

000 TALES: TEXT ADVENTURE LEARNING ENVIRONMENT 001 SUITE

002 **Anonymous authors**

003 Paper under double-blind review

004 ABSTRACT

005 Reasoning is an essential skill to enable Large Language Models (LLMs) to interact
006 with the world. As tasks become more complex, they demand increasingly sophis-
007 ticated and diverse reasoning capabilities for sequential decision-making, requiring
008 structured reasoning over the context history to determine the next best action.
009 We introduce TALES, a diverse collection of synthetic and human-written text-
010 adventure games designed to challenge and evaluate diverse reasoning capabilities.
011 We present results over a range of LLMs, open- and closed-weights, performing a
012 qualitative analysis on the top performing models. Despite an impressive showing
013 on synthetic games, even the top LLM-driven agents fail to achieve 20% on games
014 designed for human enjoyment. Visualization of the experiments can be found at
015 <https://github.com/tale-suite/tale-suite-anonymized>.

016 1 INTRODUCTION

017 Reasoning is crucial in sequential decision-making tasks where optimal actions depend on previous
018 choices whose effects may only emerge later. In complex tasks, the agent often needs to leverage a
019 variety of reasoning skills to make the best decision. This becomes more challenging in grounded en-
020 vironments, where the causal constraints between actions are fixed and cannot be violated. Therefore,
021 the ability of a Large Language Model (LLM) to perform this structured thinking and follow these
022 constraints across long contexts is critical for real-world application (Trivedi et al., 2024).

023 Through the lens of the Problem Space Hypothesis for complex problem solving (Newell, 1979), we
024 identify four core reasoning skills vital to an LLM-driven agent’s ability to interface with applications
025 in real-world settings where there is limited human intervention: **Deductive reasoning**, to act upon
026 general principles (Johnson-Laird, 1999); **Inductive reasoning**, to draw conclusions from interaction
027 and observation (Heit, 2000); **Spatial reasoning**, to efficiently navigate and understand the spatial
028 relationship between objects (Byrne & Johnson-Laird, 1989); and **Grounded reasoning**, to identify
029 relevant information and perform admissible actions in a given context (Endsley et al., 2000). For
030 an LLM agent to be successful in real-world applications, it must be able to continuously use and
031 combine these core reasoning skills at every step.

032 Text-adventure games are a prime test-bed to evaluate an LLM agent’s ability to reason in these
033 modes due to the need **to** apply commonsense principles to perform directed exploration (Deductive),
034 discover implicit dynamics through trial and error (Inductive), and operate in a situated environment
035 (Spatial and Grounded). Figure 1 illustrates an agent navigating through a text-adventure game. At
036 each step, one or more of the core reasoning skills may be required for optimal decision making,
037 while a single failure in any reasoning skill can dramatically reduce overall performance. Success
038 in these environments require the consistent and compositional use of the core reasoning skills,
039 mirroring the challenges faced by LLM agents in real-world applications. While previous works have
040 explored using text-adventure games as a metric of an LLM agent’s capabilities, they either focus on
041 one specific framework, provide significant scaffolds for the LLM, or drastically change the scope of
042 the original task (Paglieri et al., 2024; Chang et al., 2024).

043 To evaluate an LLM agent’s comprehensive reasoning capabilities, we introduce TALES, the first
044 benchmark that unifies TEXTWORLD (TW), TEXTWORLDEXPRESS (TWX), ALFWORLD (AW),
045 SCIENCEWORLD (SW) and JERICHO in their canonical forms. Unlike other benchmarks, we
046 remove all environment-specific scaffolding in favor of a standardized evaluation. This creates a
047 challenging and comprehensive evaluation suite of 122 tasks for better understanding the agent’s
048

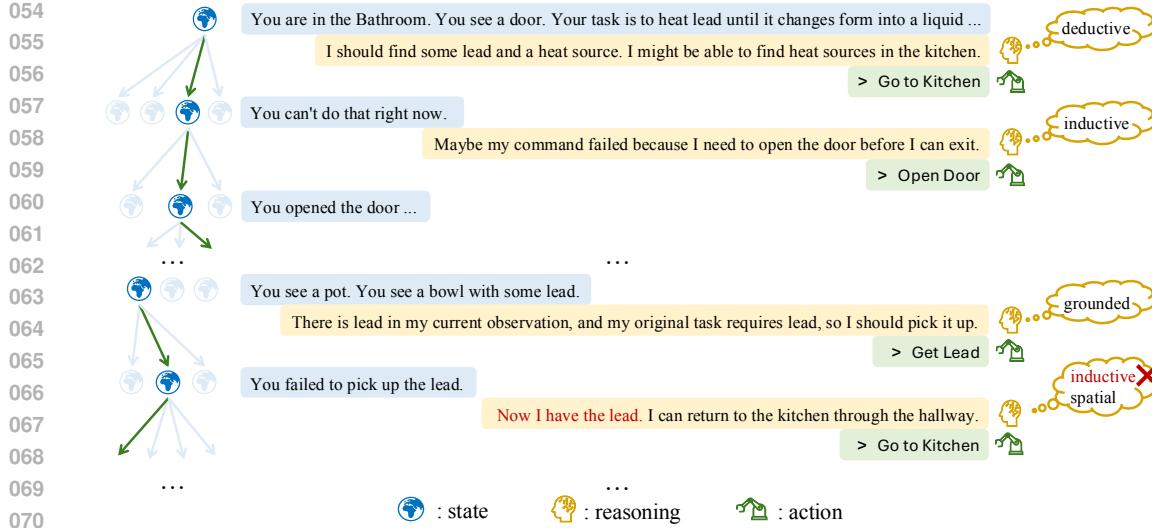


Figure 1: Example of a gameplay trajectory presenting the conversation between the game engine and an agent. We additionally fabricate the agent's reasoning to demonstrate the reasoning types this work concerns, detailed in Section 2. Here, the agent made a **mistake** in its inductive reasoning, which further caused the generation of a sub-optimal action.

baseline composite reasoning skills. We introduce TALES by following the ECB framework (Liu et al., 2024) that outlines the key benchmark design decisions.

