APPENDIX

А

871 872 A.1 WECHSLER INTELLIGENCE TEST

BENCHMARK INFORMATION

Our study focuses on the Wechsler Intelligence Scale, specifically the Wechsler Intelligence Scale 873 for Children (WISC), an individually administered test designed for children aged 6 to 16. Originally 874 developed by the American psychologist Dr. David Wechsler, the test undergoes periodic updates 875 and revisions every few decades to maintain its accuracy and relevance. The latest and most widely 876 used version is WISC-V, which measures five primary cognitive domains: Verbal Comprehension 877 Index (VCI), Visual Spatial Index (VSI), Fluid Reasoning Index (FRI), Working Memory Index 878 (WMI), and Processing Speed Index (PSI). 879

MLLMs are still in the early stages of development. While assessments for adults typically em-880 phasize specialized knowledge or skills, children's tests focus more on fundamental cognitive abil-881 ities that underpin general intelligence. By evaluating MLLMs against these foundational abilities, 882 we can more effectively gauge their adaptability and developmental potential, mirroring the stages 883 of human cognitive growth. This approach allows us to gain a comprehensive understanding of 884 MLLMs' strengths and limitations, providing clear guidance for improvements and ultimately con-885 tributing to the advancement of AGI.

- 886 887

A.2 CAPABILITY DESIGN

889 890 891

893

Wechsler believed that intelligence could be viewed as a combination of various components or abil-892 ities that differ in quantity. Here's an explanation of how each capability discussed in the Section 3 connects to the specific cognitive domains measured by the Wechsler Intelligence Test.

894 Execution in MLLMs is connected to the Processing Speed Index (PSI) of the WISC, which assesses 895 the speed and accuracy with which a child completes simple tasks. However, for MLLMs, speed is 896 not the primary focus; we place greater emphasis on accuracy. Therefore, it signifies the model's 897 ability to perform whether simple or complex tasks aligned with specific goals, ensuring that the 898 outcomes are correct and precise. 899

Perception Reasoning in MLLMs aligns with the Visual Spatial Index (VSI) and Fluid Reasoning 900 Index (FRI) of the WISC. While VSI assesses a child's ability to interpret and organize visual infor-901 mation, and FRI evaluates problem-solving and abstract thinking, Perception Reasoning in MLLMs 902 focuses on interpreting input data and making logical inferences, which allows the model to under-903 stand context and relationships. 904

Memory in MLLMs corresponds to the Working Memory Index (WMI) of the WISC, which mea-905 sures a child's capacity to temporarily hold and manipulate information. For MLLMs, Memory 906 involves retaining contextual information over interactions and recalling relevant data when needed. 907 This ensures coherent and contextually appropriate responses. 908

Learning in MLLMs is underpinned by the Verbal Comprehension Index (VCI) from the WISC, 909 which assesses a child's ability to understand and process verbal information. VCI serves as a 910 foundational prerequisite for Learning, as it enables the comprehension necessary for acquiring new 911 knowledge. In designing MLLM capabilities, we considered VCI to ensure that the models have the 912 linguistic understanding required to effectively learn from and adapt to new data. 913

914 **Planning** in MLLMs functions similarly to executive processes—cognitive abilities that, although 915 not explicitly isolated in the WISC, are essential for complex tasks. It involves strategizing and sequencing actions to achieve specific objectives while considering potential outcomes. This capacity 916 enables MLLMs to generate coherent, goal-oriented responses or actions, reflecting higher-order 917 thinking skills akin to those that enhance children's performance across various WISC subtests.

В **EXPERIMENT DETAILS**

B.1 RESOLUTION SETUP

The game environment consists of a 9×9 grid, where each cell is rendered at a resolution of 64×64 pixels, resulting in a total resolution of 576×576 pixels. Within this environment, the central 8×7 grid area is artistically designed as a background, while a 5×5 square serves as the active area where the agent can move and interact. The remaining spaces mainly function as decorative elements to provide semantic context. The bottom 8×1 grid displays the agent's backpack, while the left 9×2 grid shows hints and task-related information. All item and backpack labels, are rendered at 16×16 pixels, ensuring clarity and efficient use of space.

B.2 MODEL PARAMETERS

In this study, we set the temperature parameter of all language models to 0 for all experimental tasks. By doing so, we enforced deterministic behavior, ensuring that the models' outputs were exclusively determined by their learned probability distributions. This configuration minimizes the influence of stochasticity and provides a controlled environment for evaluating model performance.

B.3 RANDOM BASELINE

To provide a benchmark for comparison, we established a random baseline in which actions were selected entirely at random. The results of the random baseline were obtained either through analyt-ical computation or experimental estimation, depending on the complexity of the task. For simpler tasks, such as SE, DE, FI, PU, DE*, SO, PI* and PL, probabilistic methods were used to compute the expected performance of random actions. For more complex tasks, such as CL, MA, MA* and CO, where analytical solutions are impractical, the random baseline was approximated by running 500 iterations of random experiments. This methodology allows the random baseline to serve as a meaningful point of reference across a diverse range of tasks.

- B.4 RAW DATA

Table 4: Comparison of raw success rate of different MLLMs on Level1 test of Classification, Selection, Decode, Maze, Filling, Puzzle.

Level1	CL	SE	DE	MA	FI	PU
Human GPT-40 Gemini Qwen2 Internvl DeepSee	1.00 0.88 0.97 0.70 0.61	1.00 0.48 0.26 0.41 0.32 <u>0.24</u>	1.00 0.72 0.52 0.34 0.26 0.35	1.00 1.00 0.99 <u>0.71</u> 0.90 0.89	0.94 0.49 0.42 0.47 0.36 0.39	1.00 0.26 0.24 0.25 <u>0.22</u> 0.27
Phi3.5 Llava	0.42 <u>0.37</u>	0.26 0.25	$\frac{0.25}{0.25}$	0.83 0.88	0.36 <u>0.32</u>	0.25 0.25
InternLN Random		0.25 0.25	<u>0.25</u> 0.25	0.92 ≈0.79	0.42 0.25	0.25 0.25

