkish under Claude Sonnet 4.6.

Turkish tasks 01–10					Turkish tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
02	R0	UNSAT.	UNSAT.	UNSAT.	12	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
03	R0	UNSAT.	UNSAT.	UNSAT.	13	R5	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
04	R0	UNSAT.	UNSAT.	UNSAT.	14	R3	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
05	R0	UNSAT.	UNSAT.	UNSAT.	15	R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R6	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
07	R0	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R8	UNSAT.	UNSAT.	UNSAT.
08	R0	UNSAT.	UNSAT.	UNSAT.	18	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
09	R0	UNSAT.	UNSAT.	UNSAT.	19	R5	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
10	R0	UNSAT.	UNSAT.	UNSAT.	20	R3	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.

Turkish tasks 21–30					Turkish tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	UNSAT.	UNSAT.	UNSAT.	31	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	UNSAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	UNSAT.	UNSAT.
23	R5	SAT.	SAT.	SAT.	33	R6	UNSAT.	UNSAT.	UNSAT.
	R6	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
24	R4	UNSAT.	UNSAT.	UNSAT.	34	R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R5	UNSAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
26	R2	UNSAT.	UNSAT.	UNSAT.	36	R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
27	R1	UNSAT.	UNSAT.	UNSAT.	37	R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	UNSAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
30	R5	UNSAT.	UNSAT.	UNSAT.	40	R4	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
31	R4	UNSAT.	SAT.	SAT.	41	R3	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.

Turkish tasks 41–50					Turkish tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	UNSAT.	UNSAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
42	R0	UNSAT.	UNSAT.	UNSAT.	53	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
43	R0	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	UNSAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R0	UNSAT.	UNSAT.	UNSAT.	55	R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
45	R0	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
46	R0	UNSAT.	UNSAT.	UNSAT.		R9	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R11	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R12	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R13	UNSAT.	UNSAT.	SAT.
	R5	SAT.	SAT.	SAT.					
47	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	SAT.	SAT.	SAT.					
48	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	SAT.	SAT.	SAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
49	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
50	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
50	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
50	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
50	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					

Table 13: Claude Sonnet 4.6 / Turkish: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	22.37	1746	230	15.17	1760	232	14.66	1705	216
02	18.08	1522	190	15.50	1513	192	15.99	1519	190
03	17.75	1674	252	17.72	1633	257	16.32	1562	225
04	13.86	1046	173	14.63	1098	172	14.36	1042	169
05	15.20	1521	212	14.38	1499	202	15.56	1518	200
06	14.58	1400	240	13.77	1354	226	15.58	1439	248
07	16.18	1542	256	17.29	1585	233	15.28	1475	220
08	14.99	1468	201	14.98	1446	203	17.65	1438	204
09	14.53	1435	233	14.19	1433	217	13.99	1434	203
10	16.07	1508	220	16.66	1532	216	17.95	1518	222
11	15.03	1580	221	13.80	1552	218	16.61	1559	222
12	16.14	1607	230	15.13	1570	226	16.78	1566	224
13	15.18	1436	214	16.23	1421	233	17.09	1417	223
14	15.49	1583	225	14.59	1561	236	16.99	1559	245
15	14.28	1511	207	14.88	1554	219	16.33	1523	170
16	15.77	1557	223	14.69	1539	213	15.56	1527	191
17	15.26	1624	197	16.02	1656	211	17.36	1627	204
18	15.09	1438	244	15.56	1438	195	18.35	1454	217
19	14.42	1537	241	15.96	1533	233	16.87	1548	186
20	13.05	1435	183	13.90	1442	193	14.20	1482	197
21	14.81	1468	220	16.26	1529	230	14.42	1450	195
22	14.03	1539	197	15.64	1558	214	14.51	1524	195
23	13.71	1629	219	14.33	1603	200	15.43	1634	217
24	15.40	1744	231	14.04	1745	214	15.38	1726	207
25	14.74	1506	251	14.89	1483	244	15.77	1513	232
26	15.29	1552	252	16.01	1573	261	16.02	1555	242
27	15.05	1733	229	15.29	1764	232	14.83	1718	211
28	14.78	1461	243	14.21	1448	217	14.86	1435	215
29	15.44	1480	244	14.89	1459	230	14.80	1501	210
30	15.04	1496	216	14.84	1486	198	14.96	1498	188
31	16.24	1574	213	14.68	1573	224	15.01	1620	202
32	17.39	1523	222	15.06	1495	219	16.15	1492	230
33	15.70	1519	243	14.70	1553	251	15.16	1523	259
34	17.90	1536	228	16.81	1521	221	17.07	1567	214
35	15.20	1450	212	16.64	1583	217	16.26	1546	204
36	16.50	1512	235	15.01	1504	252	16.61	1494	253
37	16.27	1569	227	14.66	1575	205	15.05	1553	212
38	16.73	1693	205	14.71	1661	205	15.16	1691	208
39	17.17	1789	200	14.15	1826	204	17.30	1965	199
40	16.84	1493	253	15.47	1496	251	16.05	1504	244
41	17.62	1592	212	14.88	1526	229	16.05	1571	222
42	15.78	1472	217	14.44	1486	225	14.84	1468	200
43	17.76	1734	232	16.49	1785	244	15.56	1710	241
44	17.11	1462	221	15.92	1524	199	17.35	1477	233
45	16.16	1379	244	13.45	1388	222	14.93	1392	240
46	17.75	1633	225	15.09	1615	218	16.34	1638	222
47	20.26	1964	222	17.44	1817	235	17.68	1886	212
48	17.38	1537	210	14.82	1550	228	15.36	1527	208
49	16.33	1485	196	17.61	1420	227	17.97	1425	238
50	14.85	1493	224	16.68	1495	211	17.27	1452	216
51	16.32	1111	279	18.08	1154	247	17.35	1134	284
52	16.00	1114	239	21.60	1142	254	17.82	1114	227
53	14.65	1492	209	18.25	1479	219	16.15	1463	198
54	14.98	1662	240	17.52	1614	261	17.05	1716	224
55	17.57	1663	341	18.28	1646	306	18.14	1640	323

1.3.4 Chinese

Requirement-level satisfaction for Chinese under Claude Sonnet 4.6.

Chinese tasks 01–10					Chinese tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
02	R4	UNSAT.	UNSAT.	UNSAT.	12	R4	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
03	R3	SAT.	SAT.	SAT.	13	R2	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
04	R2	UNSAT.	UNSAT.	UNSAT.	14	R6	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
05	R1	UNSAT.	UNSAT.	UNSAT.	15	R3	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
07	R4	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
08	R3	UNSAT.	UNSAT.	UNSAT.	18	R8	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
09	R1	UNSAT.	UNSAT.	UNSAT.	19	R3	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
10	R5	UNSAT.	UNSAT.	UNSAT.	20	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.					

Chinese tasks 21–30					Chinese tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	UNSAT.	UNSAT.	UNSAT.	31	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	UNSAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
23	R5	UNSAT.	SAT.	SAT.	33	R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
24	R4	UNSAT.	UNSAT.	UNSAT.	34	R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
26	R2	UNSAT.	UNSAT.	UNSAT.	36	R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
27	R1	UNSAT.	UNSAT.	UNSAT.	37	R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
30	R5	UNSAT.	UNSAT.	UNSAT.	40	R4	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
31	R4	UNSAT.	UNSAT.	UNSAT.	41	R3	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.

