Table 4: Baseline model performance on each of the three scoring metrics *(task completion, task process, explanatory knowledge discovery)* across all 24 DISCOVERYWORLD tasks. Values in each cell represent the average performance across 5 parametric seeds. *Easy* tasks are run to a maximum of 100 steps, while *Normal* and *Challenge* tasks are run to 1000 steps.

			ReACT			Plan+Execute			Hypothesizer		
#	Topic	Task	Procedure	Completion	Knowledge	Procedure	Completion	Knowledge	Procedure	Completion	Knowledge
Proteomics		Clustering									
1	Easy	Simplified Clustering	0.87	0.20	0.20	0.80	0.00	0.00	0.90	0.40	1.00
2	Normal	Clustering (2D)	0.88	0.40	0.40	0.68	0.20	0.00	0.93	0.40	0.40
3	Challenge	Clustering (3D)	0.88	0.40	0.60	0.58	0.20	0.00	0.93	0.40	0.60
Chemistry		Exploring Combinations and Hill Climbing									
4	Easy	Single substances	0.87	1.00	1.00	0.70	0.60	0.40	0.90	0.00	0.40
5	Normal	Mix of 3 substances	0.82	0.00	0.00	0.87	0.40	0.00	0.93	0.60	0.40
6	Challenge	Mix of 4 substances	0.90	0.40	0.00	0.90	0.40	0.00	0.87	0.00	0.00
Archaeology		Correlations									
7	Easy	Simple instrument	0.27	0.60	0.00	0.33	0.20	0.00	0.60	0.20	0.50
8	Normal	Instrument Use	0.72	0.40	0.30	0.74	0.00	0.00	0.64	0.40	0.40
9	Challenge	Correlation	0.46	0.20	0.00	0.46	0.00	0.05	0.55	0.20	0.05
Reactor Lab		Regression									
10	Easy	Slope only	0.42	0.00	0.40	0.44	0.00	0.10	0.38	0.00	0.20
11	Normal	Linear regression	0.44	0.00	0.20	0.49	0.00	0.00	0.51	0.00	0.00
12	Challenge	Quadratic regression	0.43	0.00	0.20	0.39	0.00	0.00	0.39	0.00	0.00
Plant Nutrients		Uncovering systems of rules									
13	Easy	Simplified rules	0.80	0.20	0.20	0.70	0.20	0.20	0.60	0.00	0.00
14	Normal	Presence rules	0.91	0.60	0.00	0.84	0.40	0.00	0.56	0.00	0.00
15	Challenge	Logical Rules	0.89	0.40	0.00	0.73	0.40	0.00	0.62	0.00	0.00
Space Sick		Open-ended discovery									
16	Easy	Single instrument	0.78	0.60	0.00	0.68	0.40	0.10	0.80	1.00	0.60
17	Normal	Multiple instruments	0.58	0.00	0.13	0.45	0.00	0.13	0.16	0.00	0.33
18	Challenge	Novel instruments	0.55	0.00	0.00	0.26	0.00	0.00	0.20	0.00	0.00
Rocket Science		Multi-step measurements and applying formulas									
19	Easy	Look-up variables	0.33	0.00	0.00	0.53	0.00	0.07	0.13	0.40	0.00
20	Normal	Measure 2 variables	0.51	0.00	0.05	0.34	0.00	0.00	0.11	0.00	0.00
21	Challenge	Measure 5 variables	0.43	0.00	0.00	0.15	0.00	0.00	0.22	0.00	0.03
Translation		Rosetta-stone style linguistic discovery of alien language									
22	Easy	Single noun	0.40	0.40	0.20	0.30	0.00	0.00	0.20	0.20	0.00
23	Normal	Noun and verb	0.20	0.00	0.00	0.68	0.40	0.00	0.84	0.40	0.00
24	Challenge	Noun, adj., and verb	0.49	0.00	0.00	0.55	0.20	0.05	0.15	0.00	0.00
Average (Easy)			0.59	0.38	0.25	0.56	0.18	0.11	0.56	0.28	0.34
Average (Normal)			0.63	0.18	0.14	0.64	0.18	0.02	0.58	0.23	0.19
Average (Challenge)			0.63	0.18	0.10	0.50	0.15	0.01	0.49	0.08	0.08

Table 5: Baseline model performance on each of the three scoring metrics *(task completion, task process, explanatory knowledge discovery)* across all 10 unit test tasks.Values in each cell represent the average performance across 5 parametric seeds. Unit tests tasks are run to a maximum of 100 steps.

		ReACT		Plan+Execute		Hypothesizer	
#	Unit Test Topic	Procedure	Completion	Procedure	Completion	Procedure	Completion
25	Multi-turn dialog with an agent	1.00	1.00	1.00	1.00	1.00	1.00
26	Measure an object with an instrument	0.87	0.60	0.73	0.40	1.00	1.00
27	Pick-and-place object	0.90	0.80	0.80	0.60	1.00	1.00
28	Pick-and-give object	1.00	1.00	1.00	1.00	1.00	1.00
29	Read DiscoveryFeed posts	1.00	1.00	0.90	0.80	1.00	1.00
30	Move through doors	0.58	0.20	0.25	0.00	0.30	0.00
31	Using keys with doors	0.69	0.20	0.54	0.00	0.69	0.00
32	Navigate to a specific room in a house	0.20	0.20	0.20	0.00	0.20	0.20
33	Search an environment for an object	0.80	0.80	0.60	0.60	1.00	1.00
34	Interact with a moving agent	0.60	0.20	0.53	0.00	0.53	0.20
Average (Unit Tests)		0.76	0.60	0.66	0.44	0.77	0.64

4.2 Baseline Agent Models

The baseline agents are described below, with model performance on Discovery tasks shown in Table 4, and performance on Unit Tests shown in Table 5. We use the GPT-4O model for all our agents due to its higher performance and lower cost compared to other models. For space we provide