	Level1	MA*	DE*	SO	FI*	PL	CO
	Human	0.94	1.00	0.83	0.89	0.72	1.00
	GPT-40	0.93	0.95	0.58	0.73	0.06	0.51
	Gemini	0.96	0.80	0.47	0.46	0.16	<u>0.42</u>
	Qwen2	0.10	0.38	0.72	0.35	0.14	0.50
	Internvl	0.43	0.27	0.96	0.28	0.11	0.49
	DeepSeek	0.16	0.26	0.57	0.24	0.19	<u>0.42</u>
	Phi3.5	0.17	0.26	0.56	0.36	0.12	0.43
	Llava	0.18	0.24	0.42	0.27	0.12	0.42
	InternLM	0.18	0.26	0.45	0.17	0.17	0.44
	Random	≈ 0.08	0.25	0.50	0.25	0.13	≈ 0.43
Table 6: Compari	son of raw	uccess r	ate of di	fferent	MIIM	[s on I i	ovel? tec
Selection, Decode,				nerent			
Selection, Decode,	, 111120, 1 11111	5, 1 uzzie	•				
	Level2	CL	SE	DE	MA	FI	PU
-	Level2	CL	SE	DE	MA	ГІ	FU
	Human	1.00	1.00	1.00	1.00	0.94	4 1.00
	GPT-40	0.67	0.47	0.72	0.91		
	Gemini	0.63	0.16	0.45	0.64		
	Qwen2	0.27	0.10	0.21	0.23		
	Internvl	0.18	0.10	0.19	0.62		
	DeepSeek	0.14	0.06	0.20	0.34		
	Phi3.5	0.16	0.09	0.18	0.50	0.15	5 0.08
	Llava	0.21	0.07	0.17	0.35		6 0.08
	Intown I M	0.34	<u>0.06</u>	<u>0.16</u>	0.74	0.10	5 <u>0.07</u>
	InternLM						
	Random	≈ 0.36	≈ 0.07	0.17	≈ 0.6	7 0.08	3 0.08
			≈ 0.07	0.17		7 0.08	3 0.08
			≈0.07	0.17		7 0.08	3 0.08
			≈0.07	0.17		7 0.08	3 0.08
			≈0.07	0.17		7 0.08	3 0.08
			≈0.07	0.17		7 0.08	3 0.08
Table 7: Comparis	Random	≈0.36			≈0.6		
Table 7: Comparis	Random son of raw su	≈0.36			≈0.6		
Table 7: Comparis Sorting, Filling*, F	Random son of raw su	≈0.36			≈0.6		
	Random son of raw su Placement, Co	≈0.36	e of diffe	erent M	≈0.6	on Leve	12 test of
	Random son of raw su	≈0.36			≈0.6		
	Random son of raw su Placement, Co Level2	≈0.36 cccess rate punting. MA*	e of diffe	erent M	≈0.6 ILLMs FI*	on Leve PL	12 test of
	Random son of raw su Placement, Co Level2 Human	≈ 0.36 excess rate pointing. MA^* 0.83	e of diffe DE* 1.00	erent M <u>SO</u> 0.72	≈0.6 ILLMs <u>FI*</u> 0.89	on Leve <u>PL</u> 0.67	12 test of <u>CO</u> 1.00
	Random son of raw su Placement, Co Level2 Human GPT-40	≈ 0.36 cccess rate ounting. MA* 0.83 0.65	e of diffe DE* 1.00 0.98	erent M SO 0.72 0.13	≈0.6 LLMs <u>FI*</u> 0.89 0.27	on Leve PL 0.67 0.17	12 test of <u>CO</u> 1.00 0.48
	Random son of raw su Placement, Co Level2 Human GPT-40 Gemini	≈0.36 cccess rate ounting. <u>MA*</u> 0.83 0.65 0.55	e of diffe DE* 1.00 0.98 0.83	erent M SO 0.72 0.13 <u>0.02</u>	≈0.6 LLMs FI* 0.89 0.27 0.18	on Leve PL 0.67 0.17 0.10	12 test of <u>CO</u> 1.00 0.48 0.52
	Random son of raw su Placement, Co Level2 Human GPT-40 Gemini Qwen2	≈0.36 cccess rate ounting. <u>MA*</u> 0.83 0.65 0.55 0.05	e of diffe DE* 1.00 0.98 0.83 0.29	erent M SO 0.72 0.13 <u>0.02</u> 0.26	≈0.6 ELLMs FI* 0.89 0.27 0.18 0.16	on Leve PL 0.67 0.17 0.10 0.20	12 test of <u>CO</u> 1.00 0.48 0.52 0.54
	Random Son of raw su Placement, Co Level2 Human GPT-40 Gemini Qwen2 Internvl	≈0.36 cccess rate ounting. <u>MA*</u> 0.83 0.65 0.55 0.05 0.24	e of diffe DE* 1.00 0.98 0.83 0.29 0.17	erent M SO 0.72 0.13 <u>0.02</u> 0.26 0.39	≈0.6 ELLMs FI* 0.89 0.27 0.18 0.16 0.11	PL 0.67 0.17 0.10 0.20 0.08	12 test of <u>CO</u> 1.00 0.48 0.52 0.54 0.45
	Random Son of raw su Placement, Co Level2 Human GPT-40 Gemini Qwen2 Internvl DeepSeek	≈0.36 cccess rate ounting. <u>MA*</u> 0.83 0.65 0.55 0.05 0.24 0.04	e of diffe DE* 1.00 0.98 0.83 0.29 0.17 0.17	SO 0.72 0.13 0.26 0.39 0.27	≈0.6 ELLMs FI* 0.89 0.27 0.18 0.16 0.11 0.16	PL 0.67 0.17 0.10 0.20 <u>0.08</u> 0.12	12 test of <u>CO</u> 1.00 0.48 0.52 0.54 0.45 0.45
	Random Son of raw su Placement, Co Level2 Human GPT-40 Gemini Qwen2 Internvl DeepSeek Phi3.5	≈0.36 cccess rate ounting. MA* 0.83 0.65 0.55 0.05 0.24 0.04 0.14	e of diffe DE* 1.00 0.98 0.83 0.29 0.17 0.17 0.20	SO 0.72 0.13 0.26 0.27 0.27 0.34	≈0.6 FI* 0.89 0.27 0.18 0.16 0.11 0.16 0.13	PL 0.67 0.17 0.10 0.20 <u>0.08</u> 0.12 0.09	12 test of <u>CO</u> 1.00 0.48 0.52 0.54 0.45 0.45 0.47
	Random Son of raw su Placement, Co Level2 Human GPT-40 Gemini Qwen2 Internvl DeepSeek	≈0.36 cccess rate ounting. <u>MA*</u> 0.83 0.65 0.55 0.05 0.24 0.04	e of diffe DE* 1.00 0.98 0.83 0.29 0.17 0.17	SO 0.72 0.13 0.26 0.39 0.27	≈0.6 ELLMs FI* 0.89 0.27 0.18 0.16 0.11 0.16	PL 0.67 0.17 0.10 0.20 <u>0.08</u> 0.12	12 test of <u>CO</u> 1.00 0.48 0.52 0.54 0.45 0.45

Table 5: Comparison of raw success rate of different MLLMs on Level1 test of Maze*, Decode*,
Sorting, Filling*, Placement, Counting.