Chinese tasks 41–50					Chinese tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	UNSAT.	UNSAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
43	R5	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R5	SAT.	SAT.	SAT.	55	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
45	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R9	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	UNSAT.
46	R5	UNSAT.	UNSAT.	UNSAT.	50	R11	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R12	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R13	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
47	R5	UNSAT.	SAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
48	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	SAT.	SAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
49	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
50	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					

Table 14: Claude Sonnet 4.6 / Chinese: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot				OpenHands			
	Time	Input	Output Tokens	Time	Input	Output Tokens		Time	Input	Output Tokens	
01	13.93	1556	330	16.20	1545	369		15.99	1531	331	
02	13.35	1320	292	15.96	1340	273		17.68	1350	307	
03	15.74	1391	323	18.73	1366	342		16.79	1374	328	
04	11.68	857	280	14.73	844	299		12.55	854	276	
05	13.66	1331	246	15.90	1325	263		15.67	1314	333	
06	13.15	1253	322	14.80	1266	284		15.11	1246	319	
07	15.34	1248	337	16.97	1208	329		16.51	1246	313	
08	14.05	1256	351	16.44	1259	337		17.20	1246	342	
09	11.29	1178	243	13.97	1194	235		13.08	1168	227	
10	14.86	1300	327	17.42	1302	304		16.85	1292	319	
11	12.82	1368	255	16.04	1386	256		15.07	1369	290	
12	16.14	1332	326	16.47	1346	314		15.69	1334	311	
13	15.01	1247	310	17.98	1228	314		15.70	1208	300	
14	12.16	1346	283	15.72	1347	278		15.52	1353	275	
15	12.82	1342	295	14.43	1346	302		14.46	1343	316	
16	12.21	1292	239	14.81	1292	288		15.41	1298	287	
17	14.34	1418	303	15.89	1415	323		16.73	1440	306	
18	13.78	1289	320	16.05	1228	307		15.19	1220	300	
19	13.94	1306	326	16.84	1303	306		16.55	1327	299	
20	11.82	1306	227	15.08	1296	297		13.55	1278	237	
21	13.63	1316	300	16.45	1314	297		15.93	1321	312	
22	14.67	1381	334	15.71	1361	310		15.87	1350	275	
23	12.78	1420	287	15.37	1419	283		14.63	1400	271	
24	13.35	1580	250	16.92	1572	277		15.87	1548	263	
25	12.47	1283	289	15.09	1277	305		15.99	1290	297	
26	13.10	1260	289	15.90	1272	280		14.17	1262	283	
27	15.52	1561	370	16.45	1563	342		17.79	1565	314	
28	13.24	1240	277	14.76	1235	278		14.23	1225	282	
29	13.15	1238	294	15.21	1234	269		14.79	1228	273	
30	14.73	1294	352	16.62	1300	329		17.11	1292	327	
31	14.20	1435	352	15.54	1438	329		17.87	1418	346	
32	14.29	1261	336	15.65	1259	337		15.78	1242	309	
33	12.98	1346	330	15.30	1330	347		12.92	1323	331	
34	15.16	1348	329	17.13	1366	306		14.54	1343	315	
35	11.98	1293	233	15.74	1280	284		13.52	1307	232	
36	13.54	1325	334	15.15	1299	329		13.85	1307	344	
37	12.74	1320	314	15.05	1346	308		13.73	1320	311	
38	13.24	1494	257	15.23	1496	303		12.36	1500	263	
39	11.58	1564	296	13.76	1554	314		12.50	1546	306	
40	13.79	1262	326	16.61	1267	339		14.56	1270	352	
41	13.66	1315	316	15.97	1316	296		13.03	1292	282	
42	14.33	1296	278	17.05	1299	317		14.06	1304	297	
43	13.54	1500	305	16.81	1507	318		13.13	1484	252	
44	14.07	1253	300	16.77	1249	353		15.56	1269	341	
45	12.40	1201	296	17.80	1224	318		13.24	1206	281	
46	13.68	1403	304	18.53	1434	324		12.63	1386	272	
47	13.99	1412	271	16.59	1393	293		13.30	1379	274	
48	14.30	1336	304	14.66	1337	250		12.53	1318	258	
49	15.97	1253	287	15.57	1237	321		13.43	1247	294	
50	14.92	1320	306	14.75	1321	267		13.82	1324	260	
51	16.53	994	342	17.01	1037	315		17.68	1023	378	
52	15.99	957	261	14.63	952	280		15.29	949	259	
53	14.60	1258	289	13.93	1267	256		15.73	1249	292	
54	16.75	1426	455	17.15	1422	389		16.91	1408	396	
55	18.20	1462	412	18.06	1462	416		17.83	1465	375	

1.3.5 Hindi

Requirement-level satisfaction for Hindi under Claude Sonnet 4.6.

Hindi tasks 01–10					Hindi tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
02	R4	UNSAT.	UNSAT.	UNSAT.	12	R4	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
03	R3	SAT.	SAT.	SAT.	13	R2	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
04	R2	UNSAT.	UNSAT.	UNSAT.	14	R6	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
05	R1	UNSAT.	UNSAT.	UNSAT.	15	R3	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
07	R4	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
08	R3	UNSAT.	UNSAT.	UNSAT.	18	R8	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
09	R1	UNSAT.	UNSAT.	UNSAT.	19	R3	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
10	R5	UNSAT.	UNSAT.	UNSAT.	20	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.					

Hindi tasks 21–30					Hindi tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	UNSAT.	UNSAT.	UNSAT.	31	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	UNSAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	UNSAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	SAT.
23	R5	SAT.	SAT.	SAT.	33	R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
24	R4	UNSAT.	UNSAT.	UNSAT.	34	R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
26	R2	UNSAT.	UNSAT.	UNSAT.	36	R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
27	R1	UNSAT.	UNSAT.	UNSAT.	37	R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	SAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
30	R5	UNSAT.	UNSAT.	UNSAT.	40	R4	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
31	R4	UNSAT.	SAT.	SAT.	41	R3	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.

Hindi tasks 41–50					Hindi tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	UNSAT.	UNSAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
43	R5	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R5	UNSAT.	UNSAT.	UNSAT.	55	R6	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
45	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R9	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	UNSAT.
46	R5	UNSAT.	UNSAT.	UNSAT.		R11	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R12	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R13	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
47	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
48	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
49	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
50	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					

Table 15: Claude Sonnet 4.6 / Hindi: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	14.62	1903	146	16.07	1909	145	15.48	1897	142
02	14.67	1688	153	17.16	1668	160	18.62	1668	170
03	13.74	1713	183	16.49	1666	177	17.73	1644	168
04	13.65	1186	205	16.51	1190	212	16.49	1225	223
05	14.80	1609	171	13.13	1570	177	16.07	1562	163
06	13.54	1623	160	12.35	1599	151	15.15	1539	143
07	13.67	1622	162	13.18	1499	166	16.07	1583	162
08	14.38	1580	166	14.63	1547	172	17.23	1511	161
09	11.84	1552	146	12.36	1606	148	14.47	1490	128
10	13.48	1605	184	13.37	1588	186	16.12	1582	157
11	13.60	1732	163	13.40	1687	151	17.59	1717	153
12	13.43	1692	155	13.50	1643	157	17.40	1617	156
13	14.56	1577	163	14.79	1531	161	16.83	1509	154
14	12.97	1727	167	12.52	1685	147	15.92	1658	152
15	11.94	1599	157	12.77	1580	146	15.87	1567	163
16	13.33	1664	149	13.40	1644	153	17.18	1668	177
17	12.79	1719	161	12.97	1714	156	15.51	1693	170
18	12.12	1508	158	13.20	1506	193	15.86	1510	173
19	14.04	1684	158	13.57	1604	156	17.93	1729	157
20	12.32	1632	140	13.13	1602	141	15.23	1594	160
21	13.89	1668	154	13.54	1607	144	15.94	1597	160
22	13.73	1698	162	13.96	1667	153	16.18	1611	166
23	13.80	1784	159	14.85	1751	158	15.31	1734	154
24	13.56	1934	154	15.87	1861	163	15.22	1816	154
25	13.27	1639	168	14.86	1593	136	15.24	1565	161
26	13.86	1692	165	16.36	1691	166	16.14	1636	136
27	14.82	1978	160	15.82	1918	154	16.65	1842	165
28	12.34	1525	145	13.33	1475	136	14.08	1479	143
29	12.80	1608	149	15.15	1592	136	14.49	1547	153
30	14.30	1657	196	16.65	1617	184	15.56	1569	179
31	13.23	1738	163	16.22	1715	174	16.61	1752	172
32	14.04	1618	171	15.35	1577	165	16.39	1569	174
33	12.18	1645	163	15.29	1647	173	15.12	1589	167
34	15.63	1586	171	15.96	1582	160	15.17	1578	175
35	13.07	1601	153	15.78	1628	155	16.21	1547	153
36	12.80	1624	158	14.50	1590	162	14.25	1548	162
37	12.96	1691	164	15.34	1636	158	14.68	1628	156
38	10.96	1776	155	14.70	1793	149	14.52	1748	148
39	12.97	1899	149	14.24	1838	153	14.05	1809	157
40	13.25	1573	159	14.42	1527	155	16.61	1523	162
41	14.63	1646	153	16.93	1618	160	14.88	1570	147
42	16.04	1664	154	15.08	1611	158	15.39	1606	157
43	15.84	1792	160	15.30	1740	191	15.21	1714	176
44	18.66	1650	178	17.80	1680	173	16.89	1615	173
45	14.03	1503	161	14.53	1484	165	14.43	1452	153
46	14.74	1686	175	16.04	1670	169	14.35	1646	168
47	15.46	1703	157	15.42	1650	158	15.99	1653	157
48	13.75	1586	161	16.57	1616	154	15.31	1578	162
49	14.69	1605	173	14.68	1579	172	15.43	1550	162
50	14.14	1637	152	13.08	1614	160	14.41	1603	142
51	17.67	1378	260	14.25	1356	196	14.96	1332	230
52	16.98	1256	210	15.34	1289	223	15.09	1189	227
53	15.23	1606	184	12.89	1604	164	14.18	1552	181
54	15.94	1860	170	13.22	1836	175	15.13	1770	184
55	16.52	1810	218	14.06	1772	191	17.55	1799	250