As an initial litmus test of an agent's compositional reasoning skills, we introduce the game **SIMON SAYS** to assess whether the agent has the baseline capabilities required to challenge TALES. In this classic children's game, players must follow instructions only when prefaced with "Simon says" - making it fundamentally an instruction-following task. The simplest formulation of our text-adventure implementation gives the player a direct walkthrough of required actions, similar to the iconic copy task (Graves et al., 2014) where models must reproduce given sequences. Despite this programmatic simplicity, we find that even advanced models struggle with this straightforward instruction-following challenge. We discover that success in this elementary task strongly predicts (Pearson $r = 0.83$) a model's ability to make meaningful progress in the more complex environments of TALES.

We show the performance of 42 models, open- and closed-weights, in a zero-shot setting on **TALES and perform extensive Reinforcement Learning (RL) fine-tuning experiments to quantify the impact of privileged or domain knowledge commonly used in prompts to scaffold agent performance**. We analyze game transcripts to identify common behaviors and failure modes from top models. While thinking LLMs tend to outperform non-thinking models overall, both model types struggle to reason across extremely long-horizon contexts where important information is sparsely scattered throughout. This limitation significantly hinders their ability to progress through the **JERICHO** framework (Hausknecht et al., 2020), a collection of games meant to be played by humans, slowly and iteratively over extended periods of time. Despite strong evidence of data contamination (See Appendix R) and success on easier frameworks, no agent is capable of completing the gauntlet of games in TALES in a zero-shot setting with minimal inductive bias.

In summary, our contributions are as follows:

- We introduce TALES, a unified framework **suite** for evaluating agents in text-adventure games.
- We outline the reasoning skills required for an agent to be able to successfully complete any text-adventure game **task** in the benchmark.
- We introduce the new **SIMON SAYS** game mode where the agent must echo a command sequence.
- We investigate the game transcripts of AW, SW and the iconic ZORK1 to find where even the top models fail to progress in games meant to be enjoyed and solved by human ingenuity.
- We provide zero-shot results averaged over 5 runs for 42 of the top LLMs as of the time of writing.

108 • We examine the effects of common prompt scaffolds through the RL fine-tuning of Qwen3-8B.
 109

110 **2 REASONING**
 111

112 The Problem-Space Hypothesis posits that humans solve problems through the creation and manipulation
 113 of problem spaces (Newell, 1979). A problem space is often represented by initial, intermediate,
 114 and goal states, possible operators, and environmental constraints. Complex tasks typically require
 115 a combination of reasoning skills to navigate these spaces effectively. Consider an LLM agent for
 116 computer use: it must identify its current state (screen content) and goal state (task completion),
 117 while discovering operators, both familiar and unclear interface affordances. The agent must apply
 118 existing knowledge to infer environmental constraints (e.g., what actions are permissible). As tasks
 119 grow complex, the problem space expands with both known and unknown operators and constraints.
 120 Success requires the agent to leverage a composite of different reasoning skills. Lacking a specific
 121 reasoning skill would lead to task failure as the error will cascade and be difficult to recover.
 122

123 From these insights on the Problem-Space Hypothesis, we identified four reasoning skills critical for
 124 LLM agents to succeed in complex, grounded tasks. Those skills comprise the capability module
 125 (Liu et al., 2024) measured by TALES.

126 **Deductive reasoning:** The ability to derive valid actions through the logical application of general
 127 principles within a specific environmental context (Johnson-Laird, 1999). Deductive reasoning is
 128 particularly critical when environmental interactions are limited or when action has substantial costs
 129 and irreversible consequences. In such scenarios, the agent must leverage pre-existing knowledge to
 130 understand the affordance and constraints of the context and make correct actions towards the goal.
 131

132 **Inductive reasoning:** The ability to draw conclusions through interactions and observations. This is
 133 a critical skill for agents that interface with complex, interactive systems. Given the diversity of tasks,
 134 the environment's affordance may be unknown or contradict with general principles (e.g., a software
 135 interface element behaving inconsistently across operating systems). The inductive reasoning skill
 136 allows the agent to discover operators (actions) and rules about the environment through exploration.
 137 This skill encompasses both adjustments to API calls to adhere to strict function signatures as well as
 138 learning from system feedback (Zhong et al., 2024).

139 **Spatial reasoning:** The ability to navigate effectively and understand the spatial relationship among
 140 game objects, including path finding, backtracking, and locating items (Momennejad et al., 2023). An
 141 agent with spatial reasoning skills can integrate information about locations to identify intermediary
 142 states between the initial and goal states.