 ≈ 0.10 0.17 0.17 0.08 0.13 ≈ 0.48

Random

1025

		Level3	CL	SE	DE	M	A	FI	PU
	-								
		Human	0.83	0.89	1.00	1.0	0	0.89	0.94
		GPT-40	0.58	0.38	0.59			0.07	0.04
		Gemini	0.41	0.05	0.35			0.11	0.04
		Qwen2	0.16	0.02	0.16			0.08	0.03
		Internvl	0.04	$\frac{0.00}{0.02}$	0.13			0.05	$\frac{0.02}{0.02}$
		DeepSeek	$\frac{0.00}{0.11}$	0.02	0.14			0.06	0.0
		Phi3.5	0.11	0.02	0.14			0.08	0.04
		Llava InternLM	0.09 0.15	$\begin{array}{c} 0.01 \\ 0.01 \end{array}$	$\frac{0.12}{0.12}$			0.09 0.04	0.05 0.0
			≈ 0.13	≈ 0.02		≈ 0.4		0.04	0.04
		Random	~ 0.25	~ 0.02	0.15	~ 0	.05	0.0-	0.0
Tabl	e 9: Compari	son of raw su	ccess rat	te of diff	erent M	ILLMs	on Le	evel3	test c
		Placement, Co					D		
-	U, 0, 1	, -	6						
		Level3	MA*	DE*	SO	FI*	PL	C	CO
			10111		50		12		
		Human	0.72	0.94	0.67	0.83	0.67	1.	.00
		GPT-40	0.63	0.95	0.08	0.14	0.17		.48
		Gemini	0.39	0.83	<u>0.04</u>	0.09	0.10		.52
		Qwen2	0.08	0.24	0.08	0.08	0.20		.54
		Internvl	0.17	0.15	0.13	$\underline{0.05}$	0.08		.45
		DeepSeek	$\frac{0.04}{0.12}$	$\frac{0.13}{0.14}$	0.07	0.06	0.12		.45
		Phi3.5	0.13	0.14	0.17	0.06	0.09		.47
		Llava InternLM	$\frac{0.04}{0.16}$	$\frac{0.13}{0.13}$	$0.05 \\ 0.04$	0.09 0.07	0.12 0.11		.42 .50
		Random	0.13	$\frac{0.13}{0.13}$	0.04	0.04	0.13).48
С	Evaluati	on Proce	DURE						
C.1	CODE STRU	UCTURE							
N 7.4	f f a f	f 11'		A		£	0		
		user-friendline th pip. The ov							
Ja311	y mistaneu wi	in pip. The O			WOIK IU	i Uvaiu	aung	•1 	1 15
lgo	orithm 1 Mod	lel Evaluation							
Req		m environme		e task, m	odel: N	1LLM	mode	, hist	tory:
1: 2:		GeneratePro	mnt(an	<i>a</i> ,)					
2: 3:		$Generater for history \leftarrow C$			nnt hie	tory			
3. 4:		ProcessAnsw			.p., 1118	<i>cory</i>)			
ч. 5:	env.step(a		••• (respe						
6:		s over or reac	h the ma	ximum r	umber	of step	s then	1	
7:		while loop				r			
8:	end if								
0.	and while								

1026 Table 8: Comparison of raw success rate of different MLLMs on Level3 test of Classification, Selection, Decode, Maze, Filling, Puzzle. 1027 1028

9: end while

1080 1081	C.2 PROMPT DESIGN
1082 1083	To ensure fairness across all MLLMs, we conducted experiments using exactly the same prompts that included both instructions on the game rules and defined goals for each task.
1084 1085	Game Rule
1085 1086 1087 1088 1089	You are currently playing as a character in a 2D game, as depicted in the image. The game rules are as follows: Items in the game scene are all identified by numerical labels, such as 0, 1, 2, 3, etc. The black squares at the bottom of the screen represent your backpack, labeled as A, B, C, and D,
1090 1091 1092 1093	each capable of holding only one item. You cannot move to or interact with an item if there is anything between you and that item. Before each step, you will be presented with a series of action options and you should select the letter corresponding to the action you believe is the right choice to achieve the goal.
1094	Task Description
1095	In this task, your goal is $\langle GOAL \rangle$.
1096 1097 1098	Now, game starts! What is the first action you will choose? / What is the next action you will choose? The actions you can choose from are: $\langle ACTIONS \rangle$.
1099 1100	Invalid Answer
1100 1101 1102	Your answer is invaild, please tell me the action letter you choose. (e.g. A)
1103	C.3 ACTION CHOICE
1104 1105	We convert the high-dimensional actions into random options for MLLMs to choose from.
1106	Example
1107 1108	The action list is: ['pick up apple', 'pick up banana', 'pick up orange'].
1109	The action choice will be: A) 'pick up orange', B) 'pick up apple', C) 'pick up banana'.
1110 1111	C.4 ANSWER DECODE
1112 1113 1114 1115 1116	Even though we repeatedly emphasized in the prompts that the MLLM should respond in the speci- fied format of the options, we were still unable to strictly standardize their response formats. There- fore, we collected a large number of responses to analyze and summarize the characteristics of MLLMs when answering our questions. The final decoding method was determined as follows.
1117	Algorithm 2 Process Answer
1118 1119 1120 1121 1122 1123 1124 1125 1126 1127	Require: answer: string, actions: list of strings 1: for each action in actions do 2: if action is in answer then 3: return index of action 4: end if 5: end for 6: match \leftarrow first uppercase letter found in answer 7: if match is not None and the index of match in alphabet is less than length of actions then 8: index \leftarrow index of match in alphabet 9: return index 10: end if
1128 1129	11: return None
1120	

1131 Explanation

First, we iterate through the list of actions and check whether any action appears in the response generated by the MLLM. If a match is found, the index of the matching action is returned immedi-

ately. If no match is found, we then search for the first single uppercase letter, and check whether
its corresponding index falls within the valid range of the action list. If it does, the corresponding
index is returned; otherwise, the response is considered invalid.

- 1137 1138 Example
- 1139 $response: A \rightarrow return A$

1140 response: I choose action letter B) 'pick up item with label 2'. \rightarrow return B

1142 response: Based on all of the information, I choose action $C. \rightarrow return C$

1143 response: I'm sorry, but I can't provide the correct answer as the image does not contain a dog. It 1144 appears to be a game with various animals, but none of them are dogs. \rightarrow return NONE

1145 1146 $response: ...? = = ..n n The l n The l \rightarrow return NONE$

1147 1148

D HUMAN BASELINE

1149 1150

1150 D.1 SETUP

The human experiment included 18 college students who had no prior exposure to our GridAgent environment. The participants were divided into three groups of six, with each group assigned to test one of the three difficulty levels. To minimize the potential influence of repeated attempts and accumulated experience, which could artificially inflate success rates, each participant completed only a single iteration of their assigned tasks. This experimental setup closely mirrored the conditions for the MLLM, which also approached each task as a first-time experience.

1158

1159 D.2 RAW RESULTS

The results of the human test are presented in Table 10. Despite the limited number of human testers (18), the design of the options in this study provides limited freedom for players, resulting in minimal strategy variation among testers. Furthermore, the convergence of strategy choices indicates that the results of this task do not require a large sample size for reliability. Consequently, the sample size of 18 testers is sufficient for establishing a reliable human baseline for comparison.

1166

166													
167		Ta	ble 10:	Accurac	ey rates	for task	s of level	1/2/3 diffi	culty in h	uman tes	ting.		
168 169 170 171	CL1 1.00	CL2 1.00	CL3 0.83	SE1 1.00	SE2 1.00	SE3 0.89	DE1 1.00	DE2 1.00	DE3 1.00	MA1 1.00	MA2 1.00	MA3 1.00	
172 173 174	FI1 0.94	FI2 0.94	FI3 0.89	PU1 1.00	PU2 1.00	PU3 0.94	MA*1 0.94	MA*2 0.83	MA*3 0.72	DE*1 1.00	DE*2 1.00	DE*3 0.94	
75 76 77 78	SO1 0.83	SO2 0.72	SO3 <u>0.67</u>	FI*1 0.89	FI*2 0.89	FI*3 0.83	PL1 0.72	PL2 0.67	PL3 <u>0.44</u>	CO1 1.00	CO2 1.00	CO3 1.00	

1179

1180

From the results, it was observed that some testers made simple mistakes due to carelessness, such as
 misreading the option letters or misinterpreting item and backpack labels. Additionally, distractions
 during the test led to testers forgetting previously memorized information.

Surprisingly, the testers did not perform well on tasks requiring learning (only 0.67 in SO3 and 0.44
in PL3). After discussing with them, we identified that he phrasing of the new rules was somewhat
convoluted, making it harder for testers to quickly and accurately internalize the instructions. Thus,
some testers struggled to fully grasp the new rules introduced in the tasks, leading to confusion during execution.