1.4 Gemini 3 Flash Preview

1.4.1 English

Requirement-level satisfaction for English under Gemini 3 Flash Preview.

English tasks 01–10					English tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
02	R4	UNSAT.	UNSAT.	UNSAT.	12	R4	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
03	R3	SAT.	SAT.	SAT.	13	R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	SAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
04	R2	UNSAT.	UNSAT.	UNSAT.	14	R6	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
05	R1	UNSAT.	UNSAT.	UNSAT.	15	R3	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R6	SAT.	SAT.	SAT.
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
07	R4	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
08	R3	SAT.	SAT.	SAT.	18	R2	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
09	R2	UNSAT.	UNSAT.	UNSAT.	19	R0	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	SAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
10	R0	UNSAT.	UNSAT.	UNSAT.	20	R4	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
						R4	UNSAT.	UNSAT.	UNSAT.
						R5	SAT.	SAT.	SAT.
						R0	UNSAT.	UNSAT.	UNSAT.
						R1	UNSAT.	UNSAT.	UNSAT.
						R2	UNSAT.	UNSAT.	UNSAT.
						R3	UNSAT.	UNSAT.	UNSAT.
						R4	UNSAT.	UNSAT.	UNSAT.
						R5	SAT.	SAT.	SAT.
						R0	UNSAT.	UNSAT.	UNSAT.
						R1	UNSAT.	UNSAT.	UNSAT.

English tasks 21–30					English tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	SAT.	SAT.	SAT.	31	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	UNSAT.	UNSAT.	UNSAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
23	R5	SAT.	SAT.	SAT.	33	R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
24	R4	UNSAT.	UNSAT.	UNSAT.	34	R5	UNSAT.	SAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
26	R2	UNSAT.	UNSAT.	UNSAT.	36	R3	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
27	R1	UNSAT.	UNSAT.	UNSAT.	37	R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	SAT.
	R5	SAT.	SAT.	SAT.		R6	SAT.	SAT.	SAT.
	R6	UNSAT.	SAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
30	R5	UNSAT.	UNSAT.	UNSAT.	40	R4	UNSAT.	UNSAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R6	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.

English tasks 41–50					English tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	UNSAT.	UNSAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	SAT.	UNSAT.	SAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	SAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
42	R0	SAT.	UNSAT.	SAT.	53	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
43	R0	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R0	UNSAT.	UNSAT.	UNSAT.	55	R6	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
45	R0	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	SAT.
46	R5	UNSAT.	UNSAT.	UNSAT.		R9	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R11	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R12	SAT.	SAT.	SAT.
	R3	SAT.	SAT.	SAT.		R13	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.					
47	R0	UNSAT.	UNSAT.	UNSAT.	48	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	SAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
48	R6	UNSAT.	UNSAT.	UNSAT.	49	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
49	R5	SAT.	SAT.	SAT.	50	R5	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
50	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					

Table 16: Gemini 3 Flash Preview / English: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	9.57	1412	113	7.68	1434	109	7.98	1423	124
02	9.63	1207	111	8.34	1190	113	8.28	1231	120
03	8.70	1236	126	8.48	1243	124	7.82	1246	125
04	6.10	786	79	7.95	784	76	6.17	788	86
05	8.21	1205	107	7.23	1211	104	7.58	1201	105
06	6.92	1196	113	7.50	1180	112	7.30	1198	110
07	7.48	1169	104	7.74	1179	106	8.01	1180	115
08	7.54	1134	105	8.06	1138	110	7.11	1142	106
09	7.64	1159	109	6.97	1162	103	6.33	1145	105
10	7.34	1184	110	7.26	1200	119	6.75	1200	116
11	7.98	1251	124	8.10	1254	107	7.41	1260	119
12	7.74	1216	119	8.11	1221	109	7.74	1231	120
13	8.04	1138	112	7.71	1155	109	8.03	1144	97
14	7.23	1199	113	7.44	1211	112	7.18	1204	109
15	7.82	1207	109	7.46	1212	104	7.76	1205	98
16	7.07	1175	101	7.32	1179	110	6.83	1170	99
17	6.93	1276	99	6.95	1272	105	6.52	1264	105
18	7.67	1163	94	9.36	1188	91	6.44	1163	100
19	7.47	1202	108	7.47	1201	117	7.04	1191	104
20	7.49	1160	112	6.89	1173	107	6.58	1158	111
21	7.06	1174	99	7.35	1190	105	6.66	1187	109
22	7.18	1213	104	7.24	1209	93	6.89	1211	95
23	6.86	1293	107	7.09	1291	110	6.52	1280	106
24	7.94	1394	117	7.14	1417	104	6.53	1390	117
25	7.20	1212	106	7.16	1202	108	6.53	1191	104
26	6.89	1152	108	7.36	1161	111	6.58	1151	108
27	7.83	1429	123	8.28	1402	121	7.43	1417	109
28	6.58	1092	101	6.96	1090	97	6.93	1089	105
29	6.98	1155	94	6.98	1128	100	6.98	1133	105
30	6.77	1145	97	6.83	1150	103	6.94	1156	85
31	8.30	1303	110	7.58	1325	122	7.34	1305	112
32	6.89	1144	116	7.26	1144	105	7.35	1150	111
33	6.94	1220	104	7.46	1244	98	7.22	1253	105
34	7.59	1175	112	7.58	1202	109	7.43	1180	109
35	7.26	1161	98	7.62	1164	106	7.70	1166	104
36	7.11	1230	115	6.98	1210	114	7.15	1205	113
37	6.71	1228	105	6.99	1211	103	6.66	1226	101
38	7.12	1377	103	6.99	1384	96	6.58	1371	104
39	7.02	1427	126	7.20	1429	125	7.88	1552	141
40	7.17	1143	110	7.52	1154	105	7.31	1138	109
41	6.94	1171	101	7.56	1196	113	6.99	1169	112
42	7.51	1179	117	7.96	1174	105	7.05	1160	92
43	7.79	1357	111	7.81	1369	121	7.36	1352	118
44	7.24	1114	99	7.38	1112	107	6.95	1110	102
45	6.78	1082	101	7.69	1088	101	6.76	1094	100
46	6.73	1269	105	7.03	1264	101	6.73	1264	100
47	8.22	1244	129	7.81	1255	129	6.65	1253	123
48	7.41	1224	108	7.91	1210	108	6.90	1214	104
49	6.86	1123	100	7.62	1131	102	6.67	1118	104
50	6.91	1179	104	6.87	1170	101	6.85	1185	101
51	8.01	962	106	8.51	987	109	7.66	1031	118
52	7.49	846	110	7.23	843	100	6.92	860	102
53	6.31	1170	94	7.32	1197	95	7.12	1191	96
54	6.90	1328	111	7.26	1334	112	7.11	1328	116
55	7.20	1322	128	7.78	1370	122	6.72	1342	115

1.4.2 Arabic

Requirement-level satisfaction for Arabic under Gemini 3 Flash Preview.

