

## A ALL AGENT AVERAGE SCORES PER FRAMEWORK

In Table 5, we include the average scores per framework and average-per-game score of all LLM agents.

Model	TEXTWORLD	TEXTWORLDEXPRESS	ALFWORLD	SCIENCEWORLD	JERICO	Average Score
o3 (medium)	100	91.9	88.3	93.0	15.7	58.7
o3 (high)	100	89.6	81.7	93.1	16.1	58.0
gpt-5 (thinking)	100	75.5	93.3	91.8	17.2	57.5
o3 (low)	99.1	89.8	70.0	88.3	14.2	54.8
claude-3.7-sonnet (thinking)	97.3	91.3	83.3	76.5	12.5	52.5
claude-3.7-sonnet	97.3	95.8	81.7	72.4	13.0	52.1
claude-3.5-sonnet-latest	95.5	81.6	75.0	82.3	9.6	50.4
gpt-4.1	95.3	92.5	83.3	76.1	6.8	49.9
gpt-5-mini (thinking)	94.7	61.9	61.7	82.7	9.5	46.5
o1	97.8	70.2	28.3	80.1	10.3	44.2
gpt-4o	83.6	80.6	56.7	61.4	5.6	40.6
claude-3.5-haiku	94.9	79.8	26.7	67.3	5.0	39.6
Llama-3.1-405B-Instruct	90.9	79.2	31.7	51.8	6.1	36.4
gemini-2.0-flash	80.8	76.1	20.0	57.1	5.4	35.0
Qwen3-32B	79.5	68.9	48.3	49.8	4.0	34.3
Llama-3.7-70B-Instruct	69.6	77.2	15.0	55.1	4.5	32.8
Llama-3.1-70B-Instruct	65.6	81.9	8.3	51.9	5.3	32.0
Qwen2.5-72B-Instruct	76.5	83.8	36.7	35.0	2.9	30.7
Mistral-Large-Instruct-2407	82.4	68.3	6.7	46.1	5.8	30.3
gpt-4.1-mini	62.1	74.5	5.0	41.9	3.4	27.1
gpt-4o-mini	56.5	73.6	0.0	27.2	1.8	21.8
Llama-4-Scout-17B-16E-Instruct	41.1	68.4	0.0	27.0	1.8	19.8
gpt-5-nano	50.1	41.3	1.7	32.0	1.7	18.3
Llama-4-Maverick-17B-128E-Instruct-FP8	43.5	56.1	8.3	11.5	2.0	15.5
Mistral-Small-Instruct-2409	56.1	27.3	0.0	24.4	1.4	14.8
Llama-3.1-8B-Instruct	29.7	50.3	0.0	15.7	2.3	13.9
DeepSeek-R1	37.1	38.6	0.0	15.8	1.0	12.4
Qwen2.5-7B-Instruct	27.7	45.6	0.0	12.6	0.7	11.7
Llama-3.2-3B-Instruct	21.4	42.0	0.0	10.0	1.5	10.4
phi-4	20.8	43.8	0.0	8.9	1.6	10.3
gpt-4.1-nano	12.8	38.7	0.0	9.4	3.6	10.0
Mistral-Small-24B-Instruct-2501	15.8	23.0	0.0	15.8	1.4	8.8
DeepSeek-R1-Distill-Llama-70B	8.7	39.8	0.0	7.7	1.3	8.4
Ministral-8B-Instruct-2410	10.9	22.8	0.0	2.3	0.4	4.6
Mistral-Small-3.1-24B-Instruct-2503	2.5	10.3	0.0	10.5	0.8	4.5
Mixtral-8x22B-Instruct-v0.1	17.1	8.4	0.0	4.0	0.4	3.7
Llama-3.2-1B-Instruct	0.0	19.0	0.0	2.4	0.6	3.3
Phi-3-mini-128k-instruct	2.7	9.4	0.0	2.4	0.3	2.2
Phi-3.5-MoE-instruct	0.0	7.0	0.0	2.3	0.4	1.7
Phi-4-mini-instruct	0.0	5.5	0.0	2.3	0.5	1.5
Mixtral-8x7B-Instruct-v0.1	0.0	1.6	0.0	4.0	0.3	1.3
Phi-3.5-mini-instruct	0.0	2.0	0.0	2.4	0.5	1.0
Phi-3-medium-128k-instruct	0.0	0.0	0.0	2.3	0.3	0.7

Table 5: Average scores per framework and total TALES score.

## B JERICHO WALKTHROUGH SCORES

Table 6 shows the percent of achievable score when using the walkthrough for all JERICHO for 50, 100, 200, 300, 400, 500 and 1000 steps.

Game	50 Steps	100 Steps	200 Steps	300 Steps	400 Steps	500 Steps	1000 Steps
JerichoEnv905	100.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvAcorncourt	100.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvAdvent	26.300	42.600	63.100	100.000	100.000	100.000	100.000
JerichoEnvAdventureland	21.000	42.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvAfflicted	46.700	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvAnchor	5.000	11.000	29.000	41.000	52.000	64.000	99.000
JerichoEnvAwaken	60.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvBalances	58.800	58.800	98.000	98.000	98.000	98.000	98.000
JerichoEnvBallyhoo	15.000	30.000	50.000	75.000	95.000	100.000	100.000
JerichoEnvCurses	3.800	5.600	12.700	28.200	38.200	47.500	81.800
JerichoEnvCutthroat	12.000	28.000	36.000	44.000	100.000	100.000	100.000
JerichoEnvDeephyme	20.700	28.000	60.000	76.000	100.000	100.000	100.000
JerichoEnvDetective	100.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvDragon	24.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvEnchanter	11.300	31.200	70.000	100.000	100.000	100.000	100.000
JerichoEnvEnter	35.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvGold	12.000	30.000	51.000	75.000	100.000	100.000	100.000
JerichoEnvHhgg	8.300	21.200	40.000	50.000	100.000	100.000	100.000
JerichoEnvHuntdark	0.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvInfidel	12.500	20.000	70.000	100.000	100.000	100.000	100.000
JerichoEnvInhumane	33.300	77.800	100.000	100.000	100.000	100.000	100.000
JerichoEnvJewel	15.600	26.700	77.800	100.000	100.000	100.000	100.000
JerichoEnvKarn	5.900	23.500	38.200	67.600	100.000	100.000	100.000
JerichoEnvLibrary	100.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvLoose	100.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvLostpig	28.600	42.900	85.700	85.700	85.700	85.700	85.700
JerichoEnvLudicorp	13.300	25.300	58.700	92.700	100.000	100.000	100.000
JerichoEnvLurking	10.000	25.000	55.000	100.000	100.000	100.000	100.000
JerichoEnvMoonlit	0.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvMurdac	6.800	18.000	18.000	48.000	99.600	99.600	99.600
JerichoEnvNight	60.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvOmniquest	40.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvPartyfoul	0.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvPentari	100.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvPlanetfall	7.500	26.300	35.000	60.000	100.000	100.000	100.000
JerichoEnvPlundered	16.000	44.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvReverb	60.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvSeastalker	28.000	44.000	90.000	100.000	100.000	100.000	100.000
JerichoEnvSherlock	23.000	37.000	55.000	84.000	100.000	100.000	100.000
JerichoEnvSnacktime	100.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvSorcerer	23.700	37.500	53.700	100.000	100.000	100.000	100.000
JerichoEnvSpellbrkr	13.300	26.700	42.500	65.000	91.700	100.000	100.000
JerichoEnvSpirit	2.400	3.200	9.600	14.400	18.800	27.200	71.200
JerichoEnvTemple	28.600	57.100	100.000	100.000	100.000	100.000	100.000
JerichoEnvTrinity	15.000	22.000	32.000	47.000	58.000	78.000	100.000
JerichoEnvTryst205	2.900	14.300	24.300	41.400	58.600	74.300	100.000
JerichoEnvWeapon	0.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvWishbringer	24.000	50.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvYomomma	25.700	97.100	97.100	97.100	97.100	97.100	97.100
JerichoEnvZenon	40.000	100.000	100.000	100.000	100.000	100.000	100.000
JerichoEnvZork1	18.000	29.100	41.700	77.400	100.000	100.000	100.000
JerichoEnvZork2	6.200	22.500	47.500	100.000	100.000	100.000	100.000
JerichoEnvZork3	28.600	42.900	100.000	100.000	100.000	100.000	100.000
JerichoEnvZtuu	47.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 6: Max score percentage reached by following the provided walkthrough for each JERICHO game.

## C ALL AGENT AVERAGE FINAL TOKENS USED PER FRAMEWORK

In Table 7, we include the average final tokens used per game for each framework of all agents.