These observations highlight natural occurrences, as similar errors such as misreading instructions, forgetting prior information or misunderstanding the task rule are also observed in MLLMs. These errors reflect the challenges inherent in processing and applying new information within a limited context, whether for humans or models. Consequently, such factors must be carefully considered during the evaluation process to ensure a fair and comprehensive comparison of performance.

1194 1195 D.3 NORMALIZED SCORE

Given the success rate R_{mlm} of the MLLM, we compare its performance with the human baseline (see in Table 10), and eliminate the impact of absolute values through normalization.

$$S_{\rm mllm} = 1 - \frac{R_{\rm human} - R_{\rm mllm}}{R_{\rm human}} = \frac{R_{\rm mllm}}{R_{\rm human}}$$

E ADDTIONAL EXPERIMENTS

1206 E.1 ITEM RECOGNIZE

Table 11: Experiment 1209 1210 1211 MLLM Recognize 1212 GPT-40 1213 0.86 Gemini 0.91 1214 Qwen2-VL-7b 0.95 1215 Internvl-chat-v1-5 0.86 1216 DeepSeek-v1-7b-chat 0.83 1217 Phi3-5-vision-instruct 0.61 1218 Llava-v1.6-Mistral-7b 0.31 1219 InternLM-Xcomposer2-7b-chat 0.31 1220

1221

1193

1199

1201 1202 1203

1204 1205

1208

1222

1223 In the recognize test, we provide the large language model with four different images, which could 1224 be of animals, fruits, toys, etc. Then we ask the model, "Tell me which one is $\langle OBJ.NAME \rangle$?" 1225 The table above records the accuracy of the responses given by the large language models. From this, 1226 we can see that the accuracy of most large language models exceeds 0.6, with some even reaching 1227 0.8 or 0.9. This suggests that most large language models are capable of recognizing various items 1228 in the game scene.

However, we also noticed that a few large language models performed poorly in this test. For instance, Llava-v1.6-Mistral-7b's accuracy was relatively low. By analyzing the responses from the language model, we found that many of Llava-v1.6-Mistral-7b's answers were considered invalid(see Section C.4), which explains its low accuracy.(InternLM-Xcomposer2-7b-chat is the same.) So generally speaking, most large language models possess a strong ability to recognize objects.

1235

1237

1236 E.2 RESOLUTION TEST

In our initial setup, each grid cell had a resolution of 64x64. To further enhance our analysis, we conducted additional experiments by testing grid cells at resolutions of 0.5x(32*32), 0.75x(48*48), 1.25x(80*80), and 1.5x(96*96) the original size. These supplementary tests were designed to explore the impact of varying grid resolutions on our results, providing deeper insights into how resolution influences performance and outcomes in our experiments.

Resolution	32*32	48*48	64*64	80*80	96*96
GPT-40	0.51	0.46	0.49	0.62	0.48
			0,		
Gemini	0.38	0.42	0.42	0.41	0.44
Qwen2	0.29	0.40	0.47	0.43	0.45
Internvl	0.34	0.31	0.36	0.42	0.34
DeepSeek	0.45	0.43	0.39	0.39	0.45
Phi3.5	0.32	0.31	0.36	0.43	0.32
Llava	0.33	0.37	0.32	0.44	0.38
InternLM	0.32	0.34	0.42	0.37	0.35

Table 12: The results of the success rate of all models in Resolution Test on FI.

E.3 EXAMPLE TEST

> We incorporated a fully correct solution process for the tasks PI, SO, and PL and used it as the initial prompt provided to the MLLMs. This approach was aimed at investigating whether including an accurate solution as a guiding example would influence the model's performance. By doing so, we sought to assess the extent to which the model leverages such examples to improve its problem-solving capabilities and overall response quality.

Table 13: The results of the success rate of all models in Example Test.

Enomalo	F	I	S	С	PL		
Example	Origin	New	Origin	New	Origin	New	
GPT-40	0.49	0.59	0.58	0.76	0.06	0.31	
Gemini	0.42	0.56	0.47	0.69	0.16	0.17	
Qwen2	0.47	0.36↓	0.72	0.91	0.14	0.11↓	
Internvl	0.36	0.41	0.96	0.63↓	0.11	0.17	
DeepSeek	0.39	0.41	0.57	0.98	0.19	0.10↓	
Phi3.5	0.36	0.44	0.56	0.91	0.12	0.10↓	
Llava	0.32	0.46	0.42	0.84	0.12	0.17	
InternLM	0.42	0.36↓	0.45	0.72	0.17	0.11↓	

Example of Filling Task

To help you better understand the game rules, let's walk through a correct example. At the start of the game, you will see the following setup: the target image is displayed on the left side, while the incomplete image is shown inside a frame on the right side. Your goal is to make the framed image on the right identical to the target image by selecting and placing the correct puzzle pieces in the specified positions. In this example, the target image on the left shows a sheep, while the framed image on the right is missing the lower-left corner, specifically the front feet of the sheep. By examining the four available puzzle pieces, we can determine that piece C matches the missing feet. In addition to shape, the color of the piece further confirms that piece C is correct, as its color matches that of the target image, unlike the other options. Thus, the correct choice is: Place piece from backpack C in grid I.

<u>ଜ</u>୍ଞା 😽 🕟 Figure 4: "Example image for Filling Task" **Example of Placement Task** To help you better understand the game rule, let's walk through a correct example. At the

start of the game, you will see the following setup: Around the hamburger, there are eight grids marked with Roman numerals from I to VIII. Then, you will get the rule: 'Place the item in the opposite direction, such as if it requires you to put it on the east side, you need to place it on the west side.' After you understand the rule, it needs you to 'Place the chicken leg on the east side of the hamburger.' Since the new rule tells that you need to put the item in the opposite direction, the opposite direction of east is west. Therefore, you need to place the chicken leg on the west side of the hamburger, that is label VII. Thus, the correct choice is: place chicken leg at grid VII.



(VI) (IV) (V)= 1 (VII)

Example of Sorting Task

To help you better understand the game rules, let's walk through a correct example. At the start of the game, you will see the following setup: There are two animals in your backpack, a mouse in backpack A and a pig in backpack B. Then, you will get the rule: 'In this world, the lighter the animal is, the slower it is.' After understanding the rule, the task requires you to 'Rank the animals in the backpack from slow to fast by speed in position I, II.' According to the rule, 'the lighter the animal is, the slower it is,' you should know that a mouse is lighter than a pig, so the mouse runs slower than the pig. Although this contradicts common sense, we need to follow the new rules. Thus, the slower animal, the mouse in backpack A, should be placed at grid I, and the faster animal, the pig in backpack B, should be placed at grid II. Therefore, the first correct choice is: 'place animal from backpack A at grid I'.