143 **Grounded reasoning:** The ability to make decisions based on relevant information and current
 144 context. An agent with grounded reasoning skills can accurately perceive and understand the current
 145 state of the problem space. This reasoning skill is analogous to situational awareness in humans.
 146 Although an LLM may be pre-trained on a vast amount of world knowledge, it has to attend to
 147 task-specific information when making its decisions. As agents **may** have access to the full interaction
 148 history at every step, the ability to correctly identify what information is relevant to the current state
 149 and reason over said information becomes **critically** important as the length of the history grows.
 150

151 The ability to leverage all of these skills is critical to the success of agents as the complexity of the
 152 task increases. Within longer contexts, these reasoning skills often become compositional with a
 153 failure in one skill leading to failures in the others later on.

154 We believe text-adventure games are an ideal environment to simulate the challenges of real-world
 155 tasks, i.e., evaluating an agent on all four core reasoning skills at the same time. A deductive failure
 156 may lead an agent to waste a large number of steps trying to accomplish a sub-goal that does not
 157 progress towards the objective. Inductive reasoning failures may cause an agent to repeatedly attempt
 158 the same action over and over without success. Figure 1 illustrates a simple task in a text-adventure
 159 game where multiple reasoning skills **may be** required at each step and a single failure leads to later
 160 errors. We discuss explicit failure modes that result from the lack of these core reasoning skills in
 161 Section 5 and design choices meant to allow for better evaluation in Section 6

162 Broadly speaking, interactive text-game environments are ideal test-beds for exploring highly complex,
 163 interactive and grounded scenarios where previous work has shown the potential for cross-domain
 164 transfer of knowledge and experiences (Ammanabrolu & Riedl, 2021).

	TEXTWORLD	TEXTWORLDEXPRESS	ALFWORLD	SCIENCEWORLD	JERICHO
Properties	(TW)	(TWX)	(AW)	(SW)	(JERICHO)
# Games	10	16	12	30	54
Avg. walkthrough steps	13.70	33.06	5.83	41.67	87.15
Informative feedback	✓	✓	✗	✓	✓
Intermediate rewards	✓	✓	✗	✓	✓
Nearest-neighbor parser	✓	✗	✗	✓	✓
Dead States	✓	✓	✗	✓	✓

Table 1: Key properties of each framework. Informative feedback: when an action fails, specific details about the nature of the failure are returned. Nearest-neighbor parser: environment can understand similar action phrases to mean the same thing. Dead States: actions can prevent completing the task until a reset. **All frameworks aside from JERICHO are synthetic: created specifically to train embodied AI agents.**

3 TALES

All frameworks included in TALES are *text-adventure game* environments where players are provided a textual observation, and sometimes an explicit goal, and are able to interact with the environment through short action phrases. If these action phrases are invalid, the parser will typically return an error message indicative of whether the action has been understood by the environment but cannot be done, or if the parser just does not understand the action. **Table 1 contrasts properties for each framework.** We provide a short description of each environment and any notable characteristics about the environment or rewards in the following section. We organize the following sections in the rough order of difficulty with the recommendation that users avoid testing on later frameworks without an agent that is able to succeed in earlier environments due to the scaling difficulty **and computational costs for evaluation.** We do not use all environments within each framework. See Appendix P for more details.

3.1 SIMON SAYS: YOU SHALL NOT PASS UNLESS YOU CAN SOLVE THIS TASK

For all frameworks included in TALES, there is a requirement for the agent to be at least minimally proficient in all reasoning skills to make any non-trivial progress. With the release of TALES, we also introduce a new TWX game in the form of “SIMON SAYS”. The basic SIMON SAYS simply provides the agent an action to repeat at each turn while SIMON SAYS WITH MEMORY provides a list of actions to follow at the start of the game. Both versions award a point for every correct action. The game restarts if any action is performed out of order or is wrong. SIMON SAYS is unique compared to other games in TALES as it requires minimal reasoning to complete. However, we find it serves as a good heuristic to evaluate whether an agent is likely to succeed in TALES with a Pearson correlation coefficient of .83 when taken with respect to the average of all SIMON SAYS games against the entirety of TALES. A prerequisite to success in TALES is the ability to at least properly attend to information over a long horizons. SIMON SAYS is the simplest form of this, posed in a straightforward, instruction-following task. We include a graph visualizing the correlation between success in SIMON SAYS and success in TALES in Appendix K.

3.2 FRAMEWORKS

TEXTWORLD (TW) (Côté et al., 2018) is a framework originally designed for training agents with Reinforcement Learning (RL) on text-based games. It can generate synthetic text-adventure games of varying complexity. In TALES, we integrate the “CookingWorld” games that were used as part of the NeurIPS 2018 Competition¹. The task involves following a recipe that requires finding ingredients and processing them according to said recipe. We selected one game per difficulty ranging from level 1 (with one location and a recipe of 1 ingredient) to level 10 (having 12 locations and a recipe with 3 ingredients). The player receives 1 point after completing sub-goals related to the task in the game.

