4 A4. Detailed Results of Base LLM Evaluation

Detailed results of base LLM evaluation are in Table 3. In experiments on ARC-Bench, we mainly use Python 3.12.4 and Sympy 1.14.0.

Table 3: Detailed Results of Base LLM Evaluation

Method	ACR-Bench			Simple Single-step			Simple Multi-step			Complex Single-step			Complex Multi-step		
	Pass@1	Pass@5	R@5	Pass@1	Pass@5	R@5	Pass@1	Pass@5	R@5	Pass@1	Pass@5	R@5	Pass@1	Pass@5	R@5
Qwen2.5-Max	0.14	0.27	0.04	0.23	0.41	0.08	0.06	0.13	0.00	0.10	0.25	0.01	0.05	0.11	0.02
Hunyuan-Turbos	0.17	0.35	0.04	0.26	0.51	0.07	0.12	0.27	0.03	0.16	0.25	0.04	0.01	0.05	0.00
Qwen3-32b	0.17	0.35	0.06	0.27	0.50	0.12	0.10	0.26	0.01	0.13	0.30	0.01	0.07	0.14	0.01
GPT4o	0.19	0.34	0.07	0.31	0.51	0.16	0.11	0.28	0.00	0.14	0.26	0.03	0.05	0.09	0.00
Doubao-Seed-1.6	0.25	0.45	0.09	0.36	0.62	0.15	0.24	0.38	0.12	0.17	0.44	0.01	0.06	0.16	0.00
Gemini-2.0-Flash	0.26	0.42	0.13	0.42	0.58	0.26	0.22	0.39	0.04	0.15	0.35	0.05	0.04	0.13	0.00
Lamma4-17B-128E	0.26	0.44	0.11	0.36	0.62	0.13	0.22	0.37	0.11	0.19	0.34	0.09	0.11	0.18	0.09
GPT4.1	0.30	0.48	0.14	0.48	0.70	0.24	0.24	0.45	0.08	0.18	0.35	0.05	0.09	0.11	0.03
Deepseek-V3	0.36	0.53	0.19	0.54	0.71	0.34	0.24	0.44	0.13	0.25	0.52	0.06	0.15	0.23	0.06