)													
							-						
)							0						
}													
ļ.						unit in the							
)						\bigcirc							
5													
								`					
						e ee e							
							1 Alexandre						
				Fig	ıre 6: "	Exampl	e image	e for Sor	ting Tas	k"			
				8		<u>Littling</u> i	e	Jei Sei					
E.	4 Mo	odel Si	IZE TE	ST									
													eriments
			2-VL-72	2b mod	el and c	compare	ed its re	esults ag	ainst the	e Qwen	2-VL-7	b mode	el on all
Le	evel1 ta	ISKS.											
	Table	14: The	results	of the s	uccess	rate of	two mo	dels in N	Aodel S	ize Test	on all	Level1	tasks.
(Qwen	CL	SE	DE	MA	FI	PU	MA*	DE*	SO	FI*	PL	CO
	72b	0.79	0.24	0.27	0.98	0.52	0.21	0.31	0.38	0.71	0.40	0.06	0.59
	7b	0.70	0.41	0.34	0.71	0.47	0.25	0.10	0.38	0.72	0.35	0.14	0.50
F	CA			TODE									
Г	CA.	PABILI	114.50	JORE									
Tł	ne score	a for eac	rh canal	hility is	calcula	tod by t							
			in capa	onney is	calcula	lied by	the follo	owing ga	ames:				
			in cupu	511109 15	carcula	lied by	the follo	owing ga	ames:				
	•]	Execution	1	2	eareura	lied by	the follo	owing ga	ames:				
	•		1	2	carcuit	lied by	the follo	owing ga	ames:				
			on: All	Games		Ĵ		owing ga	ames:				
	•	Executio Percepti	on: All ion Rea	Games	FI, PU	, FI*, P		owing ga	ames:				
	•	Executio	on: All ion Rea	Games	FI, PU	, FI*, P		owing ga	ames:				
	•]	Executio Percepti Memory	on: All ion Rea y: SE, N	Games soning: MA*, D	FI, PU E*, FI*	, FI*, P		owing ga	ames:				
	•]	Executio Percepti	on: All ion Rea y: SE, N	Games soning: MA*, D	FI, PU E*, FI*	, FI*, P		owing ga	ames:				
	•	Executio Percepti Memory Learnin	on: All ion Rea y: SE, N g: DE,	Games soning: MA*, D DE*, S	FI, PU E*, FI* O, PL	, FI*, P		owing ga	ames:				
	•	Executio Percepti Memory	on: All ion Rea y: SE, N g: DE,	Games soning: MA*, D DE*, S	FI, PU E*, FI* O, PL	, FI*, P		owing ga	ames:				
	•	Executio Percepti Memory Learnin	on: All ion Rea y: SE, N g: DE,	Games soning: MA*, D DE*, S	FI, PU E*, FI* O, PL	, FI*, P		owing ga	ames:				
Fc	•	Execution Perception Memory Learnin Plannin	on: All ion Rea y: SE, M g: DE, g: MA,	Games soning: MA*, D DE*, S MA*, S	FI, PU E*, FI* O, PL SO, CO	, FI*, P	PL, CO			luating	its perf	ormanc	e across
	• • •	Execution Perception Memory Learnin Plannin, MLLM	on: All ion Rea y: SE, M g: DE, g: MA, , we con	Games soning: MA*, D DE*, S MA*, S MA*, S	FI, PU E*, FI* O, PL SO, CO he score	, FI*, P	PL, CO		c by eva	luating	its perf	ormanc	e across
	• • •	Execution Perception Memory Learnin Plannin, MLLM	on: All ion Rea y: SE, M g: DE, g: MA, , we con	Games soning: MA*, D DE*, S MA*, S MA*, S	FI, PU E*, FI* O, PL SO, CO he score	, FI*, P	PL, CO	pability	c by eva	luating	its perf	ormanc	e across
	• • •	Execution Perception Memory Learnin Plannin, MLLM	on: All ion Rea y: SE, M g: DE, g: MA, , we con	Games soning: MA*, D DE*, S MA*, S MA*, S	FI, PU E*, FI* O, PL SO, CO he score	, FI*, P	PL, CO	pability	c by eva	luating	its perf	ormanc	e across
	• • •	Execution Perception Memory Learnin Plannin, MLLM	on: All ion Rea y: SE, M g: DE, g: MA, , we con	Games soning: MA*, D DE*, S MA*, S MA*, S	FI, PU E*, FI* O, PL SO, CO he score	, FI*, P	PL, CO ;iven caj 2 score i	pability s calcula	c by eva	luating	its perf	ormanc	e across
	• • •	Execution Perception Memory Learnin Plannin, MLLM	on: All ion Rea y: SE, M g: DE, g: MA, , we con	Games soning: MA*, D DE*, S MA*, S MA*, S	FI, PU E*, FI* O, PL SO, CO he score	, FI*, P	PL, CO ;iven caj 2 score i	pability s calcula	c by eva	luating	its perf	ormanc	e across
	• • •	Execution Perception Memory Learnin Plannin, MLLM	on: All ion Rea y: SE, M g: DE, g: MA, , we con	Games soning: MA*, D DE*, S MA*, S MA*, S	FI, PU E*, FI* O, PL SO, CO he score	, FI*, P	PL, CO	pability s calcula	c by eva	luating	its perf	ormanc	e across
	• • •	Execution Perception Memory Learnin Plannin, MLLM	on: All ion Rea y: SE, M g: DE, g: MA, , we con	Games soning: MA*, D DE*, S MA*, S MA*, S	FI, PU E*, FI* O, PL SO, CO he score	, FI*, P	PL, CO ;iven caj 2 score i	pability s calcula	c by eva	luating	its perf	ormanc	e across
all	or each tasks t	Execution Perception Memory Learnin Plannin MLLM t_i associ	on: All ion Rea y: SE, N g: DE, g: MA, , we con iated w:	Games soning: MA*, D DE*, S MA*, S MA*, S	FI, PU E*, FI* O, PL SO, CO he score capabil	; FI*, P e for a g ity. The $S_c =$	PL, CO given cap e score i = $\frac{1}{n} \sum_{i=1}^{n}$	pability s calcula R_{t_i}	<i>c</i> by eva ated as:	-	-		
all	or each tasks the tasks the tasks the task structure tasks the task structure task structure tasks the task structure task st	Execution Perception Memory Learnin Plannin MLLM t_i associ	on: All ion Rea y: SE, N g: DE, g: MA, , we con iated w:	Games soning: MA*, D DE*, S MA*, S mpute th ith that	FI, PU E*, FI* O, PL SO, CO he score capabil	f, FI*, P e for a g ity. The $S_c =$ e for ca	PL, CO given cap e score i = $\frac{1}{n} \sum_{i=1}^{n}$	pability s calcula R_{t_i} v c, n is	<i>c</i> by eva ated as:	-	-		e across ted to c,