¹<https://competitions.codalab.org/competitions/21557>

216 TEXTWORLDEXPRESS (TWX) (Jansen & Côté, 2022) is a highly optimized re-implementation of
 217 many TW game scenarios that runs approximately three orders of magnitudes faster compared to the
 218 TW counterparts. We opt to use TWX over TW for the performance improvement where applicable.
 219 While significantly faster, an arguable drawback of using TWX over TW is also in its stricter parser.
 220 TWX simplifies its parser for speed and thus does not allow for nearest-neighbor action phrases.
 221 **Compared to other environments, the feedback from `help` only re-iterates the game objective.**

222 ALFWORLD (AW) (Shridhar et al., 2021) is a multi-modal framework, combining complementary
 223 visual and textual observations, where agents are asked to navigate and perform tasks in a household
 224 setting. All tasks provide only a terminal reward of 1 upon task completion. For TALES, we only
 225 use its textual modality as it has become the standard in the LLM literature when evaluated on AW
 226 (Yao et al., 2023; Shinn et al., 2023). The AW environments are unique in their lack of informative
 227 feedback. Where other environments have a predefined error message relating to the type of error,
 228 whether it is due to the parser not recognizing the command or the action not being possible, AW has
 229 only one error message: **Nothing happens**. In the original AW framework, the visual component
 230 compensates for the lack of detailed text feedback. However, this makes it significantly harder for
 231 agents relying solely on text-based interactions, compounded by the limitation that in ALFWORLD
 232 you can hold one object at a time.

233 SCIENCEWORLD (SW) (Wang et al., 2022) is a framework focused on the completion of elementary-
 234 level science curriculum tasks. Notably for many of its tasks, SW emulates an open-world setting
 235 where the player can complete the task in different ways that do not follow one expected trajectory.
 236 When it comes to heating objects, this part of the task can be completed by either the oven in the
 237 kitchen or the blast furnace in the workshop. Similarly, SW also allows the player the freedom to
 238 reset the game on command. This is especially important as a number of SW games have failure
 239 modes where it is no longer possible to complete the assigned task in that playthrough.

240 JERICHO (Hausknecht et al., 2020) is a suite of 54² human-written, interactive fiction games. We
 241 consider JERICHO to be the most difficult framework due to the length and complexity of many of
 242 the games. Some can be completed within 17 steps while some others require over 500 steps. These
 243 games also cover an extremely wide range of genres and styles and lack the consistency of many
 244 other text-game environment suites designed for evaluating agents. For example, 9:05 follows the
 245 morning of an ordinary office worker **while** ANCHORHEAD is a Lovecraftian Horror Story.

246 4 EVALUATION

249 TALES enables evaluation by customizing models with specific prompts and agentic strategies. For
 250 our initial release, we adapt examinee models by considering a minimal agent scaffolding that uses
 251 the following prompt in a zero-shot settings, i.e., without any examples of playing text-based games.

253 You are playing a text-based game and your goal is to finish it with the highest score. Upon reading
 254 the text observation, provide a *single* short phrase to interact with the game, e.g. `get lamp`
 255 (without backticks). When stuck, try using the `help` command to see what commands are available.

256 For the main results, we do not provide any other instructions to the LLMs on how to play the game.
 257 We aim to measure LLMs' raw capabilities exempt of inductive bias from **a** human expert with
 258 domain knowledge. In Section 6, we explore the effects of providing **different forms of privileged**
 259 **or domain knowledge directly** to the agent. When calling the LLMs, the observation and feedback
 260 are provided as the *user* inputs while the LLM actions are recorded as the *assistant* outputs.

261 For our results in the initial release of TALES, we cap the number of steps the agents can take in any
 262 environment to 100. In most frameworks, 100 steps serves as ample opportunity to make mistakes,
 263 self-correct, and eventually find success through directed exploration without allowing for a randomly
 264 acting agent to eventually happen upon success through chance. Even though only 57% of the total
 265 score is achievable within 100 steps in JERICHO, no agent approaches this score. This step limit thus
 266 serves an effective medium between the easier and the most challenging frameworks in TALES. As
 267 the step number is not explicitly referenced in the system prompt, this allows us to continue exploring
 268 longer horizon performances leveraging current game history in future work.

269 ²We exclude HOLLYWOOD.Z3 because of segfault errors and THREATRE.Z5 due to game engine errors.

270

271 Table 2: Average scores per framework and total TALES score for the top 10 models. **TALES score**
 272 **is calculated by averaging success over all tasks.** Almost uniformly, reasoning LLMs outperform
 273 non-reasoning LLMs in all frameworks. However, a higher thinking budget does not always lead to
 274 better overall results.

Model	TEXTWORLD	TEXTWORLD EXPRESS	ALFWORLD	SCIENCEWORLD	JERICHO	TALES 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

283

284

285 TALES captures the model’s capability evidence by the score from each game environment, ranging
 286 from 0-100. Although each game environment has its own customized scoring rules, those rules
 287 mark significant milestones in solving the game. In Appendix B, we include a breakdown of the
 288 percentage of the max score from following the game walkthrough to a certain number of steps in
 289 JERICHO. Each game is played 5 times to account for the stochastic nature of LLMs, but **we** find
 290 minimal changes in performance.³ When supported by the LLM, we set the temperature to 0, provide
 291 fixed random seeds, and limit the amount of reasoning tokens to 1024.

292

293 Table 2 shows the per-framework scores of the top 10 overall scoring models. While both thinking
 294 and non-thinking LLMs excel in synthetic environments, significant progress is still needed in AW,
 295 SW, and JERICHO, especially for non-thinking LLMs. However, more thinking is not always better.
 296 For both o3 and Claude-3.7-Sonnet, we see an increased thinking budget resulting in a lower overall
 297 score, with the non-thinking mode of Claude-3.7-Sonnet achieving the highest score in TWX.