Arabic tasks 01–10					Arabic tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	SAT.	SAT.	SAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
02	R4	UNSAT.	UNSAT.	UNSAT.	12	R4	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R0	UNSAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
03	R3	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
04	R0	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.	13	R5	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
05	R4	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	SAT.
	R0	UNSAT.	SAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
06	R2	SAT.	SAT.	SAT.	14	R4	UNSAT.	SAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R6	SAT.	SAT.	SAT.
	R0	UNSAT.	SAT.	UNSAT.		R0	SAT.	SAT.	SAT.
07	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.	15	R2	UNSAT.	SAT.	SAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
08	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.	16	R0	UNSAT.	UNSAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
09	R4	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
10	R0	UNSAT.	UNSAT.	UNSAT.	17	R5	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
11	R4	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.
12	R0	SAT.	SAT.	SAT.	18	R4	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
13	R4	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
14	R0	UNSAT.	UNSAT.	UNSAT.	19	R4	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.	20	R0	SAT.	SAT.	SAT.
15	R4	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
16	R3	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
17	R0	UNSAT.	UNSAT.	UNSAT.	20	R0	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
18	R4	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
19	R0	UNSAT.	UNSAT.	UNSAT.	20	R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
20	R4	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.

Arabic tasks 21–30					Arabic tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	SAT.	SAT.	SAT.	31	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	UNSAT.	UNSAT.	UNSAT.	32	R0	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
23	R5	SAT.	SAT.	SAT.	33	R6	UNSAT.	UNSAT.	UNSAT.
	R6	SAT.	SAT.	SAT.		R0	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
24	R4	UNSAT.	UNSAT.	UNSAT.	34	R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	SAT.
	R0	UNSAT.	SAT.	UNSAT.		R1	SAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
26	R2	UNSAT.	UNSAT.	UNSAT.	36	R3	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	UNSAT.	SAT.
	R4	SAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R6	SAT.	SAT.	SAT.
	R6	SAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
27	R1	SAT.	SAT.	UNSAT.	37	R2	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	SAT.	SAT.	SAT.
	R6	SAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	SAT.		R7	SAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R3	UNSAT.	SAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
30	R5	UNSAT.	UNSAT.	UNSAT.	40	R4	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R6	SAT.	SAT.	SAT.
	R1	UNSAT.	SAT.	UNSAT.		R0	SAT.	SAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
31	R4	SAT.	SAT.	SAT.	41	R3	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.					
	R2	SAT.	SAT.	SAT.					

Arabic tasks 41–50					Arabic tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	SAT.	SAT.	SAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
	R3	SAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	SAT.	UNSAT.	SAT.		R1	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R0	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	SAT.
43	R5	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R1	UNSAT.	SAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	SAT.
	R3	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R4	SAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
44	R5	SAT.	SAT.	SAT.	55	R6	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	SAT.	SAT.
	R0	SAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
	R3	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
45	R4	SAT.	SAT.	SAT.		R5	UNSAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	SAT.		R8	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R9	SAT.	SAT.	SAT.
	R3	SAT.	SAT.	SAT.		R10	UNSAT.	UNSAT.	UNSAT.
46	R4	UNSAT.	UNSAT.	UNSAT.		R11	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R12	SAT.	SAT.	SAT.
	R0	UNSAT.	SAT.	UNSAT.		R13	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	SAT.	SAT.					
47	R4	SAT.	SAT.	SAT.					
	R5	SAT.	SAT.	SAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	SAT.	SAT.					
	R1	UNSAT.	SAT.	SAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
48	R3	UNSAT.	SAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	SAT.	SAT.	SAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	SAT.	SAT.					
49	R0	SAT.	SAT.	SAT.					
	R1	SAT.	SAT.	SAT.					
	R2	SAT.	SAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
50	R6	SAT.	SAT.	SAT.					
	R0	UNSAT.	SAT.	SAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	SAT.	SAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					

Table 17: Gemini 3 Flash Preview / Arabic: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	7.44	1573	137	7.28	1571	142	7.14	1537	131
02	8.53	1367	108	8.59	1356	107	9.23	1354	111
03	9.76	1436	133	8.66	1439	126	7.99	1414	126
04	7.28	969	97	7.55	974	102	7.21	969	100
05	9.00	1428	114	8.30	1413	113	8.78	1435	126
06	7.79	1329	118	7.29	1328	112	7.80	1313	130
07	8.71	1329	118	7.58	1334	105	7.84	1354	120
08	8.40	1287	114	7.80	1316	112	7.58	1302	123
09	7.52	1327	137	7.69	1329	146	6.66	1308	126
10	7.73	1356	150	8.22	1356	151	7.32	1349	122
11	7.65	1364	128	7.20	1392	135	7.62	1374	134
12	8.03	1395	126	7.79	1408	132	7.97	1367	123
13	8.57	1309	105	7.24	1284	103	7.91	1292	105
14	8.06	1375	147	8.26	1387	125	7.06	1352	116
15	8.64	1384	110	7.50	1426	112	7.51	1396	99
16	7.68	1329	109	7.27	1347	135	7.31	1334	140
17	8.02	1437	116	7.39	1454	132	7.57	1454	111
18	7.68	1274	104	7.28	1280	106	7.55	1303	108
19	8.12	1394	114	7.65	1415	131	8.19	1398	120
20	7.47	1384	120	7.58	1332	109	7.66	1353	118
21	8.53	1384	135	7.41	1366	126	7.48	1359	123
22	7.89	1372	114	6.92	1375	117	7.67	1386	111
23	8.09	1474	135	7.68	1490	147	7.44	1479	124
24	8.87	1595	147	7.55	1605	136	8.00	1574	138
25	8.15	1360	121	8.25	1382	135	7.70	1377	123
26	7.66	1366	118	8.20	1362	128	7.66	1350	113
27	7.80	1586	129	8.45	1602	147	8.62	1596	134
28	7.64	1242	110	6.90	1239	130	7.12	1241	119
29	7.12	1265	122	6.56	1271	130	7.06	1271	117
30	7.36	1285	133	7.23	1329	99	7.47	1318	108
31	8.90	1488	131	7.77	1511	130	7.46	1464	124
32	8.26	1327	116	8.21	1333	127	7.70	1335	120
33	8.23	1424	120	7.77	1430	120	7.45	1410	111
34	8.21	1373	118	8.52	1389	133	8.38	1399	124
35	7.36	1304	105	8.42	1362	126	8.64	1361	109
36	7.07	1349	116	8.06	1372	121	7.43	1360	115
37	7.75	1421	108	7.47	1434	124	7.72	1437	101
38	7.38	1546	117	7.79	1548	123	7.50	1545	127
39	7.40	1647	120	7.01	1605	116	7.89	1642	133
40	8.03	1328	128	8.21	1344	138	7.75	1315	134
41	7.59	1362	110	8.20	1425	113	7.54	1400	110
42	7.81	1318	106	8.17	1323	118	8.09	1322	113
43	7.71	1526	152	7.83	1536	152	7.94	1513	142
44	7.71	1312	111	7.93	1334	108	8.23	1300	109
45	7.68	1246	107	6.97	1280	112	7.35	1253	114
46	8.12	1406	127	7.37	1445	120	8.34	1429	131
47	8.10	1452	170	7.57	1443	141	8.24	1451	156
48	8.38	1428	126	7.68	1413	108	7.75	1429	114
49	7.74	1312	101	7.52	1336	119	7.61	1352	119
50	7.47	1354	128	7.65	1358	112	7.64	1359	123
51	7.85	1121	111	8.15	1068	118	8.03	1086	119
52	8.26	1094	129	8.85	1091	115	8.00	1096	101
53	7.54	1354	114	7.64	1321	100	8.47	1370	106
54	7.53	1533	130	7.85	1540	140	8.89	1530	132
55	7.94	1502	136	7.66	1509	143	7.82	1491	133

1.4.3 Turkish

Requirement-level satisfaction for Turkish under Gemini 3 Flash Preview.