	Execution	Memory	Learning	Planning	Perception Reasoning
Human Baseline	1.00	1.00	1.00	1.00	1.00
GPT-40	0.70	0.81	0.61	0.80	0.44
Gemini	0.58	0.65	0.53	0.75	0.37
Qwen2	0.45	0.32	0.45	0.55	0.37
Internvl	0.46	0.34	0.46	0.75	0.31
DeepSeek	0.39	0.24	0.39	0.54	0.33
Phi3.5	0.38	0.28	0.34	0.53	0.33
Llava	0.35	0.25	0.29	0.50	0.30
InternLM	0.39	0.22	0.32	0.52	0.31
G TASK INFO	RMATION				
G.1 CLASSIFICAT	TION				
Introduction					
-	gent to place	items into c	lesignated co	ontainers ac	cording to given instructi
Goal					
Place $\langle ITEM_1 \rangle$ in	$\langle CONT_1 \rangle$ a	and $\langle ITEM \rangle$	$\left\langle t_{2} ight angle$ in $\left\langle CON ight angle$	$\langle T_2 \rangle$ respect	ively.
Actions					
	/ =: =				
pick up item with la put the item from ba			the basket v	vith label (C	$CONT.ID\rangle$
		,		•	
Difficulty Level					
Level1: The agent n					orresponding basket.
Level1: The agent n Level2: The agent n	eeds to put f	our items of	^F two differen	t kinds into	the corresponding basket
Level1: The agent n Level2: The agent n	eeds to put f	our items of	^F two differen	t kinds into	the corresponding baske
Level1: The agent n Level2: The agent n Level3: The agent n	eeds to put fo eeds to put fo	our items of our items of	f two differen f two differen	et kinds into et kinds into	the corresponding baske the corresponding baske
Level1: The agent n Level2: The agent n Level3: The agent n Example	eeds to put fo eeds to put fo	our items of our items of	f two differen f two differen	et kinds into et kinds into	the corresponding basket the corresponding basket
Level1: The agent n Level2: The agent n Level3: The agent n Example	eeds to put fo eeds to put fo	ification Lev	two different two different vell success	et kinds into et kinds into fully comple	the corresponding baske the corresponding baske
Level1: The agent n Level2: The agent n Level3: The agent n Example	eeds to put fo eeds to put fo	our items of our items of	f two differen f two differen	et kinds into et kinds into fully comple	the corresponding baske the corresponding baske
Level1: The agent n Level2: The agent n Level3: The agent n Example	eeds to put fo eeds to put fo	ification Lev	two different two different vell success	et kinds into et kinds into fully comple	the corresponding baske the corresponding baske
Level1: The agent n Level2: The agent n Level3: The agent n Example	eeds to put fo eeds to put fo	ification Lev	two different two different vell success	et kinds into et kinds into fully comple	the corresponding baske the corresponding baske
Level1: The agent n Level2: The agent n Level3: The agent n Example This is an example of	eeds to put fo eeds to put fo	ification Lev	two different two different vell successi	et kinds into et kinds into fully comple	the corresponding baske the corresponding baske eted by Gpt4o.
Level1: The agent n Level2: The agent n Level3: The agent n Example	eeds to put fo eeds to put fo	ification Lev	two different two different vell success	et kinds into et kinds into fully comple	the corresponding baske the corresponding baske
Level1: The agent n Level2: The agent n Level3: The agent n Example This is an example of	eeds to put fo eeds to put fo	ification Lev	two different two different vell successi	et kinds into et kinds into fully comple	the corresponding basket the corresponding basket eted by Gpt4o.
Level1: The agent n Level2: The agent n Level3: The agent n Example This is an example of	eeds to put fo eeds to put fo	ification Lev	two different two different vell successi	et kinds into et kinds into fully comple	the corresponding basket the corresponding basket eted by Gpt4o.
Level1: The agent n Level2: The agent n Level3: The agent n Example This is an example of	eeds to put fo eeds to put fo	ification Lev	two different two different vell successi	et kinds into et kinds into fully comple	the corresponding basket the corresponding basket eted by Gpt4o.
Level1: The agent n Level2: The agent n Level3: The agent n Example This is an example of	eeds to put fa eeds to put fa of task Classi	ification Lev	two different two different vell successi the successi th	t kinds into tt kinds into fully comple	the corresponding baske the corresponding baske eted by Gpt4o.
Level1: The agent n Level2: The agent n Level3: The agent n Example This is an example of	eeds to put fa eeds to put fa of task Classi	ification Lev	two different two different vell successi the successi th	t kinds into tt kinds into fully comple	the corresponding basket the corresponding basket eted by Gpt4o.
Level1: The agent n Level2: The agent n Level3: The agent n Example This is an example of Comparison of the second	eeds to put fa eeds to put fa of task Classi	ification Lev	two different two different vell successi the successi th	t kinds into tt kinds into fully comple	the corresponding basked the corresponding basked eted by Gpt4o.
Level1: The agent n Level2: The agent n Level3: The agent n Example This is an example of	eeds to put fa eeds to put fa of task Classi	ification Lev	two different two different vell successi the successi th	t kinds into tt kinds into fully comple	the corresponding baske the corresponding baske eted by Gpt4o.
Level1: The agent in Level2: The agent in Level3: The agent in Example This is an example of Comparison of the second Figure 1: Step1	eeds to put fa eeds to put fa of task Classi	ification Lev	two different two different vell successi the successi th	t kinds into tt kinds into fully comple	the corresponding baske the corresponding baske eted by Gpt4o.
Level1: The agent in Level2: The agent in Level3: The agent in Example This is an example of Figure 1 of the second secon	eeds to put fi eeds to put fi of task Classi iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii	our items of our items of ification Lev	two different two different vell successi the successi th	t kinds into tt kinds into fully comple	the corresponding baske the corresponding baske eted by Gpt4o.
Level1: The agent in Level2: The agent in Level3: The agent in Example This is an example of Figure 1 of the second secon	eeds to put fi eeds to put fi of task Classi up item with	ification Lev	two different two different vell successi the successi th	t kinds into tt kinds into fully comple	the corresponding baske the corresponding baske eted by Gpt4o.
Level1: The agent n Level2: The agent n Level3: The agent n Example This is an example of Fig • Step1 - Action List * A) 'pick * B) 'pick	eeds to put fa eeds to put fa of task Classi eeds to put fa task Classi eure 7: "The f up item with up item with	ification Lev	two different two different vell successi the successi th	t kinds into tt kinds into fully comple	the corresponding baske the corresponding baske eted by Gpt4o.
Level1: The agent n Level2: The agent n Level3: The agent n Example This is an example of Figure 1 of the second o	eeds to put fields to put fiel	our items of our items of ification Lev ification Lev progress of progress of h label 1' h label 0'	two different two different vell successi the successi th	t kinds into tt kinds into fully comple	the corresponding basked the corresponding basked eted by Gpt4o.
Level1: The agent n Level2: The agent n Level3: The agent n Example This is an example of Figure 1 of the second of the second Figure 1 of the second of the	eeds to put fa eeds to put fa of task Classi eeds to put fa task Classi eure 7: "The f up item with up item with	our items of our items of ification Lev ification Lev progress of progress of h label 1' h label 0'	two different two different vell successi the successi th	t kinds into tt kinds into fully comple	the corresponding basket the corresponding basket eted by Gpt4o.

Table 15: The Capability Score of each models.

1458	– Action List
1459	* A) 'put the item from backpack A into the basket with label 2'
1460	* B) 'put the item from backpack A into the basket with label 3'
1461	* C) 'pick up item with label 0'
1462	– Model's Response
1463	* B) 'put the item from backpack A into the basket with label 3'
1464	
1465	• Step3
1466	– Action List
1467	* A) 'pick up item with label 0'
1468	 Model's Response
1469	* A) 'pick up item with label 0'
1470	• Step4
1471 1472	– Action List
1472	* A) 'put the item from backpack A into the basket with label 3'
1474	 * A) put the item from backpack A into the basket with label 3' * B) 'put the item from backpack A into the basket with label 2'
1475	 Model's Response
1476	*
1477	* B) 'put the item from backpack A into the basket with label 2'
1478	
1479	G.2 COUNTING
1480	Introduction
1481	This task requires agent to pick up a certain number of items.
1482	
1483	Goal
1484	Colletct $\langle N \rangle \langle ITEM \rangle$.
1485	Actions
1486	Actions
1487	pick up $\langle ITEM angle$ with label $\langle ITEM.ID angle$
1488	Difficulty Level
1489	
1490	Level1: The agent needs to collect 1-3 items. Level2: The agent needs to collect 4-6 items.
1491	Level3: The agent needs to collect 7-10 items.
1492	Levels. The agent needs to conect 7-10 nems.
1493	Example
1494	This is an example of task Counting Level1 successfully completed by Gpt4o.
1495	
1496	
1497	
1498	
1499	
1500	