298

299

5 ANALYSIS

300

301

302

303

304

305

306

307 In this section, we provide the results from a qualitative analysis of the game transcripts of the top
 308 LLMs for AW, SW, and ZORK1. This analysis is meant to supplement the automatic evaluation of
 309 the TALES score with empirical evidence of the importance of the core reasoning skills outlined
 310 in Section 2. We select AW and SW due to the lower overall performance in these environments
 311 compared to TW or TWX. We use ZORK1 as a representative sample of JERICHO due to its renown
 312 as one of the most famous and influential text-adventure games, and ability to challenge current
 313 state-of-the-art despite a high likelihood of data contamination.

314

315

316 Following our qualitative analysis of top LLM behaviors, we provide a short analysis of the strengths
 317 and weaknesses of the Claude-3.7-Sonnet thinking mode in contrast to the non-thinking mode. We
 318 focus on 3.7-Sonnet due to its performance in both modes as well as accessibility to the thinking traces.
 319 **In Appendix M we** take a representative subset of the game logs from all Claude models, explicitly
 320 labeling reasoning failures in a per-step granularity over 32 hand-annotated logs, **and investigate**
 321 **whether the 100 step limit for TALES poses an artificial limit through a trial of ZORK1 where**
 322 **the best performing models are allowed to explore for up to 1000 steps in Appendix N.**

323

5.1 REASONING FAILURE MODES IN ALFWORLD, SCIENCEWORLD, AND ZORK1

324

325

326

327 **LLM Agents Waste Steps With Weak Deductive Reasoning.** Agents with weaker deductive
 328 reasoning skills often waste many steps in undirected exploration. This manifests in interactions with
 329 distractor game elements and failures to have their commands understood by the environment action
 330 parser. In our system prompt, we explicitly tell the agent to use the ‘help’ command when stuck,
 331 which provides the agent with the set of action templates for the environment. We found that stronger
 332 models often immediately used the help action and leveraged the provided action templates to avoid
 333 wasting environment steps with invalid action phrases.

³We run 5 times for stability at 100 steps, though 3 runs may suffice if budget is limited.

324 **Most LLM Agents Can Inductively Reason On the Step-Level but Not the Trajectory-Level.**
 325 While the best agents could iteratively reason over step-level attempts, they failed to improve on past
 326 trajectories when resetting. On the step-level, we found agents displaying strong inductive reasoning
 327 skills were able to have their intent understood by using minor variations at each turn rather than
 328 repeating the exact same phrase. We found this to be a baseline competency required for agents to
 329 process through the environments, regardless of whether the help command was called. The top
 330 models displayed the ability to do this iteration methodology on the *step-level* but often failed to do
 331 so on the *trajectory-level*, repeating past mistakes instead of optimizing over previous attempts.

332 **Spatial and Grounded Reasoning Failures Often Result In and From Hallucination.** The strict
 333 causal-constraints of a situated environment allow for a unique lens to view what causes an LLM
 334 agent to hallucinate. Failures in these reasoning skills manifested through the LLM being unable to
 335 differentiate between what elements had appeared within the context and what elements were within
 336 the agent’s current scope, or a failure to adjust to an error message from the environment, similar to
 337 Figure 1. While the top LLMs would not fabricate entirely new game elements, they would often
 338 attempt to directly interact with those mentioned previously in context but no longer in the agent’s
 339 scope. This includes elements mentioned in the task description but not in observations.

340 **5.2 TO THINK OR NOT TO THINK? A CLAUDE-CASE STUDY**

342 Due to overall performance and availability of the thinking traces, we analyze the game transcripts
 343 and thinking traces of Claude-3.7-Sonnet, comparing the model’s performance with and without the
 344 ability to think. (White et al.) records a significant increase in performance for Claude-3.7-Sonnet
 345 with thinking enabled versus the base model. Despite this, we see an increase of less than 1% in
 346 TALES score for the thinking versus non-thinking modes of Claude-3.7-Sonnet.

347 **Thinking scaffolds different reasoning skills at different stages of progression.** We find Claude-
 348 3.7-Sonnet often exhibits one or more reasoning skills in its thinking traces throughout the progression
 349 of the game. Early, the primary reasoning skill explicitly exhibited is often deductive reasoning when
 350 identifying relevant game elements to interact with or areas to explore. Spatial reasoning appeared
 351 intermittently when a sub-goal was **completed and the agent had to navigate back to another**
 352 **location.** Later steps mainly focused on inductive reasoning to synthesize implicit knowledge gained
 353 from exploring the environment or grounded reasoning in evaluating the current state of the agent in
 354 the environment. In particular, this appeared to aid the reasoning agent in avoiding falling into the
 355 cyclical behavior patterns that appeared in even top models in more complex environments like SW
 356 or ZORK1.

357 **Thinking LLMs still fail when required to integrate multiple reasoning skills simultaneously.**
 358 Although Claude-3.7-Sonnet displayed all core reasoning skills individually at various points through-
 359 out the games, failures still **occurred** when multiple reasoning skills were required at key steps.
 360 Similar to the example shown in Figure 1, we see correct reasoning about one aspect of the current
 361 state only to neglect some other critical detail. In some cases when the missed detail leads to an
 362 immediate error, the LLM is capable of self-correcting. However, errors occurring later in the horizon
 363 often results in the agent being unable to identify and backtrack to the original point of failure.