Turkish tasks 01–10					Turkish tasks 11–20					
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands	
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.	
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.	
	R2	SAT.	SAT.	SAT.		R2	SAT.	UNSAT.	UNSAT.	
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.	
	R4	SAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.	
02	R0	SAT.	UNSAT.	UNSAT.	12	R5	SAT.	SAT.	SAT.	
	R1	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.	
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.	
	R3	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.	
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.	
03	R0	SAT.	SAT.	SAT.	13	R4	UNSAT.	UNSAT.	UNSAT.	
	R1	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.	
	R2	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.	
	R3	UNSAT.	UNSAT.	UNSAT.		14	R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.			R1	UNSAT.	SAT.	SAT.
04	R0	SAT.	SAT.	SAT.	R2		SAT.	SAT.	SAT.	
	R1	SAT.	SAT.	SAT.	R3		SAT.	SAT.	SAT.	
	R2	SAT.	SAT.	SAT.	R4		UNSAT.	SAT.	UNSAT.	
	R3	SAT.	SAT.	SAT.	R5	SAT.	UNSAT.	UNSAT.		
	R4	SAT.	SAT.	SAT.	R6	SAT.	SAT.	SAT.		
05	R0	UNSAT.	UNSAT.	UNSAT.	15	R0	UNSAT.	SAT.	UNSAT.	
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.	
	R2	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.	
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.	
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.	
06	R0	SAT.	UNSAT.	UNSAT.	16	R5	UNSAT.	UNSAT.	UNSAT.	
	R1	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.	
	R2	UNSAT.	UNSAT.	UNSAT.		17	R0	SAT.	SAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.			R1	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.			R2	UNSAT.	UNSAT.	UNSAT.
07	R0	UNSAT.	UNSAT.	UNSAT.	R3		UNSAT.	UNSAT.	UNSAT.	
	R1	UNSAT.	SAT.	UNSAT.	R4		SAT.	SAT.	SAT.	
	R2	SAT.	SAT.	SAT.	R5	SAT.	SAT.	SAT.		
	R3	SAT.	SAT.	SAT.	18	R0	UNSAT.	SAT.	SAT.	
	R4	SAT.	SAT.	SAT.		R1	UNSAT.	SAT.	SAT.	
08	R0	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.	
	R1	UNSAT.	SAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.	
	R2	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.	
	R3	UNSAT.	UNSAT.	UNSAT.	R5	SAT.	SAT.	SAT.		
	R4	SAT.	SAT.	SAT.	R6	UNSAT.	UNSAT.	UNSAT.		
09	R5	UNSAT.	UNSAT.	UNSAT.	19	R0	UNSAT.	SAT.	SAT.	
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.	
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.	
	R2	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.	
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.	
10	R4	UNSAT.	UNSAT.	UNSAT.	20	R5	SAT.	SAT.	SAT.	
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.	
	R0	UNSAT.	SAT.	SAT.		21	R0	UNSAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.			R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.			R2	UNSAT.	UNSAT.	UNSAT.
R3	UNSAT.	UNSAT.	UNSAT.	R3	UNSAT.		UNSAT.	UNSAT.		
R4	UNSAT.	UNSAT.	UNSAT.	R4	UNSAT.		UNSAT.	UNSAT.		
11	R5	SAT.	SAT.	SAT.	22	R5	SAT.	SAT.	SAT.	
	R0	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.	
	R1	UNSAT.	UNSAT.	UNSAT.		23	R0	UNSAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.			R1	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.			R2	UNSAT.	UNSAT.	UNSAT.
R4	UNSAT.	UNSAT.	UNSAT.	R3	UNSAT.		UNSAT.	UNSAT.		
12	R5	SAT.	SAT.	SAT.	24		R4	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.	
	R1	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.	
	R2	UNSAT.	UNSAT.	UNSAT.		25	R0	UNSAT.	SAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.			R1	UNSAT.	UNSAT.	UNSAT.
R4	UNSAT.	UNSAT.	UNSAT.	R2	UNSAT.		UNSAT.	UNSAT.		
R5	SAT.	SAT.	SAT.	R3	UNSAT.		UNSAT.	UNSAT.		
13	R0	UNSAT.	UNSAT.	UNSAT.	26		R4	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.	
	R2	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.	
	R3	UNSAT.	UNSAT.	UNSAT.		27	R0	UNSAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.			R1	UNSAT.	UNSAT.	UNSAT.
R5	SAT.	SAT.	SAT.	R2	UNSAT.		UNSAT.	UNSAT.		
R0	UNSAT.	UNSAT.	UNSAT.	R3	UNSAT.		UNSAT.	UNSAT.		
R1	UNSAT.	UNSAT.	UNSAT.	R4	UNSAT.		UNSAT.	UNSAT.		
R2	UNSAT.	UNSAT.	UNSAT.	R5	SAT.	SAT.	SAT.			
R3	UNSAT.	UNSAT.	UNSAT.	R6	UNSAT.	UNSAT.	UNSAT.			
R4	UNSAT.	UNSAT.	UNSAT.	R7	SAT.	UNSAT.	SAT.			
R5	SAT.	SAT.	SAT.	R8	SAT.	SAT.	SAT.			

Turkish tasks 21–30					Turkish tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	SAT.	SAT.	SAT.	31	R0	SAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R0	SAT.	SAT.	SAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
23	R0	SAT.	SAT.	SAT.	33	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	SAT.
24	R0	SAT.	SAT.	SAT.	34	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
25	R0	SAT.	SAT.	SAT.	35	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
26	R0	UNSAT.	UNSAT.	UNSAT.	36	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	UNSAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	SAT.	UNSAT.	UNSAT.
27	R0	UNSAT.	UNSAT.	UNSAT.	37	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	SAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	SAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	UNSAT.		R4	SAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
29	R0	SAT.	UNSAT.	SAT.	39	R0	UNSAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	UNSAT.	SAT.
	R5	UNSAT.	SAT.	UNSAT.		R5	SAT.	SAT.	SAT.
30	R0	UNSAT.	UNSAT.	UNSAT.	40	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
31	R0	SAT.	SAT.	SAT.	41	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.

Turkish tasks 41–50					Turkish tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	SAT.	SAT.	SAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	UNSAT.	SAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	SAT.	SAT.	SAT.
	R0	SAT.	UNSAT.	SAT.		R0	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.	54	R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
43	R5	UNSAT.	UNSAT.	UNSAT.	55	R0	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
44	R2	UNSAT.	UNSAT.	UNSAT.	55	R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	SAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.	55	R6	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
45	R1	UNSAT.	UNSAT.	UNSAT.	55	R9	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	SAT.	UNSAT.
	R3	UNSAT.	SAT.	SAT.		R11	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.	55	R12	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R13	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.					
46	R0	UNSAT.	UNSAT.	UNSAT.	55				
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
47	R3	SAT.	SAT.	SAT.	55				
	R4	SAT.	SAT.	SAT.					
	R5	SAT.	SAT.	SAT.					
48	R6	UNSAT.	UNSAT.	UNSAT.	55				
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
49	R2	UNSAT.	UNSAT.	UNSAT.	55				
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
50	R5	SAT.	SAT.	SAT.	55				
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					
51	R0	UNSAT.	UNSAT.	UNSAT.	55				
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
52	R3	UNSAT.	UNSAT.	UNSAT.	55				
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
53	R6	UNSAT.	UNSAT.	UNSAT.	55				
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	UNSAT.	UNSAT.	UNSAT.					
54	R2	UNSAT.	UNSAT.	UNSAT.	55				
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	UNSAT.	UNSAT.	UNSAT.					
55	R5	UNSAT.	UNSAT.	UNSAT.	55				
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					