• Step1

- Action List

AA 🔊

ا 📣 AA • 🖶

^{*} A) 'pick up pizza with label 1'

1512	* B) 'pick up pizza with label 2'
1513	* C) 'pick up pizza with label 0'
1514	- Model's Response
1515	* C) 'pick up pizza with label 0'
1516	
1517	• Step2
1518	– Action List
1519	* A) 'pick up pizza with label 2'
1520 1521	* B) 'pick up pizza with label 1'
1521	 Model's Response
1523	* B) 'pick up pizza with label 1'
1524	
1525	G.3 SELECTION
1526	Introduction
1527 1528 1529	This task requires agent to memorize the items in the hint bar and collect all of them in the game frame.
1530	Goal
1531 1532	Memorize the item(s) on the left side of the game scene, then based on your memory, select the item(s) that have appeared before.
1533 1534	Actions
1535	choose $\langle ITEM \rangle$ with label $\langle ITEM.ID \rangle$
1536	Difficulty Level
1537	•
1538	Level1: The agent needs to remember one item in the hint bar and choose the right thing.
1539	Level2: The agent needs to remember two items in the hint bar and choose the right things. Level3: The agent needs to remember three items in the hint bar and choose the right things.
1540	
1541	Example
1542	

This is an example of task Selection Level1 successfully completed by Gpt40.



* A) 'choose toy with label 3'

- 1566 * B) 'choose toy with label 1' 1567
 - * C) 'choose toy with label 0'
 - * D) 'choose toy with label 2'
- 1569 - Model's Response 1570
 - * C) 'choose toy with label 0'

1572 G.4 DECODE 1573

1574 Introduction

1575 This task requires agent to learn the association rules from a provided code table and using these 1576 learned correspondences to select the correct item. 1577

Goal 1578

1568

1571

1579 Learn the association rules between items from the code table on the left of the game scene. Then 1580 you need to select the corresponding item to the target item inside the black square on the top of the 1581 code table.

1582 Actions 1583

choose $\langle ITEM \rangle$ with label $\langle ITEM.ID \rangle$ 1584

1585 **Difficulty Level** 1586

Level1: One set of correspondences is displayed in the hint bar on the left. 1587

Level2: Two sets of correspondences are displayed in the hint bar on the left. 1588

Level3: Three sets of correspondences are displayed in the hint bar on the left. 1589

This is an example of task Decode Level1 successfully completed by Gpt4o.

1590 Example

1591

1592 1593 1594

1595 1596 1597

1598

1601

1604

1607

1608

1609

1610

1611

1612

1613

1614

1615

1616 1617

1619



1602 1603

Figure 10: "The progress of Gpt4o finishing Decode Level1"

1605 • Step1

- Action List

- * A) 'choose item with label 3'
- * B) 'choose item with label 2'
- * C) 'choose item with label 0'
- * D) 'choose item with label 1'
- Model's Response
 - * To achieve the goal, we need to find the item associated with the target item (Pineapple) according to the code table. The code table shows that the cow (item 2) is associated with the Pineapple. Therefore, the first action letter I will choose is B) 'choose item with label 2'.

```
1618
      G.5 DECODE*
```

Introduction

This task requires agent not only to learn but to remember the association rules from a provided code table and using these learned correspondences to select the correct item.

1623 Goal

Memorize the association rules between items from the code table on the left of the game scene.
Then, based on your memory, you need to select the corresponding item to the target item inside the black square on the left side of the game scene.

1627 Actions 1628

1629 choose $\langle ITEM \rangle$ with label $\langle ITEM.ID \rangle$

¹⁶³⁰ Difficulty Level

1631 1632 Level1: One set of correspondences is displayed in the hint bar on the left.

Level2: Two sets of correspondences are displayed in the hint bar on the left.

Level3: Three sets of correspondences are displayed in the hint bar on the left.

1635 Example



This task requires agent to reconstruct a abstract target image by assembling scattered puzzle pieces from its backpack based on the visual reference provided in the hint bar.

1671 Goal

1673 Use the four puzzle pieces from your backpack to form the shape shown on the left of the game scene, pieces need to be properly placed in grid I, II, III, IV.



This is an example of task Filling Level1 successfully completed by Gpt4o.

1728	
1729	
1730 1731	
1732	
1733	hand the second se
1734	
1735	
1736	
1737	
1738	Figure 13: "The progress of Gpt40 finishing Filling Level1"
1739	Figure 15. The progress of Opt40 julishing Fulling Level1
1740	
1741	• Step1
1742	– Action List
1743	* A) 'place piece in backpack D at grid I'
1744 1745	* B) 'place piece in backpack B at grid I'
1745	* C) 'place piece in backpack C at grid I'
1747	* D) 'place piece in backpack A at grid I'
1748	 Model's Response
1749	* I choose action C) 'place piece in backpack C at grid I'.
1750	
1751	G.8 FILLING*
1752	Introduction
1753	
1754 1755	This task requires agent to remember a figurative animal in the hint bar and reconstruct it by assembling scattered pieces from its backpack.
1756	Goal
1757	Memorize the diagram on the left side of the game scene. Then, based on your memory, select a
1758 1759	picture piece from the backpack to fill in the missing part of the picture frame on the wall.
1760	Actions
1761 1762	place piece in backpack $\langle BAG.ID angle$ at grid $\langle GRID.ID angle$
1763	Difficulty Level
1764	Level1: The agent needs to select one picture piece to fill in the missing part.
1765	Level2: The agent needs to select two picture pieces to fill in the missing part.
1766	Level3: The agent needs to select three picture pieces to fill in the missing part.
1767	Example
1768	This is an example of task Filling* Level1 successfully completed by Gpt40.
1769 1770	This is an example of task I ming Leven successfully completed by Opt40.
1770 1771	
1772	
1773	
1774	
1775	
1776	
1777	
1778	
1779	
1780	
1781	Figure 14: "The progress of Gpt40 finishing Filling* Level1"

33

1782	• Step1
1783	– Action List
1784	* A) 'skip'
1785	– Model's Response
1786	* A) 'skip'.
1787	
1788	• Step2
1789 1790	– Action List
1791	* A) 'place piece in backpack D at grid I'
1792	* B) 'place piece in backpack A at grid I'
1793	 * C) 'place piece in backpack C at grid I' * D) 'place piece in backpack B at grid I'
1794	- Model's Response
1795	 * C) 'place piece in backpack C at grid I'.
1796	* C) place piece in backpack C at griu I.
1797	G.9 Maze
1798	
1799	Introduction
1800	This task requires agent to use the keys to unlock corresponding doors to get the diamond.
1801 1802	Goal
1803	
1804	Obtain the diamond in the dungeon. When your path is blocked by a door, you can use a key of the
1805	same color to unlock it.
1806	Actions
1807	obtain object with label $\langle ITEM.ID \rangle$
1808	use the key in backpack $\langle KEY.ID angle$ to unlock door with label $\langle DOOR.ID angle$
1809	Difficulty Level
1810	
1811 1812	Level1: The agent needs to open no more than one door to get to the diamond. Level2: The agent needs to open no more than two doors to get to the diamond.
1813	Level3: The agent needs to open no more than three doors to get to the diamond.
1814	Example
1815	-
1816	This is an example of task Maze Level1 successfully completed by Gpt4o.
1817	
1818	
1819	والالتكاف والالتجاف فالتلاف والتلق والتلق فالمراج الجافي والتلق فالتراج
1820	
1821	
1822	
1823	
1824	
1825 1826	
1827	
1828	Figure 15: "The progress of Gpt40 finishing Maze Level1"
1829	
1830	• Step1
1831	•
1832	 Action List * A) 'obtain object with label 1'
1000	\uparrow A) Ultani Ulteti wili lauel l

- * A) 'obtain object with label 1'
- * B) 'obtain object with label 2'
- * B) obtain obje
 * B) obtain obje
 * B) obtain obje
 * B) obtain obje
 * B) obtain obje

* I choose action B) 'obtain object with label 2'.