Model	TEXTWORLD	TEXTWORLDEXPRESS	ALFWORLD	SCIENCEWORLD	JERICO
o3 (medium)	41342.0	107026.3	128740.7	92422.7	378168.5
o3 (high)	32148.6	51361.7	68275.0	51323.4	251760.8
gpt-5 (thinking)	38774.8	377563.2	145110.7	197999.9	649059.0
o3 (low)	51609.6	84218.5	119657.4	88557.9	318262.7
claude-3.7-sonnet (thinking)	69138.9	63072.8	74516.3	128752.7	311684.2
claude-3.7-sonnet	72072.9	46948.4	65131.1	152130.6	298533.1
claude-3.5-sonnet-latest	60766.1	68812.7	78765.4	106749.2	291950.3
gpt-4.1	53378.9	46454.8	54107.6	86547.5	218123.2
gpt-5-mini (thinking)	151984.2	391444.6	508585.4	318636.6	878673.2
o1	47765.6	113492.9	127460.2	74300.1	211958.7
gpt-4o	106863.0	49536.2	77946.0	107121.9	209712.8
claude-3.5-haiku	119839.9	85136.1	267643.6	205751.0	269458.9
Llama-3.1-405B-Instruct	66476.2	52624.6	106290.0	137657.2	226078.6
gemini-2.0-flash	142937.1	66075.5	138048.3	142883.3	230182.0
Qwen3-32B	198390.4	188065.3	190900.1	229708.1	374514.2
Llama-3.3-70B-Instruct	166373.8	70165.0	127348.7	128860.9	205362.8
Llama-3.1-70B-Instruct	133253.4	51885.9	106925.7	144615.6	210914.8
Qwen2.5-72B-Instruct	112658.0	52096.1	97211.2	168057.3	197628.3
Mistral-Large-Instruct-2407	107788.5	110228.6	118395.4	163232.0	243256.0
gpt-4.1-mini	184516.3	92775.4	130758.2	125310.5	188824.3
gpt-4o-mini	159840.9	60210.7	145236.3	172875.7	182620.7
Llama-4-Scout-17B-16E-Instruct	289709.8	120173.9	172633.9	222464.1	229947.2
gpt-5-nano	770352.7	623055.1	821024.1	730904.7	825754.3
Llama-4-Maverick-17B-128E-Instruct-FP8	287547.2	213139.5	354183.9	394875.5	372902.6
Mistral-Small-Instruct-2409	163334.9	304510.9	107549.3	150730.7	208261.8
Llama-3.1-8B-Instruct	222239.7	358837.4	96582.5	152293.0	165505.8
DeepSeek-R1	393654.5	398322.7	496328.4	431997.9	439399.3
Qwen2.5-7B-Instruct	143127.1	214926.3	91334.4	163021.2	171107.7
Llama-3.2-3B-Instruct	230950.5	79878.3	84620.4	195397.2	152544.7
phi-4	189031.6	100363.9	126068.1	153395.2	178713.4
gpt-4.1-nano	545577.7	171767.5	277643.8	201505.9	182678.8
Mistral-Small-24B-Instruct-2501	399093.8	500484.8	479125.0	418284.9	475649.7
DeepSeek-R1-Distill-Llama-70B	453695.7	637384.1	719404.2	482819.3	407401.8
Ministral-8B-Instruct-2410	220157.9	337447.5	112710.5	108916.4	118104.9
Mistral-Small-3.1-24B-Instruct-2503	448764.0	507986.4	477505.8	397054.6	514733.5
Mixtral-8x22B-Instruct-v0.1	158782.2	137583.5	92832.7	134827.6	156515.8
Llama-3.2-1B-Instruct	567691.8	279214.8	457857.3	138285.6	201648.5
Phi-3-mini-128k-instruct	245215.0	429993.4	257852.2	253989.5	237881.5
Phi-3.5-MoE-instruct	274848.9	295190.9	240007.5	252055.6	271680.0
Phi-4-mini-instruct	231947.3	199299.1	195407.4	190887.4	212508.9
Mixtral-8x7B-Instruct-v0.1	612791.9	555281.3	520434.6	560994.6	564967.6
Phi-3.5-mini-instruct	426125.5	476218.4	410459.6	327584.9	457434.4
Phi-3-medium-128k-instruct	620235.4	585925.5	581721.6	513787.5	595335.5

Table 7: Avg final tokens used per LLM per game for each framework. Ordering is based on the agent’s cumulative average score shown in Table 5.

## D AGENT SCORE STANDARD DEVIATIONS

In Table 8, we include the average standard deviation across seeds per framework of all LLM agents.

Model	TEXTWORLD	TEXTWORLDEXPRESS	ALFWORLD	SCIENCEWORLD	JERICO
o3 (medium)	0.0	2.7	4.6	2.2	0.5
o3 (high)	0.0	3.2	9.1	1.1	1.3
gpt-5 (thinking)	0.0	5.5	7.0	2.5	1.1
o3 (low)	2.0	6.9	9.5	0.8	1.1
claude-3.7-sonnet (thinking)	2.8	4.7	10.2	2.9	0.9
claude-3.7-sonnet	0.0	1.4	3.7	3.7	1.1
claude-3.5-sonnet-latest	0.0	2.9	5.9	3.4	1.0
gpt-4.1	2.6	1.9	11.8	2.3	0.8
gpt-5-mini (thinking)	3.6	2.8	9.5	6.1	1.4
o1	1.2	4.4	4.6	5.0	1.7
gpt-4o	6.1	0.4	14.9	2.8	0.6
claude-3.5-haiku	5.3	0.0	3.7	2.6	0.6
Llama-3.1-405B-Instruct	5.0	4.9	10.9	4.5	0.5
gemini-2.0-flash	8.6	1.3	4.6	3.4	0.4
Qwen3-32B	6.8	1.9	10.9	3.2	0.4
Llama-3.3-70B-Instruct	2.8	3.4	3.7	2.3	0.1
Llama-3.1-70B-Instruct	3.5	1.9	5.9	4.5	0.2
Qwen2.5-72B-Instruct	2.0	2.5	4.6	3.8	0.7
Mistral-Large-Instruct-2407	8.2	2.6	3.7	8.1	0.9
gpt-4.1-mini	6.1	1.7	7.5	3.6	0.3
gpt-4o-mini	5.4	1.7	0.0	1.5	0.2
Llama-4-Scout-17B-16E-Instruct	0.0	0.0	0.0	0.0	0.0
gpt-5-nano	7.7	5.1	3.7	4.1	0.3
Llama-4-Maverick-17B-128E-Instruct-FP8	1.3	0.0	0.0	0.1	0.3
Mistral-Small-Instruct-2409	5.1	0.0	0.0	2.2	0.0
Llama-3.1-8B-Instruct	4.7	2.9	0.0	0.9	0.1
DeepSeek-R1	3.9	0.0	0.0	2.2	0.1
Qwen2.5-7B-Instruct	0.0	0.0	0.0	0.7	0.1
Llama-3.2-3B-Instruct	2.6	2.9	0.0	1.6	0.3
phi-4	0.4	0.0	0.0	1.3	0.0
gpt-4.1-nano	2.1	4.5	0.0	1.0	2.4
Mistral-Small-24B-Instruct-2501	3.1	1.0	0.0	1.1	0.3
DeepSeek-R1-Distill-Llama-70B	2.8	0.3	0.0	0.4	0.1
Ministral-8B-Instruct-2410	4.2	0.0	0.0	0.0	0.0
Mistral-Small-3.1-24B-Instruct-2503	0.0	0.0	0.0	0.3	0.0
Mixtral-8x22B-Instruct-v0.1	3.0	2.3	0.0	1.7	0.1
Llama-3.2-1B-Instruct	0.0	0.0	0.0	0.0	0.0
Phi-3-mini-128k-instruct	2.0	0.0	0.0	0.3	0.0
Phi-3.5-MoE-instruct	0.0	2.7	0.0	0.0	0.1
Phi-4-mini-instruct	0.0	0.0	0.0	0.0	0.0
Mixtral-8x7B-Instruct-v0.1	0.0	0.0	0.0	0.0	0.0
Phi-3.5-mini-instruct	0.0	1.1	0.0	0.1	0.0
Phi-3-medium-128k-instruct	0.0	0.0	0.0	0.0	0.0

Table 8: Standard deviation statistics for different LLMs Ordering is based on the agent’s cumulative average score shown in Table 5.

## E ALL GAMES

In Table 9 and Table 10 we list all tasks and games in their respective frameworks.

Table 9: Games Organized by Framework. Part 1.

Jericho		
1. 905	19. HuntDark	37. Reverb
2. AcornCourt	20. Infidel	38. Seastalker
3. Advent	21. Inhumane	39. Sherlock
4. Adventureland	22. Jewel	40. Snacktime
5. Afflicted	23. Karn	41. Sorcerer
6. Anchor	24. Library	42. Spellbrkr
7. Awaken	25. Loose	43. Spirit
8. Balances	26. Lostpig	44. Temple
9. Ballyhoo	27. Ludicorp	45. Theatre
10. Curses	28. Lurking	46. Trinity
11. Cutthroat	29. Moonlit	47. Tryst205
12. Deephome	30. Murdac	48. Weapon
13. Detective	31. Night	49. Wishbringer
14. Dragon	32. Omniquest	50. Yomomma
15. Enchanter	33. Partyfoul	51. Zenon
16. Enter	34. Pentari	52. Zork1
17. Gold	35. Planetfall	53. Zork2
18. Hhgg	36. Plundered	54. Zork3
		55. Ztuu

ScienceWorld	
1. Boil	16. InclinedPlaneFrictionNamedSurfaces
2. ChangeTheStateOfMatterOf	17. InclinedPlaneFrictionUnnamedSurfaces
3. ChemistryMix	18. LifespanLongestLived
4. ChemistryMixPaintSecondaryColor	19. LifespanLongestLivedThenShortestLived
5. ChemistryMixPaintTertiaryColor	20. LifespanShortestLived
6. FindAnimal	21. MeasureMeltingPointKnownSubstance
7. FindLivingThing	22. MeasureMeltingPointUnknownSubstance
8. FindNonLivingThing	23. Melt
9. FindPlant	24. MendelianGeneticsKnownPlant
10. Freeze	25. MendelianGeneticsUnknownPlant
11. GrowFruit	26. PowerComponent
12. GrowPlant	27. PowerComponentRenewableVsNonrenewableEnergy
13. IdentifyLifeStages1	28. TestConductivity
14. IdentifyLifeStages2	29. TestConductivityOfUnknownSubstances
15. InclinedPlaneDetermineAngle	30. UseThermometer

Table 10: Games Organized by Framework. Part 2.

<b>ALFWorld</b>	
1. LookAtObjInLightSeen	7. PickCoolThenPlaceInRecepSeen
2. LookAtObjInLightUnseen	8. PickCoolThenPlaceInRecepUnseen
3. PickAndPlaceSimpleSeen	9. PickHeatThenPlaceInRecepSeen
4. PickAndPlaceSimpleUnseen	10. PickHeatThenPlaceInRecepUnseen
5. PickCleanThenPlaceInRecepSeen	11. PickTwoObjAndPlaceSeen
6. PickCleanThenPlaceInRecepUnseen	12. PickTwoObjAndPlaceUnseen
<b>TextWorld</b>	
1. CookingLevel1	6. CookingLevel6
2. CookingLevel2	7. CookingLevel7
3. CookingLevel3	8. CookingLevel8
4. CookingLevel4	9. CookingLevel9
5. CookingLevel5	10. CookingLevel10
<b>TWX</b>	
1. Arithmetic	9. SimonSaysWithMemory10
2. CoinCollector	10. SimonSaysWithMemory50
3. CookingWorld	11. SimonSaysWithMemory100
4. MapReader	12. SimonSaysWithMemory10Verbose
5. PeckingOrder	13. SimonSaysWithMemory50Verbose
6. SimonSays10	14. SimonSaysWithMemory100Verbose
7. SimonSays50	15. Sorting
8. SimonSays100	16. TextWorldCommonsense

## F ALL SCORES PER GAME: TEXTWORLD

Table 11 shows the per-game scores of all models in TEXTWORLD across all seeds.

## G ALL SCORES PER GAME: TEXTWORLDEXPRESS

Table 12 shows the average per-game scores of all models in TEXTWORLDEXPRESS across all seeds.

## H ALL SCORES PER GAME: ALFWORLD

Table 13 shows the average per-game scores of all models in ALFWORLD across all seeds.

## I ALL SCORES PER GAME: SCIENCEWORLD

Tables 14 and 15 shows the per-task average scores of all models in SCIENCEWORLD across all seeds.