Table 18: Gemini 3 Flash Preview / Turkish: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	9.45	1410	117	7.18	1417	144	7.29	1418	118
02	8.74	1223	122	8.21	1224	129	7.53	1189	119
03	8.58	1273	142	8.73	1278	122	8.01	1275	121
04	7.55	857	95	8.06	862	97	7.85	865	100
05	8.29	1192	116	7.51	1186	133	7.80	1211	118
06	7.15	1186	121	7.86	1186	124	7.96	1189	119
07	8.55	1164	132	8.53	1196	129	7.43	1158	137
08	7.68	1173	135	8.31	1166	124	7.38	1174	124
09	7.24	1158	118	7.93	1176	119	6.79	1142	118
10	7.38	1227	149	8.02	1244	145	7.56	1222	144
11	7.18	1226	129	7.63	1203	122	8.15	1247	138
12	7.95	1237	128	7.75	1239	126	7.96	1213	136
13	7.70	1196	121	8.19	1168	125	8.37	1178	126
14	7.55	1222	132	7.44	1244	128	7.54	1237	128
15	7.53	1221	118	7.09	1214	112	7.53	1235	120
16	6.76	1140	117	6.81	1144	114	6.75	1148	120
17	7.72	1321	129	7.35	1318	125	7.09	1301	118
18	7.25	1156	115	7.98	1148	103	7.02	1164	114
19	7.45	1212	133	8.00	1228	134	7.36	1215	130
20	7.24	1156	118	7.63	1189	129	7.57	1177	117
21	7.54	1196	126	7.48	1218	127	8.08	1203	127
22	7.43	1247	121	7.32	1237	109	7.27	1242	119
23	6.91	1276	132	7.37	1325	127	6.87	1274	118
24	7.21	1415	142	9.20	1419	141	7.54	1421	145
25	7.88	1227	129	8.01	1227	123	8.11	1188	123
26	7.42	1201	118	7.38	1238	129	7.46	1173	133
27	8.31	1458	120	8.01	1427	145	7.30	1410	131
28	7.24	1115	123	7.47	1162	120	6.49	1089	133
29	7.13	1154	136	6.96	1146	125	7.39	1189	130
30	7.55	1199	130	8.05	1209	121	8.39	1216	127
31	7.74	1303	141	7.84	1322	162	8.00	1323	163
32	8.03	1148	133	8.37	1197	133	7.71	1192	133
33	7.42	1231	128	7.19	1270	123	7.88	1272	125
34	7.71	1197	161	7.71	1208	127	9.25	1176	127
35	7.17	1158	115	7.25	1150	121	8.76	1152	117
36	7.94	1227	120	7.30	1239	147	7.87	1221	127
37	7.65	1226	128	7.10	1216	136	10.01	1205	123
38	7.06	1348	117	6.86	1335	120	7.05	1335	124
39	7.37	1399	129	6.74	1405	135	8.16	1526	130
40	7.83	1180	125	7.47	1182	119	7.08	1112	116
41	7.69	1240	124	7.36	1233	126	7.96	1243	128
42	7.95	1224	119	7.43	1222	119	8.15	1220	114
43	8.35	1381	145	7.84	1399	157	7.97	1391	144
44	7.42	1129	126	7.52	1138	122	7.42	1149	120
45	6.98	1103	118	6.88	1094	121	6.86	1083	122
46	7.55	1250	116	6.93	1260	122	7.12	1249	128
47	8.04	1307	152	7.52	1294	151	7.93	1301	127
48	7.74	1226	138	7.30	1226	123	7.47	1215	135
49	7.34	1134	126	7.33	1133	129	7.30	1141	116
50	7.25	1184	114	7.79	1242	120	7.74	1198	123
51	7.62	946	127	7.94	965	130	8.47	958	139
52	6.99	954	138	7.83	868	126	7.30	904	123
53	7.36	1190	120	8.06	1243	130	7.11	1221	124
54	7.53	1370	152	8.00	1358	141	7.53	1343	157
55	7.80	1394	153	8.05	1385	154	7.67	1377	162

1.4.4 Chinese

Requirement-level satisfaction for Chinese under Gemini 3 Flash Preview.

Chinese tasks 01–10					Chinese tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	SAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
02	R4	UNSAT.	UNSAT.	UNSAT.	12	R4	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
03	R3	SAT.	SAT.	SAT.	13	R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
04	R2	UNSAT.	UNSAT.	UNSAT.	14	R6	UNSAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
05	R1	SAT.	SAT.	SAT.	15	R3	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R4	UNSAT.	SAT.	SAT.
	R3	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
	R4	SAT.	UNSAT.	SAT.		R6	SAT.	SAT.	SAT.
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.
07	R4	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
08	R3	SAT.	SAT.	SAT.	18	R1	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
09	R2	UNSAT.	UNSAT.	UNSAT.	19	R5	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
10	R0	UNSAT.	UNSAT.	UNSAT.	20	R2	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R0	UNSAT.	SAT.	UNSAT.
						R1	SAT.	SAT.	SAT.
						R2	UNSAT.	UNSAT.	UNSAT.
						R3	SAT.	SAT.	SAT.
						R4	SAT.	SAT.	SAT.
						R5	UNSAT.	UNSAT.	UNSAT.
						R6	SAT.	SAT.	SAT.
						R0	UNSAT.	UNSAT.	UNSAT.
						R1	UNSAT.	UNSAT.	UNSAT.
						R2	UNSAT.	UNSAT.	UNSAT.
						R3	UNSAT.	UNSAT.	UNSAT.
						R4	UNSAT.	UNSAT.	UNSAT.
						R5	SAT.	SAT.	SAT.
						R6	UNSAT.	UNSAT.	UNSAT.
						R0	UNSAT.	SAT.	UNSAT.

Chinese tasks 21–30					Chinese tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	SAT.	SAT.	SAT.	31	R0	UNSAT.	SAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R0	SAT.	SAT.	SAT.	32	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	SAT.	UNSAT.		R2	SAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
23	R0	SAT.	UNSAT.	UNSAT.	33	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
24	R0	SAT.	UNSAT.	UNSAT.	34	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
25	R0	UNSAT.	UNSAT.	UNSAT.	35	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	UNSAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
26	R0	UNSAT.	UNSAT.	UNSAT.	36	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	SAT.	SAT.
27	R0	UNSAT.	UNSAT.	UNSAT.	37	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	SAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	SAT.	UNSAT.	38	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	UNSAT.		R1	SAT.	SAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	SAT.
	R3	SAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
29	R0	SAT.	SAT.	UNSAT.	39	R0	UNSAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
30	R0	UNSAT.	SAT.	UNSAT.	40	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
31	R0	UNSAT.	UNSAT.	UNSAT.	41	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.

Chinese tasks 41–50					Chinese tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	UNSAT.	SAT.	UNSAT.	51	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	UNSAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	UNSAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
43	R4	SAT.	UNSAT.	SAT.	54	R4	UNSAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
44	R3	UNSAT.	UNSAT.	UNSAT.	55	R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	SAT.	UNSAT.
	R0	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
45	R1	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R9	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R10	SAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R11	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R12	SAT.	SAT.	SAT.
46	R6	UNSAT.	UNSAT.	UNSAT.		R13	SAT.	SAT.	SAT.
	R0	UNSAT.	UNSAT.	UNSAT.	47	R0	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
47	R5	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
	R6	UNSAT.	SAT.	UNSAT.		R6	UNSAT.	SAT.	UNSAT.
48	R0	UNSAT.	UNSAT.	UNSAT.	49	R0	SAT.	UNSAT.	SAT.
	R1	UNSAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
49	R7	UNSAT.	UNSAT.	UNSAT.	50	R0	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R5	SAT.	UNSAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
50	R6	SAT.	SAT.	SAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R7	UNSAT.	UNSAT.	UNSAT.					