1836	• Step2
1837 1838	– Action List
1839	* A) 'use the key in backpack A to unlock door with label 0'
1840	* B) 'obtain object with label 1'
1841	 Model's Response
1842	* I choose action A) 'use the key in backpack A to unlock door with label 0'.
1843	• Step3
1844	– Action List
1845 1846	* A) 'obtain object with label 1'
1847	- Model's Response
1848	* I choose action A) 'obtain object with label 1'.
1849 1850	G.10 MAZE*
1851 1852	Introduction
1853 1854	This task requires agent to remember the location of the diamond and use the keys to unlock corre- sponding doors to get the diamond.
1855	Goal
1856 1857	Memorize the location of the diamond in the dungeon. Then, based on your memory, navigate to the
1858	correct treasure chest and obtain the diamond from it. When your path is blocked by a door, you can
1859	use a key of the same color to unlock it.
1860	Actions
1861 1862	obtain object with label $\langle ITEM.ID \rangle$ use the key in backpack $\langle KEY.ID \rangle$ to unlock door with label $\langle DOOR.ID \rangle$
1863 1864	Difficulty Level
1865 1866 1867	Level1: The agent needs to open no more than one door to get to the diamond. Level2: The agent needs to open no more than two doors to get to the diamond. Level3: The agent needs to open no more than three doors to get to the diamond.
1868	Example
1869 1870	This is an example of task Maze* Level1 successfully completed by Gpt40.
1871	
1872	
1873	
1874	
1875	
1876 1877	
1878	ਸ਼ਿਸ਼ ÷ ਸ ≪ ਸਮਾ ÷ ਸ ≪ ਸਮਾ ÷ ਸ ≪ ਸਮਾ ਸ ਅਤੇ ਸ ≪ ਸਮਾ ਸ * ਅ
1879	
1880	
1881	Figure 16: "The progress of Gpt40 finishing Maze* Level1"
1882	
1883	• Step1
1884 1885	-
1886	- Action List
1887	* A) 'skip'

- Model's Response ***** A) *** skip***.

1889 * A)

• Step2

1890	– Action List
1891 1892	* A) 'obtain object with label 0'
1893	 * B) 'obtain object with label 1'
1894	
1895	* C) 'obtain object with label 2'
1896	* D) 'obtain object with label 4'
1897	– Model's Response
1898	* C) 'obtain object with label 2'
1899	
1900	• Step3
1901 1902	•
1903	– Action List
1904	* A) 'use the key in backpack A to unlock door with label 3'
1905	
1906	* B) 'obtain object with label 1'
1907	* C) 'obtain object with label 0'
1908	* D) 'obtain object with label 4'
1909	- Model's Response
1910	* A) 'use the key in bag A to unlock door with label 3'.
1911 1912	* A) use the key in bag A to unlock door with laber 5.
1913	• Step4
1914	Step-
1915	– Action List
1916	
1917	* A) 'obtain object with label 1'
1918	* B) 'obtain object with label 0'
1919	* C) 'obtain object with label 4'
1920 1921	– Model's Response
1922	* A) 'obtain object with label 1'.
1923	
1924 1925	
1926	G.11 SORTING
1927	Introduction
1928	This task requires the agent to sort items based on a provided rule, even if the rule contradicts
1929 1930	real-world knowledge.
1931	Goal
1932	In this world, $\langle RULE \rangle$. Rank the $\langle TYPE \rangle$ in the backpack by $\langle PROPERTY \rangle$ in position I, II,
1933	III.
1934 1935	Actions
1936	place $\langle TYPE \rangle$ from bagpack $\langle BAG.ID \rangle$ at grid $\langle GRID.ID \rangle$
1937	Difficulty Level
1938 1939	<i>Level1: The agent needs to learn the new rule and sort the two things in corresponding order.</i>
1940	Level2: The agent needs to learn the new rule and sort the three things in corresponding order.
1941	Level3: The agent needs to learn the new rule and sort the four things in corresponding order.
1942	Example
1943	This is an example of task Sorting Level1 successfully completed by IntrenVL.

1944	
1945	
1946	
1947 1948	
1940	
1950	ادی اور
1951	
1952	
1953	
1954	Figure 17: "The progress of Internvl finishing Sorting Level1"
1955	riguie 17. The progress of miernor junshing sorting Leven
1956	
1957	• Step1
1958	5 Sup1
1959 1960	– Action List
1961	* A) 'place animal from backpack B at grid I'
1962	 * A) place animal from backpack B at grid I' * B) 'place animal from backpack A at grid II'
1963	
1964	* C) 'place animal from backpack A at grid I'
1965	* D) 'place animal from backpack B at grid II'
1966	 Model's Response
1967	* C) 'place animal from backpack A at grid I'
1968	
1969 1970	• Step2
1971	
1972	– Action List
1973	* A) 'place animal from backpack A at grid II'
1974	– Model's Response
1975	* A) 'place animal from backpack A at grid II'
1976	
1977	
1978 1979	G.12 PLACEMENT
1980	
1981	Introduction
1982	The task requires the agent to interpret the placement rules and place the item in the specified loca-
1983	tion.
1984	Goal
1985	Place the item in the opposite direction, such as if it is on the east side, you need to place it on the
1986	west side. Please place the $\langle ITEM_1 \rangle$ on the $\langle ORIENTATION \rangle$ side of the $\langle ITRM_2 \rangle$.
1987 1988	Actions
1989	
1990	place $\langle ITEM angle$ at grid $\langle GRID.ID angle$
1991	Difficulty Level
1992	Level1: The agent needs to put one item in opposite positions of the given position.
1993	Level2: The agent needs to place an item either clockwise or counterclockwise from a given location.
1994	Level3: The agent needs to place an item clockwise or counterclockwise in the opposite direction of
1995	a given location.
1996 1997	Example
1331	This is an example of task Placement Level1 successfully completed by Deepseek.

This is an example of task Placement Level1 successfully completed by Deepseek.

1998 1999 2000	
2001	
2002	
2003	
2004	
2005	
2006	
2007	
2008	Figure 18: "The progress of Deepseek finishing Placement Level1"
2009	
2010	
2011	• Step1
2012	– Action List
2013	* A) 'place elephant at grid I'
2014	* B) 'place elephant at grid II'
2015	* C) 'place elephant at grid V'
2016	* D) 'place elephant at grid VI'
2017	* E) 'place elephant at grid III'
2018	* F) 'place elephant at grid IV'
2019 2020	* G) 'place elephant at grid VII'
2020	* H) 'place elephant at grid VIII'
2021	– Model's Response
2022	* G) 'place elephant at grid VII'
2023	
2024	
2025	
2027	
2028	
2029	
2030	
2031	
2032	
2033	
2034	
2035	
2036	
2037	
2038	
2039	
2040	
2041	
2042	
2043	
2044	
2045	
2046	
2047	
2048	
2049	
2050	
2051	