## J ALL SCORES PER GAME: JERICHO

Tables 16 and 17 shows the per-game scores of all models in JERICHO. \* Indicates LLM has only been run on one seed. We will update the paper once all run seeds have been completed.

Table 11: Model Performance on TEXTWORLD Tasks.

Models	CookingLevel1	CookingLevel2	CookingLevel3	CookingLevel4	CookingLevel5	CookingLevel6	CookingLevel7	CookingLevel8	CookingLevel9	CookingLevel10
o3 (medium)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
o3 (high)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
gpt-5 (thinking)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
o3 (low)	100.0	90.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
claude-3.7-sommet (thinking)	100.0	72.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
claude-3.7-sommet	100.0	72.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
claude-3.5-sommet-latest	100.0	54.5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
gpt-4.1	100.0	61.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.9
gpt-5-mini (thinking)	100.0	47.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
o1	100.0	78.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
gpt-4o	86.7	14.5	70.0	100.0	86.7	100.0	100.0	100.0	100.0	78.2
claude-3.5-haiku	100.0	58.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.9
Llama-3.1-405B-Instruct	100.0	54.5	85.0	100.0	100.0	100.0	100.0	100.0	100.0	69.1
gemini-2.0-flash	100.0	21.8	25.0	100.0	100.0	86.7	100.0	100.0	84.0	90.9
Qwen3-32B	100.0	27.3	85.0	100.0	100.0	46.7	86.7	100.0	68.0	81.8
Llama-3.3-70B-Instruct	100.0	0.0	25.0	100.0	100.0	0.0	100.0	100.0	92.0	90.9
Llama-3.1-70B-Instruct	100.0	0.0	25.0	100.0	100.0	0.0	100.0	100.0	84.0	47.3
Qwen2.5-72B-Instruct	100.0	0.0	25.0	100.0	100.0	100.0	100.0	100.0	100.0	40.0
Mistral-Large-Instruct-2407	86.7	32.7	100.0	100.0	86.7	100.0	100.0	100.0	100.0	18.2
gpt-4.1-mini	60.0	5.5	70.0	100.0	46.7	86.7	100.0	40.0	76.0	36.4
gpt-4o-mini	100.0	0.0	25.0	100.0	46.7	100.0	0.0	52.0	100.0	41.8
Llama-4-Scout-17B-16E-Instruct	33.3	27.3	25.0	100.0	100.0	0.0	0.0	0.0	40.0	45.5
gpt-5-nano	60.0	12.7	85.0	100.0	33.3	0.0	0.0	60.0	20.0	45.5
Llama-4-Maverick-17B-128E-Instruct-FP8	100.0	0.0	25.0	100.0	20.0	0.0	100.0	32.0	16.0	41.8
Mistral-Small-Instruct-2409	100.0	0.0	25.0	100.0	100.0	20.0	0.0	100.0	76.0	40.0
Llama-3.1-8B-Instruct	73.3	1.8	25.0	0.0	100.0	0.0	0.0	0.0	60.0	9.1
DeepSeek-R1	66.7	0.0	25.0	75.0	100.0	0.0	0.0	28.0	40.0	36.4
Qwen2.5-7B-Instruct	33.3	18.2	25.0	100.0	0.0	0.0	0.0	100.0	0.0	0.0
Llama-3.2-3B-Instruct	40.0	0.0	25.0	100.0	0.0	0.0	0.0	24.0	16.0	9.1
phi-4	0.0	0.0	0.0	0.0	33.3	33.3	0.0	40.0	100.0	1.8
gpt-4.1-nano	33.3	0.0	25.0	0.0	0.0	0.0	13.3	56.0	0.0	0.0
Mistral-Small-24B-Instruct-2501	46.7	0.0	25.0	25.0	33.3	0.0	0.0	20.0	8.0	0.0
DeepSeek-R1-Distill-Llama-70B	6.7	0.0	25.0	55.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-8B-Instruct-2410	33.3	0.0	40.0	0.0	0.0	0.0	0.0	36.0	0.0	0.0
Mistral-Small-3.1-24B-Instruct-2503	0.0	0.0	25.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-8x22B-Instruct-v0.1	6.7	0.0	0.0	100.0	0.0	0.0	0.0	64.0	0.0	0.0
Llama-3.2-1B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3-mini-128k-instruct	0.0	0.0	10.0	5.0	0.0	0.0	0.0	12.0	0.0	0.0
Phi-3.5-MoE-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-4-mini-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-8x7B-Instruct-v0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3.5-mini-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3-medium-128k-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 12: Model Performance on TEXTWORLD EXPRESS tasks.

Models	Arithmetic	CodeCollector	CookingWorld	MapReader	PackingOrder	SimulSisy-10	SimulSisy-50	SimulSisy-100	SimulSisy-500	SimulSisy-WithMemory10	SimulSisy-WithMemory100	SimulSisy-WithMemory1000	SimulSisy-WithMemory10000	SimulSisy-WithMemory50	SimulSisy-WithMemory500	Sorting	TextWorldCommaence
o3 (medium)	100.0	100.0	76.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.2	100.0	100.0	100.0
o3 (low)	80.0	100.0	55.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	92.2	100.0	100.0	100.0
gpt-5 (thinking)	40.0	100.0	42.0	100.0	100.0	100.0	100.0	90.2	100.0	100.0	100.0	100.0	100.0	78.4	60.0	60.0	50.0
o3 (low)	90.0	100.0	42.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	60.0	60.0	80.0
claude-3.7-sonnet (thinking)	100.0	100.0	39.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	70.0
claude-3.7-sonnet-latest	100.0	100.0	39.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0
gpt-4.1	80.0	100.0	30.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	70.0
gpt-5-mini (thinking)	0.0	100.0	33.6	100.0	100.0	100.0	100.0	68.0	88.0	100.0	100.0	100.0	100.0	32.2	100.0	100.0	50.0
o1	60.0	100.0	42.0	90.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	46.0	100.0	0.0	50.0
gpt-4o	0.0	100.0	32.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	50.0
gpt-5.5-baihu	100.0	100.0	32.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	50.0
Llama-3.5-8B-Instruct	80.0	100.0	33.6	100.0	100.0	100.0	100.0	21.0	42.0	100.0	100.0	100.0	100.0	100.0	100.0	40.0	50.0
gemini-2.0-flash	0.0	100.0	30.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	70.4	100.0	12.0	50.0
Qwen3.3-72B-Instruct	20.0	100.0	28.0	80.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	86.4	100.0	0.0	50.0
Llama-3.3-70B-Instruct	0.0	100.0	28.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	50.0
Qwen2.5-72B-Instruct	0.0	100.0	22.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	68.0	50.0
Mistral-Large-Instruct-2407	20.0	100.0	42.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	19.2	100.0	100.0	50.0
gpt-4.1-mini	50.0	100.0	42.0	50.0	30.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	50.0
gpt-4o-mini	0.0	100.0	42.0	50.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	50.0
gpt-5.5-sonnet-17B-16E-Instruct	0.0	100.0	42.0	100.0	100.0	100.0	100.0	70.0	86.8	100.0	100.0	100.0	100.0	100.0	100.0	0.0	50.0
Llama-4-Maverick-17B-128E-Instruct-FP8	0.0	100.0	0.0	100.0	60.0	100.0	100.0	55.0	100.0	100.0	100.0	100.0	100.0	26.0	100.0	0.0	30.0
Llama-4-Maverick-17B-128E-Instruct-FP8	0.0	100.0	42.0	100.0	100.0	100.0	100.0	55.0	100.0	100.0	100.0	100.0	100.0	26.0	100.0	0.0	50.0
Mistral-Small-Instruct-2409	0.0	100.0	28.0	0.0	0.0	100.0	100.0	21.0	42.0	100.0	100.0	100.0	100.0	14.0	100.0	0.0	50.0
Llama-3.1-8B-Instruct	10.0	100.0	2.8	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	2.0	100.0	0.0	50.0
Qwen2.5-72B-Instruct	0.0	100.0	0.0	0.0	25.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	30.0	100.0	0.0	50.0
Llama-3.2-3B-Instruct	0.0	80.0	14.0	0.0	85.0	0.0	40.0	8.0	33.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.0
phi-4	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.0
gpt-4.1-turbo	0.0	100.0	0.0	0.0	0.0	100.0	100.0	21.0	42.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.0
gpt-5.5-sonnet-17B-16E-Instruct-2501	0.0	100.0	0.0	0.0	0.0	100.0	100.0	21.0	42.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.0
DeepSeek-R1-Distill-Llama-70B	0.0	100.0	28.0	50.0	50.0	100.0	100.0	34.0	68.0	100.0	100.0	100.0	100.0	4.0	100.0	0.0	50.0
Mistral-8B-Instruct-2410	0.0	0.0	0.0	0.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	2.0	100.0	0.0	0.0
Mistral-Small-3.1-24B-Instruct-2503	0.0	0.0	0.0	0.0	0.0	20.0	2.0	2.0	4.0	100.0	100.0	100.0	100.0	4.0	100.0	0.0	0.0
Mistral-8x2B-Instruct-v0.1	0.0	20.0	0.0	0.0	0.0	17.0	34.0	0.0	0.0	100.0	100.0	100.0	100.0	4.0	100.0	0.0	0.0
Phi-3.5-MoE-Instruct	0.0	0.0	0.0	0.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	4.0	100.0	0.0	0.0
Phi-3-mini-128k-instruct	0.0	80.0	0.0	0.0	0.0	20.0	4.0	0.0	0.0	100.0	100.0	100.0	100.0	0.0	100.0	0.0	0.0
Phi-3.5-MoE-instruct	0.0	80.0	0.0	0.0	0.0	6.0	3.2	2.0	3.2	100.0	100.0	100.0	100.0	0.4	100.0	0.0	0.0
Phi-4-mini-instruct	0.0	0.0	0.0	0.0	25.0	0.0	21.0	42.0	0.0	100.0	100.0	100.0	100.0	0.0	100.0	0.0	0.0
Mistral-8x7B-Instruct-v0.1	0.0	0.0	0.0	0.0	0.0	20.0	4.0	0.0	0.0	100.0	100.0	100.0	100.0	0.0	100.0	0.0	0.0
Phi-3.5-mini-128k-instruct	0.0	0.0	0.0	0.0	0.0	17.0	34.0	0.0	0.0	100.0	100.0	100.0	100.0	0.0	100.0	0.0	0.0
Phi-3-medium-128k-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	100.0	100.0	0.0	100.0	0.0	0.0



Table 13: Model Performance on ALFWORLD tasks.