364 **6 PRIVILEGED OR DOMAIN KNOWLEDGE AS ENVIRONMENT
 365 MODIFICATIONS**

366 (Liu et al., 2024) define an adaption module as "how test conditions are constructed", noting a need
 367 for fairness in evaluation condition across all objects of evaluation. **Providing excessive information**
 368 **during training and evaluation can lead to reported improvements of a method being unreliable**
 369 **when compared to baselines due to the scaffolding this information provides.** In TALES, we
 370 motivate a minimal adaption module to evaluate the **baseline** reasoning capabilities of LLM agents in
 371 complex, long-horizon, situated environments with text-adventure games. A number of other works
 372 explore individual frameworks within TALES, but all of these works, to our knowledge, introduce
 373 significant amounts of **privileged information** or domain knowledge through their implementation
 374 of the adaption module(Chang et al., 2024; Paglieri et al., 2024; Lu et al., 2025; Feng et al., 2025).

375 Table 3 illustrates common modifications to the adaption module. Admissible Actions are the actions
 376 the agent can take at any step, directly queried from the environment; Action Templates are those

378
 379 Table 3: Zero-shot scores in TW for Qwen3-8B when different forms of **privileged or** domain
 380 knowledge are included in the prompt. Experiment settings are **kept the** same as main results outside
 381 of changes to **the** system prompt. We use Textworld as a zero-shot proxy, and perform more extensive
 382 RL fine-tuning experimentation below.

In Prompt	Admissible Actions	Action Templates	Environment Dynamics	Textworld Score
BALROG	✗	✓	✓	28%
AgentBoard	✗	✓	✗	32%
IGE	✗	✓	✓	28%
verl-agent	✓	✗	✗	39%
TALES	✗	✗	✗	18%

389
 390
 391 otherwise provided by the **help** command; and Environment Dynamics refer to information such
 392 as “**The BBQ is for..., the key is for...**” that the agent would otherwise need to discover on its own.
 393 Each are problematic when taken as implicit knowledge the model should already have at the start
 394 rather than effective modifications to the environment via the adaption module.

396 6.1 PRIVILEGED AND DOMAIN INFORMATION

397
 398 We define “privileged information” as knowledge the agent has no access to within the scope
 399 of the environment, and “domain information” as knowledge the agent would otherwise need
 400 to discover through interaction and feedback. For example, using a ground-truth solution to
 401 perform reward assignment would be considered privileged information. If this solution were
 402 to be intentionally exposed in the environment for the agent to discover, it would be considered
 403 domain information. The per-step *Admissible Actions* are privileged information, given they are
 404 explicitly generated and verified by the framework. Within the scope of the environment itself,
 405 the model has no direct access to this set of commands.

406 While the action templates are accessible via the **help** command and environmental dynamics
 407 may be eventually discovered by the agent on its own, we argue this information should not
 408 be directly provided to the model at the start. The need to discover domain information
 409 through exploration represents one of the most challenging aspects of text-adventure games.
 410 Synthesizing this information requires strong inductive reasoning in parsing over game-play
 411 history and identifying the relationships between the agent, its actions, and feedback from the
 412 environment. While the argument could be made that this knowledge would be readily available
 413 in real-world applications, there are also many applications where the breadth of scope may
 414 make such pre-defined knowledge less useful. For example, trial and error debugging may
 415 require a coding agent to identify potential quirks about a code base purely through exploratory
 416 analysis. Similarly, we argue directly including domain knowledge in the prompt should be
 417 avoided, and explicitly acknowledged if done.

418 6.2 IMPACT ON RL FINE-TUNING

419 We evaluate the impact of directly providing privileged and domain information to the model
 420 during RL fine-tuning by ablating over the inclusion of the admissible actions and the contents
 421 of **help** in the prompt across the synthetic environments TW, TWX, AW, and SW. Given the
 422 model fails to find any success in SW, we exclude JERICHO which is even more challenging.

423 We use the verl-agent (Feng et al., 2025) framework due to their pre-existing results on AW.
 424 Integration of TALES is done standalone, rather than integrated into the existing verl-agent.
 425 Hyperparameters are primarily derived from their existing AW configuration and details from
 426 their appendix. We do not extensively sweep for ideal hyperparameters per environment as these
 427 results are meant to be a rudimentary baseline. See Appendix O for the exact hyperparameters
 428 used. Qwen3-8B(Team, 2025) is used due to its strong performance as a thinking-model. All
 429 hyperparameters are kept constant across experiments. Figure 2 shows the results, averaged
 430 over three seeds.

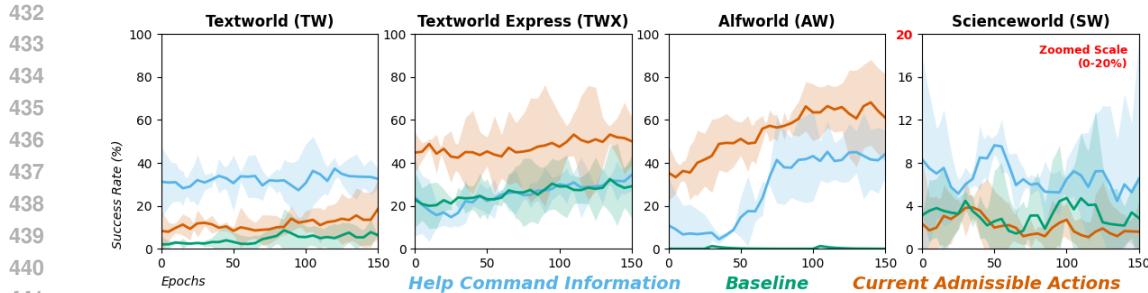


Figure 2: **Inclusion of admissible actions and the feedback from `help` in an agent’s prompt during RL finetuning can result in a significant increase in performance over an unscattered baseline.** TWX shows no performance change due to `help` providing only the task description.