Table 19: Gemini 3 Flash Preview / Chinese: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	8.37	1433	137	8.37	1459	150	11.38	1435	161
02	8.83	1223	149	9.17	1255	154	9.23	1243	155
03	8.89	1269	150	9.13	1289	163	10.17	1269	158
04	7.56	794	121	7.09	789	125	7.49	775	110
05	9.62	1282	146	8.86	1262	153	8.95	1258	171
06	8.55	1226	160	8.24	1218	126	8.68	1239	145
07	10.57	1200	161	8.55	1187	150	9.22	1193	161
08	9.22	1168	127	9.49	1194	121	10.60	1184	156
09	7.98	1141	142	8.25	1141	147	7.55	1133	140
10	8.44	1234	155	8.33	1233	157	8.68	1237	162
11	8.25	1238	137	9.67	1280	160	7.67	1247	135
12	8.64	1272	146	8.83	1269	148	8.66	1282	140
13	8.59	1152	129	9.02	1167	150	9.56	1160	152
14	8.34	1253	147	8.38	1243	159	8.67	1242	160
15	8.60	1239	160	8.87	1272	149	8.15	1274	151
16	8.38	1186	124	8.47	1195	138	7.85	1197	134
17	7.84	1298	135	8.62	1310	148	8.21	1296	148
18	7.91	1153	145	8.43	1153	156	7.81	1136	147
19	8.57	1227	155	8.59	1217	163	8.59	1224	138
20	7.96	1217	128	8.64	1226	154	8.62	1222	131
21	7.57	1181	135	8.99	1207	153	8.52	1193	140
22	7.99	1262	148	8.27	1256	126	9.31	1264	148
23	8.74	1372	157	9.49	1324	142	8.07	1316	135
24	8.91	1444	150	8.16	1441	141	8.59	1417	166
25	8.42	1227	137	9.70	1209	153	8.74	1192	140
26	8.48	1196	151	8.64	1214	158	8.46	1191	146
27	8.88	1442	142	8.62	1445	182	10.20	1463	169
28	8.05	1118	138	7.91	1137	147	8.37	1126	147
29	9.08	1126	124	7.61	1118	132	7.88	1136	123
30	8.86	1213	166	8.37	1206	167	9.43	1191	150
31	10.25	1357	154	9.21	1354	166	8.76	1327	154
32	9.59	1181	163	8.96	1190	154	8.62	1187	173
33	8.70	1249	140	8.12	1252	132	8.41	1278	158
34	8.80	1196	149	8.89	1229	160	8.73	1201	146
35	8.85	1188	135	7.79	1193	133	9.16	1170	124
36	8.09	1230	141	8.26	1210	165	8.47	1204	137
37	8.24	1263	140	8.36	1234	137	8.15	1232	144
38	8.89	1395	152	8.19	1441	147	8.00	1393	153
39	8.99	1480	194	8.48	1498	183	7.99	1481	173
40	9.68	1167	144	8.09	1160	153	8.95	1151	152
41	10.03	1235	168	8.29	1224	155	8.16	1252	164
42	8.24	1173	126	8.07	1183	129	8.41	1178	138
43	9.51	1371	157	9.24	1376	156	8.91	1360	166
44	8.27	1110	136	8.44	1124	145	8.87	1146	126
45	7.80	1091	128	7.89	1095	147	7.95	1094	130
46	8.48	1310	156	8.78	1311	139	8.20	1302	135
47	8.79	1320	145	9.09	1305	156	9.14	1300	164
48	8.71	1236	137	8.85	1245	136	8.58	1240	136
49	8.21	1156	132	7.74	1177	130	8.15	1187	139
50	8.24	1213	152	8.76	1205	129	8.55	1211	139
51	8.78	887	139	8.52	889	137	8.18	927	152
52	8.77	882	170	8.85	878	164	8.44	873	173
53	8.70	1200	134	8.31	1178	148	7.68	1161	145
54	8.32	1325	156	7.98	1362	149	8.31	1325	149
55	8.68	1348	158	8.32	1358	176	8.83	1353	157

1.4.5 Hindi

Requirement-level satisfaction for Hindi under Gemini 3 Flash Preview.

Hindi tasks 01–10					Hindi tasks 11–20				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
01	R0	UNSAT.	UNSAT.	UNSAT.	11	R0	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
02	R4	UNSAT.	UNSAT.	UNSAT.	12	R4	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
03	R3	SAT.	SAT.	SAT.	13	R2	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	UNSAT.	SAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
04	R2	UNSAT.	UNSAT.	UNSAT.	14	R6	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R2	UNSAT.	SAT.	UNSAT.
05	R1	SAT.	SAT.	SAT.	15	R3	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R4	SAT.	SAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R5	SAT.	SAT.	UNSAT.
	R4	UNSAT.	SAT.	UNSAT.		R6	SAT.	SAT.	SAT.
06	R0	UNSAT.	UNSAT.	UNSAT.	16	R0	SAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	SAT.		R2	UNSAT.	SAT.	UNSAT.
	R3	SAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.
07	R4	UNSAT.	UNSAT.	UNSAT.	17	R4	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R6	SAT.	SAT.	SAT.
	R2	UNSAT.	SAT.	SAT.		R0	SAT.	SAT.	UNSAT.
08	R3	SAT.	SAT.	SAT.	18	R1	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R0	UNSAT.	UNSAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
09	R2	UNSAT.	UNSAT.	UNSAT.	19	R5	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R0	UNSAT.	SAT.	SAT.
	R5	UNSAT.	SAT.	UNSAT.		R1	UNSAT.	UNSAT.	UNSAT.
10	R0	SAT.	SAT.	SAT.	20	R2	SAT.	SAT.	SAT.
	R1	SAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
						R1	UNSAT.	UNSAT.	UNSAT.
						R2	UNSAT.	UNSAT.	UNSAT.
						R3	UNSAT.	UNSAT.	UNSAT.
						R4	UNSAT.	UNSAT.	UNSAT.
						R5	SAT.	SAT.	SAT.
						R6	SAT.	UNSAT.	UNSAT.

Hindi tasks 21–30					Hindi tasks 31–40				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
21	R0	SAT.	SAT.	SAT.	31	R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R5	UNSAT.	UNSAT.	UNSAT.
22	R6	SAT.	SAT.	SAT.	32	R0	SAT.	UNSAT.	SAT.
	R0	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
23	R5	SAT.	SAT.	SAT.	33	R6	UNSAT.	UNSAT.	UNSAT.
	R6	SAT.	SAT.	SAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	UNSAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	SAT.	SAT.	SAT.
24	R4	UNSAT.	UNSAT.	UNSAT.	34	R5	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R0	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	UNSAT.	UNSAT.
25	R3	UNSAT.	UNSAT.	UNSAT.	35	R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	SAT.		R0	UNSAT.	UNSAT.	SAT.
	R0	UNSAT.	SAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
26	R2	UNSAT.	SAT.	SAT.	36	R3	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R4	UNSAT.	UNSAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R6	SAT.	SAT.	SAT.		R0	UNSAT.	SAT.	UNSAT.
	R0	UNSAT.	SAT.	UNSAT.		R1	UNSAT.	SAT.	UNSAT.
27	R1	UNSAT.	SAT.	UNSAT.	37	R2	UNSAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	UNSAT.	SAT.	UNSAT.
	R5	SAT.	SAT.	SAT.		R6	SAT.	SAT.	SAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
28	R0	UNSAT.	UNSAT.	UNSAT.	38	R0	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R3	UNSAT.	UNSAT.	SAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R5	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
29	R6	UNSAT.	UNSAT.	UNSAT.	39	R6	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	UNSAT.		R7	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	SAT.	SAT.
	R3	SAT.	UNSAT.	UNSAT.		R2	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
30	R5	SAT.	SAT.	SAT.	40	R4	UNSAT.	UNSAT.	UNSAT.
	R6	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	SAT.	SAT.
	R0	SAT.	UNSAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R0	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R2	SAT.	UNSAT.	UNSAT.
31	R4	SAT.	SAT.	SAT.	41	R3	SAT.	SAT.	SAT.
	R5	SAT.	UNSAT.	SAT.		R4	SAT.	SAT.	SAT.
	R6	UNSAT.	SAT.	SAT.		R5	SAT.	SAT.	SAT.
	R0	SAT.	SAT.	SAT.		R6	UNSAT.	UNSAT.	UNSAT.
	R1	UNSAT.	UNSAT.	UNSAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					