Models	LookAtObjInLightSeen	LookAtObjInLightUnseen	PickAndPlaceSimpleSeen	PickAndPlaceSimpleUnseen	PickCleanThenPlaceInReceptSeen	PickCleanThenPlaceInReceptUnseen	PickCoolThenPlaceInReceptSeen	PickCoolThenPlaceInReceptUnseen	PickHeatThenPlaceInReceptSeen	PickHeatThenPlaceInReceptUnseen	PickTwoObjAndPlaceSeen	PickTwoObjAndPlaceUnseen
o3 (medium)	100.0	100.0	100.0	100.0	60.0	100.0	100.0	100.0	60.0	80.0	80.0	80.0
o3 (high)	60.0	100.0	100.0	100.0	20.0	60.0	80.0	80.0	80.0	100.0	100.0	100.0
gpt-5 (thinking)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
o3 (low)	80.0	100.0	80.0	100.0	60.0	80.0	20.0	80.0	40.0	100.0	60.0	60.0
claude-3.7-sonnet (thinking)	100.0	100.0	100.0	100.0	80.0	100.0	100.0	100.0	60.0	80.0	80.0	80.0
claude-3.7-sonnet	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
claude-3.7-sonnet-latest	20.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	60.0	100.0	100.0	100.0
gpt-4.1	80.0	100.0	100.0	100.0	100.0	80.0	100.0	100.0	40.0	100.0	100.0	100.0
gpt-5-mini (thinking)	100.0	100.0	100.0	100.0	20.0	60.0	80.0	40.0	60.0	100.0	0.0	0.0
o1	100.0	40.0	40.0	20.0	0.0	40.0	60.0	40.0	0.0	0.0	0.0	0.0
gpt-4o	100.0	100.0	100.0	100.0	60.0	60.0	60.0	60.0	40.0	60.0	80.0	80.0
Llama-3.1-405B-Instruct	80.0	0.0	100.0	80.0	0.0	0.0	0.0	0.0	20.0	0.0	20.0	20.0
Llama-3.1-70B-Instruct	100.0	0.0	100.0	100.0	100.0	60.0	60.0	60.0	100.0	100.0	80.0	80.0
Gemma-3.2B	80.0	0.0	100.0	60.0	80.0	60.0	20.0	40.0	20.0	80.0	0.0	0.0
Llama-3.3-70B-Instruct	100.0	20.0	40.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2.5-72B-Instruct	100.0	0.0	60.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-Large-Instruct-2407	100.0	0.0	100.0	100.0	0.0	80.0	0.0	0.0	0.0	0.0	60.0	0.0
gpt-4.1-mini	0.0	0.0	0.0	40.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
gpt-4o-mini	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama-4-Scout-17B-16E-Instruct	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama-4-Maverick-17B-128E-Instruct-EP8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-Small-Instruct-2409	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3.1-8B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DeepSeek-R1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2.5-7B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3.2-3B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
gpt-4.1-nano	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-Small-24B-Instruct-2501	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DeepSeek-R1-Distill-Llama-70B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-8B-Instruct-2410	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-Small-3.1-24B-Instruct-2503	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-8x22B-Instruct-v0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3.2-1B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3.5-MoE-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3.5-MoE-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-4-mini-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-8x7B-Instruct-v0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3.5-mini-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3-medium-128k-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 14: Model Performance on SCIENCEWORLD tasks. Part 1

Models	Boil	ChangeTheStateOfMatterOf	ChemistryMix	ChemistryMixPaintSecondaryColor	ChemistryMixPaintTertiaryColor	FindAnimal	FindLivingThing	FindNonLivingThing	FindPlant	Freeze	CrowFruit	CrowPlant	IdentifyLifeStages1	IdentifyLifeStages2	InclinedPlaneDetermineAngle
o3 (medium)	48.4	100.0	100.0	100.0	100.0	85.0	100.0	100.0	100.0	48.0	100.0	85.2	89.0	100.0	100.0
o3 (high)	68.0	100.0	100.0	100.0	100.0	88.2	100.0	100.0	100.0	49.0	87.4	92.2	78.0	100.0	100.0
gpt-5 (thinking)	95.6	100.0	100.0	100.0	100.0	88.2	100.0	100.0	100.0	53.0	44.4	77.4	65.4	100.0	100.0
o3 (low)	5.8	100.0	100.0	100.0	82.0	95.0	100.0	100.0	100.0	53.0	45.2	77.0	57.6	100.0	100.0
claude-3.7-sonnet (thinking)	5.8	70.0	88.4	100.0	94.0	85.0	100.0	100.0	100.0	37.4	75.0	48.4	50.4	54.6	90.0
claude-3.7-sonnet	5.8	43.2	100.0	100.0	76.0	66.6	100.0	100.0	100.0	45.0	40.4	100.0	95.0	64.6	100.0
claude-3.5-sonnet-latest	5.0	89.4	100.0	100.0	100.0	100.0	80.0	70.0	100.0	50.0	37.8	65.4	91.6	100.0	100.0
gpt-4.1	5.0	80.0	100.0	84.0	60.0	45.0	100.0	100.0	100.0	21.2	48.8	62.6	78.0	45.2	100.0
gpt-5-mini (thinking)	52.4	43.2	100.0	100.0	52.2	65.0	100.0	81.6	100.0	48.0	72.4	78.2	71.2	70.6	100.0
o1	21.4	80.8	100.0	100.0	82.2	91.6	96.6	100.0	100.0	37.4	29.8	62.4	35.0	28.4	100.0
gpt-4o	2.8	2.6	75.0	28.0	22.0	100.0	90.0	100.0	100.0	0.0	27.0	11.4	20.0	64.8	82.0
claude-3.5-haiku	40.0	72.8	33.0	86.0	76.6	58.2	100.0	100.0	100.0	0.0	18.4	34.0	40.4	46.8	100.0
Llama-3.1-405B-Instruct	3.2	2.2	42.0	86.0	20.0	45.0	50.0	100.0	100.0	17.6	21.0	11.6	12.0	20.0	100.0
gemini-2.0-flash	6.6	80.8	42.0	60.0	9.4	25.0	50.0	100.0	100.0	40.0	26.4	100.0	45.6	48.6	100.0
Owen3-32B	5.0	16.6	88.4	48.0	45.4	25.0	70.0	40.0	25.0	2.0	34.6	13.0	19.4	37.8	100.0
Llama-3.3-70B-Instruct	9.2	1.8	75.0	62.0	53.4	50.0	70.0	100.0	100.0	0.0	41.2	10.8	65.0	35.0	100.0
Llama-3.1-70B-Instruct	1.6	2.0	33.0	74.0	70.8	25.0	85.0	100.0	100.0	0.0	23.4	9.0	50.0	20.0	100.0
Owen2.5-72B-Instruct	0.0	0.0	8.0	100.0	10.0	36.6	5.0	100.0	70.0	0.0	13.2	6.0	20.0	20.0	60.0
Mistral-Large-Instruct-2407	2.4	16.6	73.2	90.0	64.6	40.0	70.0	85.0	55.0	0.0	30.2	10.0	5.6	20.0	56.0
gpt-4o-mini	2.2	16.4	48.2	48.0	42.0	100.0	35.0	100.0	25.0	9.6	35.4	23.6	22.0	47.8	74.0
gpt-4o-mini	0.8	0.8	29.8	28.0	16.8	0.0	0.0	63.4	10.0	0.0	20.8	6.6	4.0	28.4	5.0
Llama-4-Scout-17B-16E-Instruct	0.0	0.0	33.0	20.0	7.0	25.0	25.0	83.0	0.0	0.0	12.0	7.0	4.0	20.0	5.0
Llama-4-Scout-17B-16E-Instruct	1.8	0.8	33.0	88.0	60.0	23.4	21.8	78.2	21.8	0.0	32.0	7.0	15.4	30.2	13.0
Llama-4-Maverick-17B-128E-Instruct-FP8	0.0	0.0	33.0	88.0	7.0	23.4	21.8	8.0	0.0	0.0	21.0	7.0	0.0	8.0	10.0
Llama-4-Maverick-17B-128E-Instruct-2409	0.0	0.4	43.2	72.0	8.8	25.0	25.0	95.0	36.6	0.0	13.4	7.6	0.0	9.6	0.0
Mistral-Small-Instruct-2409	0.4	0.0	8.0	24.0	19.6	20.0	18.4	48.2	10.0	0.0	29.2	7.4	4.0	14.4	3.0
DeepSeek-R1	0.0	0.0	33.0	20.0	8.8	10.2	21.8	53.4	23.4	0.0	12.0	7.0	0.0	8.0	5.0
Owen2.5-7B-Instruct	0.0	0.0	29.8	22.0	8.8	1.6	4.8	0.0	25.0	0.0	11.0	6.0	0.0	0.0	0.0
Llama-3.2-8B-Instruct	0.0	0.0	23.2	0.0	7.0	13.6	10.0	0.0	0.0	0.0	82.4	7.0	4.0	20.0	5.0
phi-4	0.0	0.0	8.0	52.0	17.4	10.0	0.0	0.0	3.4	0.0	12.0	6.0	4.0	0.0	5.0
gpt-4.1-nano	0.0	0.0	28.0	20.0	7.6	10.2	10.0	0.0	3.4	0.0	13.0	6.0	2.4	0.0	5.0
Mistral-Small-24B-Instruct-2501	0.0	0.0	33.0	20.0	7.0	17.0	15.2	15.2	3.4	0.0	13.0	6.4	2.4	14.4	3.0
DeepSeek-R1-Distill-Llama-70B	0.0	0.0	8.0	22.0	8.2	10.2	13.6	10.2	8.4	0.0	12.0	6.0	0.8	8.0	5.0
Mistral-8B-Instruct-2410	0.0	0.0	8.0	20.0	7.0	0.0	0.0	0.0	0.0	0.0	11.0	7.0	0.0	0.0	0.0
Mistral-Small-3.1-24B-Instruct-2503	0.0	0.0	8.0	20.0	7.0	0.0	0.0	0.0	0.0	0.0	11.0	6.0	0.0	8.0	0.0
Mistral-8x22B-Instruct-v0.1	0.0	0.0	11.4	36.0	10.0	10.0	0.0	0.0	15.0	0.0	11.0	6.0	0.0	6.4	0.0
Llama-3.2-1B-Instruct	0.0	0.0	8.0	20.0	7.0	0.0	0.0	0.0	0.0	0.0	12.0	6.8	0.0	0.0	0.0
Phi-3-mini-128k-instruct	0.0	0.0	8.0	20.0	7.0	0.0	0.0	0.0	0.0	0.0	11.0	6.0	0.0	0.0	0.0
Phi-3-MoE-instruct	0.0	0.0	8.0	20.0	7.0	0.0	0.0	0.0	0.0	0.0	11.0	6.0	0.0	0.0	0.0
Phi-4-mini-instruct	0.0	0.0	8.0	20.0	7.0	0.0	0.0	0.0	0.0	0.0	11.0	6.0	0.0	0.0	0.0
Mistral-8x7B-Instruct-v0.1	0.0	0.0	8.0	20.0	7.0	0.0	0.0	0.0	0.0	0.0	11.0	6.0	0.0	0.0	0.0
Phi-3.5-mini-instruct	0.0	0.0	8.0	20.0	1.6	0.0	0.0	0.0	0.0	0.0	11.0	6.0	0.0	0.0	0.0
Phi-3-medium-128k-instruct	0.0	0.0	8.0	20.0	7.0	0.0	0.0	0.0	0.0	0.0	11.0	6.0	0.0	0.0	0.0

Table 15: Model Performance on SCIENCEWORLD tasks. Part 2.