Simpler environments can be made significantly easier by the inclusion of privileged information. The requirement to explore and iterate over what action commands are accepted by the game parser can be one of the most challenging aspects of an environment. The ability to do this iteration and keep track of the current state of the problem space tests a model’s inductive, spatial, and grounded reasoning by forcing the choice of action to occur in the unbounded space of natural language. Bounding the space from which the agent must select its action to a subset of precomputed commands can drastically reduce the difficulty of the environment, as we can see by the extremely large differential in performance for TWX and AW when trained with and without admissible actions included in the prompt.

Training LLM Agents with privileged environment information should be avoided due to a failure to generalize. While capable of improving performance in simpler environments, leveraging privileged information as scaffolding can result in overfitting to the specified environment due to over-reliance. Taking admissible actions baseline as an example, we see significant improvements in TWX and AW. However, we fail to see a similar improvement in TW with performance even dropping in SW. These environments significantly feature actions that can leave the agent entirely unable to complete the given task. Without context, these actions can appear innocuous when embedded in the admissible actions. SW in particular has sufficient complexity such that the admissible actions can quickly grow combinatorially large, saturating the context with useless information. While not visible from a graph plotting success rate against epochs, computational cost on the environment can also become an issue⁴.

Domain information can generalize as a strong deductive prior. With the exception of TWX, where `help` only repeats the task objective, we see improvements in all frameworks in the `help` baseline where the information included in the prompt is information the agent would be able to otherwise discover on its own. While this still decreases the difficulty of the environment, this otherwise accessible domain information acts as more natural guidance compared to the privileged information from the admissible actions.

LLM Agents face difficulty improving performance when exposed to a large number of diverse environments simultaneously. In the `help` and admissible actions baselines, we see an upward trend of success in AW, where all tasks involve relatively uniform task specifications. However, in TW, TWX, and SW, we see little to no learning occurring due to the extremely noisy reward signal from a wide range of task variety and difficulty.

7 RELATED WORK

A large body of work exists in teaching agents to navigate and successfully complete text world games. We specifically divide this section into RL-Based agents, where the text-world is defined as a Partially Observable Markov Decision Process (POMDP) (Kaelbling et al., 1998) and LLM-based agents where information is fed to the LLM as an input with the output taken as an action.

⁴On a cluster of 8 H100s, the average wall-clock time for the training runs of SW with admissible actions was roughly 54 hours, with the baseline and `help`-baseline needing only an average of 22 hours.

486 **RL-Based:** Prior work explores text world games as benchmark for non-LLM-based agents
 487 (Narasimhan et al., 2015; Hausknecht et al., 2020). Due to the intractable action space of lan-
 488 guage, prior RL approaches used action templates to reduce the space **of** possible commands down
 489 to a subset learnable by an RL agent (Narasimhan et al., 2015; Ammanabrolu & Riedl, 2018; Yuan
 490 et al., 2018; Hausknecht et al., 2019; Ammanabrolu & Hausknecht, 2020; Ammanabrolu et al., 2020;
 491 Murugesan et al., 2021; Ryu et al., 2023). These agents are often augmented with a knowledge
 492 graph for better state tracking representation or for directing the agent (Ammanabrolu & Riedl, 2018;
 493 Hausknecht et al., 2019; Ammanabrolu & Hausknecht, 2020; Murugesan et al., 2021; Peng et al.,
 494 2023; Cui et al., 2023; 2024). Other approaches still use a base RL agent but use an LLM to guide
 495 the RL agent or generate diverse environments for generalization (Yao et al., 2020; Basavatia et al.,
 496 2024; Golchha et al., 2024).
 497

498 **LLM-Based:** Early results demonstrated that even state-of-the-art pre-trained LLMs face difficulty
 499 when playing text-adventure games meant for human players (Tsai et al., 2023). Previous work has
 500 explored leveraging an external buffer or knowledge base to guide the agent (Shinn et al., 2023; Zhu
 501 et al., 2024). Other approaches leverage task decomposition into simpler sub-goals (Lin et al., 2023;
 502 Prasad et al., 2024). (Wang et al., 2024; Zhao et al., 2024) introduce approaches that leverage LLMs
 503 to modify the provided action space. Leveraging contrasting trajectories with LLMs to improve
 504 performance has also been explored (Song et al., 2024; Yang et al., 2024; Qiao et al., 2025).
 505

506 8 CONCLUSION

507 In this work, we introduce TALES, a unified benchmark for LLM agents in text-adventure game
 508 environments. We identify a set of reasoning skills essential to agents operating through APIs to
 509 interface with outside environments.

510 We use **SIMON SAYS** to evaluate an agent’s capability of the most basic composite reasoning needed
 511 to succeed in TALES. The game transcripts from leading LLMs reveal that, despite their impres-
 512 sive language capabilities, these models still struggle with core reasoning challenges inherent to
 513 text-adventure games. The difficulty stems not only from long-horizon dependencies and implicit
 514 environmental cues but also from the need for sequential, exploratory, and commonsense reason-
 515 ing—skills that remain a bottleneck for even state-of-the-art LLMs.

516 We introduce baseline LLMs in our framework as canonical benchmarks for text-adventure environ-
 517 ments without additional prompting or training **and provide a set of baseline RL ablations that**
 518 **demonstrate the effects of including privileged or domain information in the prompt.** Future
 519 work can improve upon these through supervised fine-tuning, in-context learning, or reinforcement
 520 learning. Thinking LLMs effectively leverage thinking traces; future research could guide these
 521 traces toward specific reasoning types for improved contextual awareness.