Hindi tasks 41–50					Hindi tasks 51–55				
Task	Req	MetaGPT	GPT-Pilot	OpenHands	Task	Req	MetaGPT	GPT-Pilot	OpenHands
41	R0	SAT.	SAT.	SAT.	51	R0	SAT.	SAT.	UNSAT.
	R1	SAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	SAT.	UNSAT.	52	R0	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R1	SAT.	SAT.	SAT.
	R5	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
42	R6	UNSAT.	UNSAT.	UNSAT.	53	R3	SAT.	SAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R0	SAT.	SAT.	SAT.
	R1	SAT.	SAT.	SAT.		R1	SAT.	UNSAT.	SAT.
	R2	SAT.	SAT.	SAT.		R2	SAT.	SAT.	SAT.
	R3	UNSAT.	UNSAT.	UNSAT.		R3	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	UNSAT.		R4	UNSAT.	SAT.	UNSAT.
43	R5	UNSAT.	UNSAT.	UNSAT.	54	R0	UNSAT.	UNSAT.	UNSAT.
	R0	SAT.	SAT.	SAT.		R1	UNSAT.	UNSAT.	UNSAT.
	R1	SAT.	UNSAT.	UNSAT.		R2	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R3	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R5	SAT.	UNSAT.	UNSAT.
44	R5	SAT.	SAT.	SAT.	55	R6	SAT.	SAT.	SAT.
	R0	UNSAT.	SAT.	SAT.		R0	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R1	SAT.	UNSAT.	UNSAT.
	R2	SAT.	SAT.	SAT.		R2	SAT.	SAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R3	SAT.	SAT.	SAT.
	R4	SAT.	SAT.	SAT.		R4	SAT.	SAT.	SAT.
45	R5	SAT.	UNSAT.	SAT.		R5	SAT.	SAT.	SAT.
	R0	UNSAT.	SAT.	UNSAT.		R6	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R7	UNSAT.	UNSAT.	UNSAT.
	R2	UNSAT.	UNSAT.	UNSAT.		R8	UNSAT.	UNSAT.	UNSAT.
	R3	SAT.	SAT.	SAT.		R9	SAT.	SAT.	UNSAT.
	R4	UNSAT.	UNSAT.	UNSAT.		R10	UNSAT.	UNSAT.	UNSAT.
46	R5	UNSAT.	UNSAT.	UNSAT.		R11	SAT.	SAT.	SAT.
	R0	UNSAT.	SAT.	UNSAT.		R12	SAT.	SAT.	SAT.
	R1	UNSAT.	UNSAT.	UNSAT.		R13	SAT.	SAT.	SAT.
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	SAT.	SAT.	SAT.					
	R4	SAT.	SAT.	SAT.					
47	R5	SAT.	SAT.	SAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R0	SAT.	SAT.	SAT.					
	R1	UNSAT.	SAT.	SAT.					
	R2	UNSAT.	SAT.	UNSAT.					
	R3	SAT.	SAT.	SAT.					
48	R4	UNSAT.	UNSAT.	UNSAT.					
	R5	SAT.	SAT.	SAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					
	R0	UNSAT.	UNSAT.	UNSAT.					
	R1	SAT.	UNSAT.	UNSAT.					
49	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	SAT.	SAT.	SAT.					
	R6	SAT.	SAT.	SAT.					
	R0	UNSAT.	SAT.	UNSAT.					
50	R1	UNSAT.	UNSAT.	SAT.					
	R2	UNSAT.	UNSAT.	UNSAT.					
	R3	UNSAT.	UNSAT.	UNSAT.					
	R4	SAT.	SAT.	SAT.					
	R5	UNSAT.	UNSAT.	UNSAT.					
	R6	UNSAT.	UNSAT.	UNSAT.					
	R7	UNSAT.	UNSAT.	UNSAT.					

Table 20: Gemini 3 Flash Preview / Hindi: Task-level average judge metrics across requirements.

Task	MetaGPT			GPT-Pilot			OpenHands		
	Time	Input	Output Tokens	Time	Input	Output Tokens	Time	Input	Output Tokens
01	8.14	1426	112	8.24	1432	105	7.50	1426	99
02	9.27	1242	130	9.12	1226	123	9.07	1197	123
03	8.54	1249	133	8.18	1248	119	8.81	1244	121
04	7.45	814	91	7.75	801	104	6.90	804	101
05	8.66	1240	118	8.70	1238	111	8.33	1248	132
06	7.42	1186	126	9.00	1209	112	7.58	1182	115
07	7.77	1117	117	8.19	1120	129	8.27	1119	129
08	8.44	1146	114	8.00	1151	117	8.02	1156	122
09	7.57	1142	106	7.26	1171	109	7.23	1164	102
10	8.26	1211	125	7.88	1206	123	8.26	1201	134
11	7.99	1225	122	8.83	1225	118	9.55	1206	114
12	8.03	1236	118	8.13	1245	111	8.00	1228	108
13	8.14	1157	117	8.46	1137	118	8.20	1177	108
14	7.61	1193	115	7.83	1198	120	7.92	1213	105
15	7.96	1213	105	8.07	1248	119	8.79	1241	107
16	8.09	1185	118	8.02	1197	113	8.23	1199	112
17	7.94	1273	114	8.41	1303	122	7.85	1295	117
18	8.30	1178	100	8.04	1168	110	7.98	1150	103
19	7.78	1214	112	8.75	1206	114	8.14	1213	118
20	8.65	1194	115	8.09	1192	109	8.59	1185	114
21	8.21	1183	132	8.24	1180	120	7.48	1183	121
22	7.98	1243	124	7.94	1264	114	8.79	1261	120
23	8.22	1323	114	8.56	1318	120	7.44	1329	109
24	8.75	1418	131	7.61	1412	137	8.02	1395	133
25	8.16	1227	115	8.01	1223	108	7.54	1229	120
26	9.04	1188	112	7.84	1166	107	8.58	1185	110
27	8.31	1424	122	9.30	1430	122	8.32	1427	113
28	7.89	1098	119	7.87	1111	120	7.66	1119	113
29	8.64	1134	127	7.61	1143	116	7.77	1133	121
30	8.95	1197	115	8.06	1206	113	7.85	1204	128
31	8.59	1331	124	8.03	1371	120	7.92	1329	116
32	8.86	1186	112	7.93	1153	125	7.60	1185	117
33	7.70	1241	124	7.60	1240	127	8.92	1216	122
34	8.17	1208	115	8.89	1228	126	8.10	1228	118
35	7.73	1138	107	7.30	1160	109	7.70	1173	114
36	7.46	1190	113	7.81	1213	114	7.18	1203	112
37	7.60	1232	108	8.33	1243	121	6.80	1208	113
38	8.11	1391	111	8.53	1403	113	7.53	1403	108
39	8.21	1428	134	8.40	1444	127	8.20	1426	110
40	8.28	1150	126	7.75	1150	126	7.95	1147	120
41	8.19	1210	114	7.92	1230	126	8.33	1222	112
42	8.11	1163	113	7.64	1201	118	8.39	1180	114
43	8.35	1360	125	8.04	1376	134	7.48	1376	120
44	8.28	1132	126	7.33	1131	126	8.11	1132	116
45	8.18	1086	107	7.71	1114	117	8.00	1100	107
46	8.29	1298	123	7.30	1289	111	7.84	1286	120
47	8.23	1274	125	7.35	1296	124	8.01	1297	111
48	8.39	1220	100	7.36	1232	107	8.45	1222	109
49	8.79	1136	118	7.56	1144	114	8.20	1118	114
50	7.98	1209	114	7.99	1206	111	7.89	1203	120
51	9.07	962	110	8.05	906	109	8.19	902	111
52	7.72	884	132	7.83	939	145	9.77	883	142
53	7.72	1213	103	7.29	1219	110	8.23	1213	92
54	9.30	1384	120	7.85	1385	112	10.79	1366	113
55	8.36	1351	132	8.30	1362	138	8.13	1354	133