Models	Included@Panel-PredictedNamedSurfaces	Included@Panel-PredictedUnknownSurfaces	Lifespan:LongerLived	Lifespan:LongerLived	Lifespan:ShorterLived	MeasureMatingPoint:KnownSubstance	MeasureMatingPoint:UnknownSubstance	Melt	MolecularGenetics:KnownPlant	MolecularGenetics:UnknownPlant	PowerComponent	PowerComponent:RenewableVsNonrenewableEnergy	TestConductivity	TestConductivity:OfUnknownSubstances	UseThermometer
o3 (medium)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	81.6	53.0	100.0	100.0	100.0
gpt-5 (thinking)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	77.0	53.0	100.0	100.0	100.0
claude-3.7-sonnet (thinking)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	77.0	52.0	100.0	100.0	100.0
claude-3.7-sonnet	100.0	100.0	100.0	100.0	100.0	81.4	77.4	76.0	83.4	51.0	82.2	33.2	70.6	100.0	100.0
claude-3.5-sonnet-latest	100.0	100.0	100.0	100.0	100.0	76.0	80.6	62.0	19.0	14.6	71.6	71.8	64.4	37.0	100.0
gpt-4.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	50.6	71.4	45.0	100.0	57.2	100.0
gpt-5-mini (thinking)	100.0	100.0	100.0	100.0	100.0	100.0	95.4	43.4	100.0	100.0	90.8	71.8	100.0	100.0	100.0
o1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	60.4	82.0	69.0	38.6	100.0	100.0	100.0
claude-3.5-haiku	100.0	100.0	100.0	100.0	100.0	78.4	89.2	59.6	47.8	13.6	57.0	38.0	100.0	61.8	100.0
Llama-3.1-405B-Instruct	100.0	100.0	100.0	100.0	100.0	90.0	9.0	4.0	83.2	100.0	33.0	28.8	17.0	58.8	37.6
gpt-5-mini	100.0	100.0	100.0	100.0	100.0	100.0	9.0	4.0	83.2	100.0	33.0	28.8	17.0	58.8	37.6
Qwen3-72B	100.0	100.0	100.0	100.0	100.0	81.8	90.6	22.0	100.0	100.0	27.0	11.6	8.0	58.0	42.8
Llama-3.3-70B-Instruct	100.0	100.0	100.0	100.0	100.0	81.6	7.4	4.4	100.0	100.0	27.0	31.8	25.0	10.0	68.2
Llama-3.1-70B-Instruct	80.0	60.0	100.0	100.0	100.0	62.6	34.6	2.6	100.0	65.2	47.0	3.0	8.6	12.8	38.4
Qwen3-30B	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	65.2	31.0	21.0	5.0	65.8	64.4	28.6
Mistral-Large-Instruct-2407	4.0	70.0	100.0	100.0	100.0	100.0	10.0	2.6	65.2	64.8	23.6	10.0	15.4	48.0	45.8
gpt-4.1-mini	35.0	18.0	70.0	100.0	90.0	28.0	70.4	2.6	64.8	64.8	23.6	10.0	15.4	48.0	45.8
gpt-5-mini	35.0	18.0	70.0	100.0	90.0	28.0	70.4	2.6	64.8	64.8	23.6	10.0	15.4	48.0	45.8
Llama-4-Maverick-17B-128E-Instruct-PPH	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
gpt-5-nano	3.0	5.0	90.0	53.0	33.0	50.0	31.0	1.2	0.0	0.0	58.2	31.4	81.0	8.0	18.2
Llama-3.1-8B-Instruct	0.0	43.0	40.0	0.0	50.0	6.0	1.0	1.0	0.0	0.0	7.0	38.0	5.0	10.0	0.0
DeepSeek-R1	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	2.6	100.0	7.0	5.0	5.0	0.0	9.2
gpt-5-nano	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3.2-3B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
phi-4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
gpt-4.1-nano	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DeepSeek-R1-Distill-Llama-70B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ministral-8B-Instruct-2410	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-Small-1.1-24B-Instruct-2503	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Nemotron-4-Base-Instruct-v0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3.2-1B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3-mini-128k-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3-mini-3.5B-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3-mini-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-8x7B-Instruct-v0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3.5-mini-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3-medium-128k-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 16: Model Performance on JERICHO games(tasks), part 1.

Models	905	Acorncount	Advent	Adventureland	Afflicted	Anchor	Awaken	Balances	Ballyhoo	Curves	Cutthroat	Deepphone	Detective	Dragon	Enchanter	Enter	Gold	Hbagg	Humbledark	Infidel	Inhumane	Jewel	Karn	Library	Loose	Loospg	Ludicorp
o3 (medium)	100.0	100.0	20.6	11.2	18.9	1.6	0.0	23.5	1.0	0.8	13.6	7.0	33.3	4.8	12.3	72.0	10.8	3.2	0.0	1.2	37.8	0.9	2.4	33.3	6.0	25.7	12.0
o3 (high)	100.0	100.0	20.7	14.0	17.3	2.0	0.0	27.5	2.0	0.7	13.6	10.0	54.4	6.4	6.0	74.0	9.6	3.2	0.0	1.2	28.9	0.4	4.7	40.7	7.6	31.4	12.3
gpt-5 (thinking)	100.0	100.0	27.7	9.8	26.4	2.4	0.0	27.5	6.0	1.1	12.0	10.3	38.3	4.8	12.8	65.0	12.0	2.5	0.0	1.2	35.6	1.1	5.3	39.3	6.0	40.0	11.2
o3 (low)	80.0	40.0	23.2	4.2	19.2	1.2	0.0	29.4	0.0	0.5	16.8	13.8	36.7	4.0	13.8	72.0	6.0	2.5	0.0	0.8	42.2	1.8	2.9	26.7	3.6	22.9	10.3
claude-3.7-sonnet (thinking)	40.0	26.7	17.6	1.4	36.0	0.8	0.0	25.5	2.0	0.3	16.8	8.1	76.7	4.8	13.2	58.0	6.0	2.5	0.0	0.5	31.1	0.0	1.8	6.7	6.8	20.0	8.8
claude-3.7-sonnet	0.0	6.7	19.4	0.0	35.5	1.6	4.0	39.2	0.0	0.0	20.0	8.8	88.9	4.0	13.8	66.0	0.0	2.5	0.0	1.0	40.0	0.0	2.9	33.3	0.0	28.6	8.1
claude-3.5-sonnet-latest	0.0	33.3	12.1	5.6	15.2	0.4	0.0	25.9	0.0	0.1	20.0	8.8	67.2	4.0	12.8	75.0	0.0	2.5	0.0	0.2	37.8	0.0	3.5	33.3	0.0	2.9	8.7
gpt-4.1	0.0	0.0	10.3	0.0	4.3	0.0	0.0	23.5	0.0	0.3	8.8	5.9	65.6	4.0	5.8	39.0	9.0	2.5	0.0	0.5	4.4	0.0	0.0	13.3	0.0	2.9	7.9
gpt-5-mini (thinking)	40.0	20.0	10.3	7.0	5.1	0.4	0.0	19.6	2.0	0.2	5.6	9.3	31.1	5.6	9.8	22.0	1.2	2.5	0.0	0.2	11.1	0.9	1.8	11.3	4.0	25.7	8.3
o1	80.0	66.7	13.9	1.4	1.1	0.0	0.0	23.9	0.0	0.0	9.6	6.7	27.2	4.8	12.5	58.0	0.0	2.5	0.0	1.0	4.4	0.0	1.8	34.7	0.0	14.3	7.9
gpt-4o	0.0	0.0	10.3	0.0	1.1	0.4	0.0	13.7	0.0	0.0	0.0	6.3	34.4	0.0	4.2	38.0	7.8	2.5	0.0	1.2	4.4	0.0	1.2	13.3	0.8	5.7	0.8
claude-3.5-haiku	0.0	20.0	10.3	0.0	0.0	0.0	0.0	19.6	0.0	0.0	2.4	5.7	25.0	0.8	3.2	23.0	0.0	2.5	0.0	0.0	0.0	0.4	0.0	0.0	0.0	11.4	2.3
claude-3.40B-Instruct	0.0	0.0	11.7	0.0	4.5	0.8	0.0	19.6	0.0	0.0	5.6	5.6	28.9	4.0	4.0	51.0	0.8	1.0	0.0	1.2	0.0	0.0	0.0	20.0	0.0	14.3	2.3
gemini-2.0-flash	0.0	0.0	10.3	0.0	1.6	0.0	0.0	19.6	0.0	0.0	5.6	5.6	28.9	4.0	4.0	51.0	0.8	1.0	0.0	0.0	0.0	0.0	0.0	20.0	0.0	14.3	2.3
Qwen2.5-72B-Instruct	0.0	6.7	10.3	0.0	3.9	0.8	0.0	19.6	2.0	0.0	1.0	4.5	31.1	3.2	3.8	31.0	0.8	2.5	0.0	0.0	0.0	0.4	2.4	3.3	4.0	11.4	3.2
Llama-3.3-70B-Instruct	0.0	0.0	10.3	0.0	0.0	1.2	0.0	11.8	0.0	0.0	0.0	5.3	36.1	4.8	0.0	23.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	33.3	0.0	14.3	2.7
Llama-3.1-70B-Instruct	0.0	0.0	10.3	0.0	0.8	0.0	0.0	19.6	0.0	0.0	5.6	5.6	28.9	4.8	0.0	22.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.3	0.9
Qwen2.5-72B-Instruct	0.0	0.0	10.3	0.0	1.6	0.0	0.0	11.8	0.0	0.0	0.0	4.5	33.3	8.0	0.0	41.0	0.0	2.5	0.0	0.2	11.1	0.0	0.0	0.0	0.0	14.3	1.3
Mistral-Large-Instruct-2407	60.0	0.0	10.3	0.0	0.0	0.0	0.0	13.7	0.0	0.0	0.0	6.2	41.7	0.0	6.5	14.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
gpt-4.1-mini	0.0	0.0	10.3	0.0	0.0	0.0	0.0	9.8	0.0	0.2	2.4	6.2	3.4	8.3	0.0	2.0	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
gpt-4o-mini	0.0	0.0	10.3	0.0	0.0	0.0	0.0	9.8	0.0	0.0	3.2	3.4	8.3	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
Llama-4-Scout-17B-16E-Instruct	0.0	0.0	10.3	0.0	2.7	0.0	0.0	0.0	0.0	0.0	0.0	1.2	12.2	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
gpt-5-nano	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
Llama-4-Maverick-17B-128E-Instruct-FP8	0.0	0.0	10.3	0.0	0.0	0.0	0.0	5.9	0.0	0.0	7.2	4.1	8.3	0.0	5.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
Mistral-Small-Instruct-2409	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	8.3	3.2	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
Llama-3.1-8B-Instruct	0.0	0.0	10.3	0.0	0.5	0.0	0.0	15.7	0.0	0.0	0.8	4.3	15.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
DeepSeek-R1	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.3	11.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3
Qwen2.5-7B-Instruct	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.6	2.6	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9
Llama-3.2-3B-Instruct	0.0	0.0	10.3	0.0	0.0	0.0	0.0	3.9	0.0	0.0	1.6	0.3	8.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3
phi-4	0.0	0.0	10.3	0.0	0.0	0.0	0.0	9.8	0.0	0.0	0.0	2.7	8.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.7	0.0	0.0	0.7
gpt-4.1-nano	0.0	0.0	10.3	0.0	0.0	0.0	0.0	7.8	0.0	0.0	3.2	1.7	18.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
Mistral-Small-24B-Instruct-2501	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.6	3.2	13.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
DeepSeek-R1-Distill-Llama-70B	0.0	0.0	10.3	0.0	0.8	0.0	0.0	7.8	0.0	0.0	1.6	1.7	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
Mistral-8B-Instruct-2410	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	11.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
Mistral-Small-3.1-24B-Instruct-2503	0.0	0.0	10.3	0.0	0.0	0.0	0.0	9.8	0.0	0.0	0.0	0.3	5.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
Mistral-8x22B-Instruct-v0.1	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	6.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
Llama-3.2-1B-Instruct	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	3.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
Phi-3-mini-128k-instruct	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
Phi-3.5-MoE-instruct	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
Phi-4-mini-instruct	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
Mistral-8x7B-Instruct-v0.1	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
Phi-3.5-mini-instruct	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
Phi-3-medium-128k-instruct	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7

Table 17: Model Performance on JERICHO games(tasks), part 2.

Models	Lurking	Moonlit	Murder	Night	Omniquest	Partyfoul	Pentari	Planetfall	Plundered	Reverb	Seastalker	Sherlock	Snacktime	Sorcerer	Spellbkr	Spirit	Temple	Trinity	Trys205	Weapon	Wishbringer	Yomomma	Zenon	Zork1	Zork2	Zork3	Zhu
o3 (medium)	5.0	0.0	9.8	24.0	26.0	0.0	7.1	7.5	2.4	30.0	4.0	11.4	40.0	7.0	9.2	0.6	16.0	12.0	3.4	0.0	13.2	2.9	0.0	15.5	0.0	42.9	2.0
o3 (high)	5.0	0.0	9.0	16.0	14.0	0.0	7.1	7.5	3.2	28.8	8.0	11.6	40.0	7.0	10.3	1.1	14.3	13.4	4.0	0.0	15.2	2.9	0.0	13.3	3.0	40.0	5.0
gpt-5 (thinking)	5.0	20.0	12.5	12.0	10.0	0.0	15.7	7.5	2.4	34.0	11.0	9.4	60.0	9.2	10.0	1.3	21.1	13.0	3.7	0.0	13.0	1.7	0.0	17.8	1.0	34.3	6.0
o3 (low)	5.0	0.0	6.8	24.0	14.0	0.0	12.9	7.5	3.2	24.0	11.2	9.2	40.0	7.0	10.7	1.3	16.0	14.2	0.9	0.0	13.4	2.9	0.0	14.4	0.0	48.6	3.0
claude-3.7-sonnet (thinking)	5.0	0.0	11.3	20.0	10.0	0.0	7.1	4.5	4.0	20.6	18.6	10.6	36.0	5.3	11.5	1.0	14.3	11.8	0.6	0.0	11.6	0.0	0.0	12.2	0.0	40.0	5.0
claude-3.7-sonnet	5.0	0.0	10.8	20.0	10.0	0.0	7.1	3.8	2.4	4.0	20.6	10.6	32.0	9.2	11.7	1.4	14.3	13.8	0.3	0.0	13.0	0.0	0.0	13.5	0.0	42.9	5.0
claude-3.5-sonnet-latest	5.0	0.0	8.7	0.0	10.0	0.0	10.0	6.8	6.4	0.0	7.8	8.2	24.0	7.2	11.7	1.4	2.9	0.0	0.3	0.0	13.8	0.0	0.0	11.7	4.0	40.0	6.0
gpt-4.1	5.0	0.0	8.7	8.0	10.0	0.0	7.5	6.8	6.4	0.0	7.8	8.2	24.0	7.2	11.7	0.5	16.0	9.2	0.0	0.0	13.8	1.1	0.0	17.6	1.0	32.4	4.0
gpt-5-mini (thinking)	5.0	0.0	8.7	8.0	10.0	0.0	7.1	2.2	2.4	6.8	10.4	7.2	40.0	4.2	6.7	1.1	11.4	9.4	0.6	0.0	13.8	2.3	0.0	7.0	1.0	42.9	9.6
gpt-4o	7.0	0.0	8.6	0.0	12.0	0.0	8.6	3.8	4.0	8.8	3.8	0.4	40.0	7.8	9.2	0.0	11.4	9.4	1.7	0.0	13.8	0.6	0.0	12.5	1.0	42.9	9.6
gpt-4o	5.0	0.0	5.6	0.0	10.0	0.0	8.6	7.5	3.2	0.0	3.2	8.8	20.0	3.2	8.0	0.8	0.0	9.6	0.9	0.0	7.4	0.0	0.0	14.5	0.0	31.4	0.0
claude-3.5-haiku	5.0	0.0	5.3	0.0	10.0	0.0	5.7	6.8	2.4	14.0	10.8	0.0	8.0	2.0	4.2	0.3	0.0	9.0	0.0	0.0	11.8	1.7	0.0	13.1	0.0	31.4	0.0
claude-3.1-405B-Instruct	5.0	0.0	5.2	20.0	10.0	0.0	7.1	3.8	0.0	10.6	10.6	2.6	24.0	1.2	4.2	1.3	14.3	4.0	0.0	0.0	6.0	0.0	0.0	11.1	0.0	28.6	13.0
gemini-2.0-flash	5.0	0.0	4.4	0.0	8.0	0.0	4.3	3.8	1.6	0.0	8.2	10.4	40.0	1.2	6.7	1.3	0.0	4.8	0.0	0.0	11.6	0.0	0.0	9.4	0.0	25.7	1.0
Owen3-32B	3.0	0.0	4.4	0.0	0.0	0.0	5.7	0.0	4.0	0.0	17.4	2.4	0.0	1.2	4.2	0.0	0.0	4.4	0.0	0.0	7.4	0.0	0.0	9.4	0.0	42.9	4.0
Llama-3.3-70B-Instruct	2.0	0.0	5.6	0.0	10.0	0.0	2.9	3.8	4.0	0.0	21.0	3.8	0.0	1.7	4.2	0.0	0.0	8.4	0.0	0.0	11.6	0.0	0.0	4.3	0.0	31.4	3.0
Llama-3.1-70B-Instruct	2.0	0.0	5.6	0.0	10.0	0.0	4.3	5.3	1.6	0.0	9.6	2.2	4.0	1.2	4.2	0.0	0.0	16.0	8.4	0.0	6.0	0.0	0.0	3.1	0.0	42.9	13.0
Qwen2.5-72B-Instruct	5.0	0.0	4.6	4.0	10.0	0.0	0.0	3.8	0.0	0.0	9.8	3.6	16.0	1.2	4.2	0.0	0.0	5.0	0.0	0.0	1.0	0.0	0.0	2.3	0.0	28.6	0.0
Mistral-Large-Instruct-2407	5.0	0.0	5.6	0.0	0.0	0.0	2.9	0.8	0.0	0.0	3.2	8.8	24.0	2.0	7.2	1.6	0.0	2.8	0.0	0.0	6.0	0.0	0.0	12.8	0.0	28.6	0.0
gpt-4.1-mini	5.0	0.0	5.6	0.0	10.0	0.0	0.0	4.5	0.8	0.0	2.8	1.6	0.0	1.2	7.0	0.3	0.0	3.0	0.0	0.0	6.4	0.6	0.0	6.9	0.0	28.6	1.0
gpt-4o-mini	4.0	0.0	5.6	0.0	10.0	0.0	0.0	2.2	0.0	0.0	3.0	1.4	0.0	1.7	3.3	0.3	0.0	7.4	0.0	0.0	6.0	0.0	0.0	3.1	2.0	5.7	0.0
Llama-4-Scout-17B-16E-Instruct	5.0	0.0	0.0	0.0	0.0	0.0	0.0	3.8	0.0	0.0	3.0	0.0	0.0	1.2	4.2	0.0	0.0	4.0	0.0	0.0	6.0	0.0	0.0	1.4	0.0	42.9	0.0
gpt-5-nano	4.0	0.0	0.0	0.0	2.0	0.0	0.0	3.0	0.0	0.0	2.6	1.4	0.0	1.2	5.2	0.2	0.0	3.8	0.0	0.0	6.0	0.0	0.0	0.6	0.0	28.6	0.0
Llama-4-Maverick-17B-128E-Instruct-PP8	0.0	0.0	1.3	12.0	10.0	0.0	4.3	3.8	0.0	0.0	3.0	1.2	0.0	1.2	4.7	0.0	0.0	0.6	0.0	0.0	6.0	0.0	0.0	2.0	0.0	0.0	0.0
Mistral-Small-Instruct-2409	0.0	0.0	0.0	0.0	10.0	0.0	7.1	3.8	0.0	0.0	2.0	0.0	0.0	1.2	4.7	0.0	0.0	1.0	0.0	0.0	6.0	0.0	0.0	1.4	0.0	0.0	0.0
Llama-3.1-8B-Instruct	3.0	0.0	3.4	0.0	10.0	0.0	0.0	3.0	0.0	0.0	3.0	0.0	0.0	1.2	3.3	0.2	8.6	1.0	0.0	0.0	6.0	0.0	0.0	2.0	0.0	17.1	2.0
DeepSeek-R1	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	2.9	0.0	0.0	0.0
Qwen2.5-7B-Instruct	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	2.6	0.0	0.0	1.2	0.0	0.0	0.0	4.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3.2-3B-Instruct	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	0.0	1.2	4.2	0.0	0.0	4.0	0.0	0.0	6.0	0.0	0.0	1.4	0.0	11.4	0.0
phi-4	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.8	14.3	0.0	0.0	0.0	6.0	0.0	0.0	0.6	0.0	0.0	0.0
gpt-4.1-nano	1.0	0.0	0.0	0.0	10.0	0.0	0.0	3.0	0.0	0.0	2.6	102.0	0.0	1.2	0.0	0.2	0.0	0.0	0.0	0.0	6.0	0.0	0.0	2.0	0.0	22.9	0.0
Mistral-Small-24B-Instruct-2501	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	1.2	0.0	0.2	5.7	1.6	0.0	0.0	4.0	0.0	0.0	0.0	0.0	14.3	0.0
DeepSeek-R1-Distill-Llama-70B	1.0	0.0	0.2	0.0	10.0	0.0	0.0	0.0	0.0	0.0	2.2	0.0	0.0	1.2	0.0	0.0	2.9	0.0	0.0	0.0	4.0	0.0	0.0	2.9	0.0	14.3	0.0
Mistral-8B-Instruct-v0.1	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-8x22B-Instruct-v0.1	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3.2-7B-Instruct-v0.1	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.3	0.0	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama-3.2-128K-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3-mini-128K-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3.5-MoE-Instruct	0.0	0.0	0.0	0.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0
Phi-4-mini-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mistral-8x7B-Instruct-v0.1	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	2.4	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3.5-mini-Instruct	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3-medium-128K-Instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

## K AVERAGE SIMON SAYS SCORE VERSUS OVERALL TALES SCORE

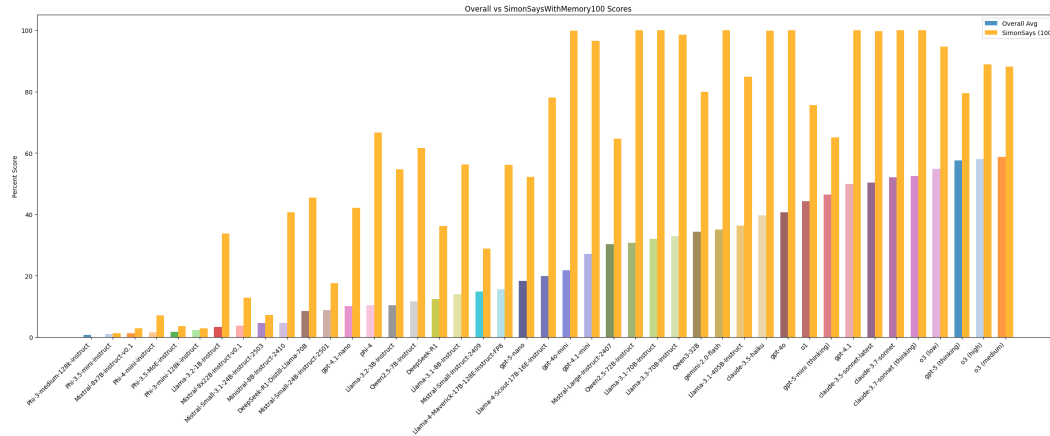


Figure 2: All TALES and average SIMON SAYS scores for each model, sorted by TALES performance. We see that an increase in performance in SIMON SAYS typically correlates with an increase in performance for TALES overall.

## L COMPUTE

All experiments were run intermittently over a course of roughly six months. Open-weight models were run on a combination of a four node cluster of 8xMI300s and one node of 8xA100s. Anthropic API experiments accrued a cost of 1,562.15 U.S. dollars. OpenAI API experiments accrued an estimated cost of 6,870.76 U.S. dollars.

## M 1000 STEPS OF ZORK1

While even the best performing LLMs make reasoning mistakes, we find what allows them to still find success in TALES is the ability to both avoid making an excessive number of these mistakes and the ability to self-correct. We argue that 100 steps are sufficient to evaluate the performance of current state-of-the-art LLMs because even the best overall LLM, o3, fails to approach the maximum possible score within 100 steps for ZORK1 (13.8% vs 29.1%). In this experiment, we explore whether any of the top models can achieve a score comparable to the walkthrough after 100 steps in ZORK1 while allowed to run for 1000 steps with the entire history kept in the context. If so, we examine the behaviors that enable this and determine the required number of steps. We select the overall top 3 performing models, o3 and both the thinking and non-thinking modes of Claude-3.7-Sonnet.

**Scores improve slightly, but the best LLMs are still far from the walkthrough score even with 10 times more steps.** That is with 1000 steps, the best LLMs fail to reach 29.3% of the total score. o3 manages to achieve a score of 20.9%, a performance increase of only 7.6% over its original score of 13.3% for 10 times the steps. Claude-3.7 non-thinking<sup>5</sup> and thinking achieve 16.9% and 15.3% respectively. The key behavior pattern we see in both thinking models is a slightly directed, random exploration of the area of the game past the bottleneck that stops other, weaker LLMs. This exploration is far less focused than the iterative search we see agents perform early in the game and in simpler environments such as AW.

We tried allowing Claude-3.7-Sonnet to think for up to 4096 tokens. However, the model never uses more than 700 tokens for its thinking, a similar value seen for the rest of the benchmark where the thinking effort is capped to 1024. This is a significantly smaller thinking effort than o3 which uses up to 5000 thinking tokens throughout its 1000 steps ZORK1 walkthrough. This suggests that o3’s

<sup>5</sup>Despite scoring higher than its thinking variant, zero-shot Claude-3.7-Sonnet suffers a catastrophic inductive reasoning failure by repeatedly issuing the quitting commands after step 479.

performance is due to a willingness to leverage many more thinking tokens at any particular step. However, the highest thinking efforts do not appear to occur at any significant points during gameplay and we are unable to verify the actual contents since we do not have access to the thinking traces.

## N FRAMEWORK ENVIRONMENT SUBSELECTION

### N.1 TEXTWORLD

For TEXTWORLD, we use the following environments:

```
test/difficulty_level_1/tw-cooking-recipe1+take1+open-0nQyHWbv6dXFPmHLKX.z8
  ↪ XFPmHLKX.z8
test/difficulty_level_2/tw-cooking-recipe1+take1+cook+open-0nQyHWbv6dXFPmHLKX.z8
  ↪ bv6dXFPmHLKX.z8
test/difficulty_level_3/tw-cooking-recipe1+take1+cut+open-0nQyHWbv6dXFPmHLKX.z8
  ↪ vh6dXFPmHLKX.z8
test/difficulty_level_4/tw-cooking-recipe1+take1+open+go6-0nQyHWbv6dXFPmHLKX.z8
  ↪ vh6dXFPmHLKX.z8
test/difficulty_level_5/tw-cooking-recipe1+take1+open+go9-0nQyHWbv6dXFPmHLKX.z8
  ↪ vh6dXFPmHLKX.z8
test/difficulty_level_6/tw-cooking-recipe1+take1+open+go12-0nQyHWbv6dXFPmHLKX.z8
  ↪ bv6dXFPmHLKX.z8
test/difficulty_level_7/tw-cooking-recipe1+take1+cook+cut+open-0nQyHWbv6dXFPmHLKX.z8
  ↪ QyHWbv6dXFPmHLKX.z8
test/difficulty_level_8/tw-cooking-recipe3+take3+open+go6-0nQyHWbv6dXFPmHLKX.z8
  ↪ vh6dXFPmHLKX.z8
test/difficulty_level_9/tw-cooking-recipe3+take3+cook+cut+open+go6-0nQyHWbv6dXFPmHLKX.z8
  ↪ 6-0nQyHWbv6dXFPmHLKX.z8
test/difficulty_level_10/tw-cooking-recipe3+take3+cook+cut+open+go12-0nQyHWbv6dXFPmHLKX.z8
  ↪ o12-0nQyHWbv6dXFPmHLKX.z8
```

### N.2 TEXTWORLDEXPRESS

For TEXTWORLDEXPRESS, we use the game parameters:

```
TASKS = [
  (
    "CookingWorld",
    "cookingworld",
    "numLocations=1, numIngredients=2, numDistractorItems=5,
    ↪ includeDoors=0, limitInventorySize=0",
  ),
  (
    "TextWorldCommonsense",
    "twc",
    "numLocations=1, numItemsToPutAway=1, includeDoors=0, limitInventorySize=0",
    ↪ nventorySize=0",
  ),
  (
    "CoinCollector",
    "coin",
    "numLocations=1, numDistractorItems=5,
    ↪ limitInventorySize=0",
  ),
  ("Arithmetic", "arithmetic", ""),
  (
    "MapReader",
    "mapreader",
  ),
]
```

```

1512         "numLocations=2, maxDistanceApart=1,
1513         ↪ maxDistractorItemsPerLocation=2, includeDoors=0,
1514         ↪ limitInventorySize=0",
1515     ),
1516     ("Sorting", "sorting", ""),
1517     ("SimonSays10", "simonsays", "gameLength=10, numDistractors=4,
1518     ↪ memorization=0"),
1519     ("SimonSays50", "simonsays", "gameLength=50, numDistractors=4,
1520     ↪ memorization=0"),
1521     ("SimonSays100", "simonsays", "gameLength=100,
1522     ↪ numDistractors=4, memorization=0"),
1523     (
1524         "SimonSaysWithMemory10",
1525         "simonsays",
1526         "gameLength=10, numDistractors=4, memorization=1,
1527         ↪ verbose=0",
1528     ),
1529     (
1530         "SimonSaysWithMemory50",
1531         "simonsays",
1532         "gameLength=50, numDistractors=4, memorization=1,
1533         ↪ verbose=0",
1534     ),
1535     (
1536         "SimonSaysWithMemory100",
1537         "simonsays",
1538         "gameLength=100, numDistractors=4, memorization=1,
1539         ↪ verbose=0",
1540     ),
1541     (
1542         "SimonSaysWithMemory10Verbose",
1543         "simonsays",
1544         "gameLength=10, numDistractors=4, memorization=1,
1545         ↪ verbose=1",
1546     ),
1547     (
1548         "SimonSaysWithMemory50Verbose",
1549         "simonsays",
1550         "gameLength=50, numDistractors=4, memorization=1,
1551         ↪ verbose=1",
1552     ),
1553     (
1554         "SimonSaysWithMemory100Verbose",
1555         "simonsays",
1556         "gameLength=100, numDistractors=4, memorization=1,
1557         ↪ verbose=1",
1558     ),
1559     ("PeckingOrder", "peckingorder", ""),
1560 ]

```

### N.3 ALFWORLD

The 12 games for ALFWORLD. Note that these are from when the "-game-seed" is not set. Changing this value would cause the games to change.

```

1563 valid_seen/pick_and_place_simple-Book-None-SideTable-329/trial_T2
1564 ↪ 0190908_050633_745514
1565 valid_seen/look_at_obj_in_light-AlarmClock-None-DeskLamp-323/tria
1566 ↪ 1_T20190909_044715_250790

```



valid\_seen/pick\_clean\_then\_place\_in\_recep-ButterKnife-None-CounterTop-8/trial\_T20190909\_105559\_983897  
 ↪ rTop-8/trial\_T20190909\_105559\_983897  
 valid\_seen/pick\_heat\_then\_place\_in\_recep-Apple-None-DiningTable-2\_6/trial\_T20190907\_060234\_011675  
 ↪ 6/trial\_T20190907\_060234\_011675  
 valid\_seen/pick\_cool\_then\_place\_in\_recep-Apple-None-CounterTop-14\_10/trial\_T20190909\_044933\_815840  
 ↪ /trial\_T20190909\_044933\_815840  
 valid\_seen/pick\_two\_obj\_and\_place-AlarmClock-None-Dresser-305/trial\_T20190907\_165826\_194855  
 ↪ al\_T20190907\_165826\_194855  
 valid\_unseen/pick\_and\_place\_simple-Mug-None-Desk-308/trial\_T20190908\_125200\_737896  
 ↪ 908\_125200\_737896  
 valid\_unseen/look\_at\_obj\_in\_light-AlarmClock-None-DeskLamp-308/trial\_T20190908\_222917\_366542  
 ↪ ial\_T20190908\_222917\_366542  
 valid\_unseen/pick\_clean\_then\_place\_in\_recep-Bowl-None-Cabinet-10\_6/trial\_T20190909\_061130\_844814  
 ↪ trial\_T20190909\_061130\_844814  
 valid\_unseen/pick\_heat\_then\_place\_in\_recep-Apple-None-Fridge-10\_182259\_116320  
 ↪ rial\_T20190906\_182259\_116320  
 valid\_unseen/pick\_cool\_then\_place\_in\_recep-Bread-None-CounterTop-10\_10/trial\_T20190908\_091747\_866951  
 ↪ 10/trial\_T20190908\_091747\_866951  
 valid\_unseen/pick\_two\_obj\_and\_place-CD-None-Safe-308/trial\_T20190907\_050942\_897916  
 ↪ 907\_050942\_897916

#### N.4 SCIENCEWORLD

We use the first variation of the test set for each of the 30 tasks.

## O ANNOTATED LOGS: CRITERIA AND EXAMPLES

### O.1 LABELING CRITERIA

For spatial reasoning failures, we primarily label those actions that involve failing navigation through path finding, backtracking, or iteratively searching spaces for necessary elements for completing the assigned task. Deductive reasoning failures were primarily actions that involved interaction with game elements that cannot be reasonably linked back to the target objective or failures to solve in-game puzzles. Inductive reasoning failures were comprised mostly of actions that the agent had previously attempted in a near-identical state to negative feedback. If there was any non-trivial change in game state compared to when the action was first attempted, we did not label the action as an inductive reasoning failure if the action did not fail due to a parser issue. Grounded reasoning failures encompassed what is commonly referred to as hallucination: attempts to interact with game elements not mentioned in observations or feedback.

### O.2 SPATIAL REASONING FAILURE EXAMPLE

Task Description:Your task is to melt lead. First, focus on the substance. Then, take actions that will cause it to change its state of matter...

...

This room is called the kitchen. In it, you see:

the agent  
 a substance called air  
 a chair. On the chair is: nothing.  
 a counter. On the counter is: a bowl (containing a red  
 ↪ apple, a banana, an orange, a potato), a drawer.  
 a cupboard. The cupboard door is closed.  
 a freezer. The freezer door is closed.  
 a fridge. The fridge door is closed.  
 a glass jar (containing a substance called sodium  
 ↪ chloride)

1620 a lighter  
 1621 a oven, which is turned off. The oven door is closed.  
 1622 a painting  
 1623 a sink, which is turned off. In the sink is: nothing.  
 1624 a substance called soap  
 1625 a stopwatch, which is deactivated.  
 1626 a stove, which is turned off. On the stove is: nothing.  
 1627 a table. On the table is: a glass cup (containing  
 1628 ↪ nothing).  
 1629 a thermometer, currently reading a temperature of 10  
 1630 ↪ degrees celsius  
 1631 You also see:  
 1632 A door to the bathroom (that is open)  
 1633 A door to the hallway (that is closed)  
 1634 A door to the outside (that is closed)  
 1635 >  
 1636 Action: open oven  
 1637 ...  
 1638 Action: go hallway  
 1639 ...  
 1640 ...  
 1641 Action: go workshop  
 1642 ...  
 1643 ...  
 1644 This room is called the workshop. In it, you see:  
 1645 the agent  
 1646 a substance called air  
 1647 a table. On the table is: a battery, a green wire, a metal  
 1648 ↪ pot (containing a substance called lead), a orange  
 1649 ↪ light bulb, which is off, a orange wire, a red wire, a  
 1650 ↪ switch, which is off, a violet light bulb, which is  
 1651 ↪ off, a yellow light bulb, which is off.  
 1652 a ultra low temperature freezer. The ultra low temperature  
 1653 ↪ freezer door is closed.  
 1654 You also see:  
 1655 A door to the hallway (that is open)  
 1656 >  
 1657 Action: pick up metal pot  
 1658 You move the metal pot to the inventory.  
 1659 Action: go kitchen  
 1660  
 1661  
 1662  
 1663 Above we provide an example of a spatial reasoning failure from Claude-3.5-Haiku in ScienceWorld-  
 1664 Melt. The agent navigates to the workshop, through the hallway and thus should realize the path is  
 1665 kitchen to hallway to workshop. However, despite this, the agent attempts the action **go kitchen**.  
 1666  
 1667 **O.3 DEDUCTIVE REASONING FAILURE EXAMPLE**  
 1668  
 1669 Task Description: Your task is to freeze lead. First, focus on the  
 1670 ↪ substance. Then, take actions that will cause it to change its  
 1671 ↪ state of matter.  
 1672  
 1673 ...

1674 Action: look around  
 1675  
 1676 This room is called the kitchen. In it, you see:  
 1677     the agent  
 1678     a substance called air  
 1679     a chair. On the chair is: nothing.  
 1680     a counter. On the counter is: a bowl (containing a red  
 1681     ↪ apple, a banana, an orange, a potato), a drawer.  
 1682     a cupboard. The cupboard door is closed.  
 1683     a freezer. The freezer door is closed.  
 1684     a fridge. The fridge door is closed.  
 1685     a glass jar (containing a substance called sodium  
 1686     ↪ chloride)  
 1687     a lighter  
 1688     a oven, which is turned off. The oven door is closed.  
 1689     a painting  
 1690     a sink, which is turned off. In the sink is: nothing.  
 1691     a substance called soap  
 1692     a stopwatch, which is deactivated.  
 1693     a stove, which is turned off. On the stove is: nothing.  
 1694     a table. On the table is: a glass cup (containing  
 1695     ↪ nothing).  
 1696     a thermometer, currently reading a temperature of 10  
 1697     ↪ degrees celsius  
 1698 You also see:  
 1699     A door to the bathroom (that is open)  
 1700     A door to the hallway (that is closed)  
 1701     A door to the outside (that is closed)  
 1702  
 1703 Action: open freezer  
 1704  
 1705 Above we provide an example of a spatial reasoning failure from Claude-3.7-Sonnet (zero-shot)  
 1706 in ScienceWorldFreeze. While looking for lead, the agent checks a kitchen freezer, an extremely  
 1707 unlikely place to find lead.  
 1708  
 1709 O.4 INDUCTIVE REASONING FAILURE EXAMPLE  
 1710  
 1711 ...  
 1712 If you pray enough, your prayers may be answered.  
 1713 Action: pray  
 1714  
 1715 If you pray enough, your prayers may be answered.  
 1716 Action: pray  
 1717  
 1718 If you pray enough, your prayers may be answered.  
 1719 Action: pray  
 1720  
 1721 If you pray enough, your prayers may be answered.  
 1722 Action: pray  
 1723  
 1724 If you pray enough, your prayers may be answered.  
 1725 Action: pray  
 1726  
 1727 ...

Above we provide an example of an inductive reasoning failure from Claude-3.5-Haiku in ZORK1. While the game environment returns a signal that praying might result in some change, Claude-3.5-Haiku continues to do this for 36 steps, far past when most reasonable players would have moved on.

## O.5 GROUNDED REASONING FAILURE EXAMPLE

This room is called the workshop. In it, you see:

the agent  
a substance called air  
a table. On the table is: a battery, a green wire, a  
→ orange light bulb, which is off, a orange wire, a red  
→ wire, a switch, which is off, a violet light bulb,  
→ which is off, a yellow light bulb, which is off.  
a ultra low temperature freezer. The ultra low temperature  
→ freezer door is closed.

You also see:

A door to the hallway (that is open)

...

Action: go to hallway

You move to the hallway.

...

Action: go to hallway

You move to the hallway.

Action: go to kitchen

You move to the kitchen.

Action: use lighter on lead

The lighter heats up the lead a small amount.

Action: take red wire

Above we provide an example of an inductive reasoning failure from Claude-3.7-Sonnet (Thinking) in ScienceWorldBoil. We see the agent attempt to interact with an item in its context despite having moved to another location where the item is no longer present.

## P DATA CONTAMINATION EVIDENCE FROM THINKING TRACES: EXAMPLES

The blue glow of my sword suggests there's a danger nearby, likely  
→ something from Zork lore (like a grue). Let me explore the  
→ available paths carefully. I have the lantern on, which should  
→ protect me from grues in dark areas. Let me try going north  
→ first.

An example of a thinking trace from the Claude-3.7-Sonnet thinking mode that leads us to strongly suspect data contamination. Most notably, the thinking traces references a *grue*, a creature from ZORK1 mythology that had not at all been previously mentioned in the game's history.

## Q HYPERPARAMETERS

The following is the command used to run experiments across 5 seeds. All arguments not included are default values.

```
for i in {{1..5}}; do
  python benchmark.py --agent agents/llm.py zero-shot --env
    ↪ jericho scienceworld textworld textworld_express alfworld
    ↪ --llm $LLM          --context 100 --nb-steps 100 --conversation
    ↪ --wandb --seed "20241106${(i)}" &
  sleep 60
done
```