522 TALES supports game history truncation; studying how limited context scope affects LLM behavior
 523 and developing methods to condense game history by eliminating redundant information could
 524 enhance performance. While we manually analyze game logs to identify behaviors and failure modes,
 525 future work could automate this process using judge LLMs.

526 **Limitations.** Qualitative analysis was primarily performed by the authors due to familiarity with the
 527 included frameworks. However this may have resulted in implicit bias in the analysis due to this prior.
 528 The development of an automated method of annotation and cross-validation with an analysis from
 529 participants less familiar with the included frameworks may help further validate observations. We
 530 also cap the number of environmental steps for our work to 100. This cap allows for examination
 531 of an LLM’s reasoning skills through their ability to perform directed exploration, but future work
 532 could explore extending this horizon and investigating what behaviors extremely long context elicit.

533 We view JERICHO as the ultimate test for an LLM agent’s reasoning capabilities, however this is
 534 complicated by data contamination issues. As shown in Appendix R, we are almost certain modern
 535 LLMs were trained on ZORK1 playthrough transcripts, and this likely extends to many, if not all,
 536 games in the JERICHO suite. This raises the question of how much of the LLMs performance is
 537 influenced by the trajectories in its pre-training.

538 Overall, while progress has been made on synthetic text-adventure games, LLM-driven agents are
 539 still far from being able to complete games meant to be played for simple, human enjoyment.

540 9 REPRODUCIBILITY STATEMENT.
541542 For the sake of anonymity, we provide an anonymized repository of all code used to gather the results
543 presented in this work. This code will be released and made available to the public. We note that
544 some of the close-sourced models may no longer be available at the time of reviewing or publication.
545 Trajectories for these models can be provided upon request.547 REFERENCES
548

549 Prithviraj Ammanabrolu and Matthew J. Hausknecht. Graph constrained reinforcement learning
550 for natural language action spaces. *ArXiv*, abs/2001.08837, 2020. URL <https://api.semanticscholar.org/CorpusID:210911499>.

552 Prithviraj Ammanabrolu and Mark O. Riedl. Playing text-adventure games with graph-based deep reinforcement learning. *ArXiv*, abs/1812.01628, 2018. URL <https://api.semanticscholar.org/CorpusID:54458698>.

556 Prithviraj Ammanabrolu and Mark O. Riedl. Situated language learning via interactive narratives. *Patterns*, 2(9):100316, 2021. ISSN 2666-3899. doi: <https://doi.org/10.1016/j.patter.2021.100316>. URL <https://www.sciencedirect.com/science/article/pii/S2666389921001598>.

560 Prithviraj Ammanabrolu, Ethan Tien, Matthew J. Hausknecht, and Mark O. Riedl. How to avoid
561 being eaten by a grue: Structured exploration strategies for textual worlds. *ArXiv*, abs/2006.07409,
562 2020. URL <https://api.semanticscholar.org/CorpusID:219687812>.

564 Shreyas Basavatia, Keerthiram Murugesan, and Shivam Ratnakar. Starling: Self-supervised training
565 of text-based reinforcement learning agent with large language models. *ArXiv*, abs/2406.05872,
566 2024. URL <https://api.semanticscholar.org/CorpusID:270372057>.

567 Ruth MJ Byrne and Philip N Johnson-Laird. Spatial reasoning. *Journal of memory and language*, 28
568 (5):564–575, 1989.

570 Ma Chang, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan,
571 Lingpeng Kong, and Junxian He. Agentboard: An analytical evaluation board of multi-turn llm
572 agents. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang
573 (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 74325–74362. Curran
574 Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/877b40688e330a0e2a3fc24084208dfa-Paper-Datasets_and_Benchmarks_Track.pdf.

577 Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James
578 Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and
579 Adam Trischler. Textworld: A learning environment for text-based games. *CoRR*, abs/1806.11532,
580 2018.

581 Christopher Cui, Xiangyu Peng, and Mark Riedl. Thespian: Multi-character text role-playing game
582 agents. 2023.

584 Christopher Z Cui, Xiangyu Peng, and Mark O Riedl. A mixture-of-experts approach to few-shot
585 task transfer in open-ended text worlds. *arXiv preprint arXiv:2405.06059*, 2024.

587 Mica R Endsley, Daniel J Garland, et al. Theoretical underpinnings of situation awareness: A critical
588 review. *Situation awareness analysis and measurement*, 1(1):3–21, 2000.

590 Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm
591 agent training. *arXiv preprint arXiv:2505.10978*, 2025.

592 Hitesh Golchha, Sahil Yerawar, Dhruvesh Patel, Soham Dan, and Keerthiram Murugesan. Language
593 guided exploration for rl agents in text environments, 2024. URL <https://arxiv.org/abs/2403.03141>.

594 Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines, 2014. URL <https://arxiv.org/abs/1410.5401>.
 595
 596

597 Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. Interactive
 598 fiction games: A colossal adventure. In *Proceedings of the AAAI Conference on Artificial
 599 Intelligence*, volume 34, 2020.

600 Matthew J. Hausknecht, Ricky Loynd, Greg Yang, Adith Swaminathan, and J. Williams. Nail:
 601 A general interactive fiction agent. *ArXiv*, abs/1902.04259, 2019. URL <https://api.semanticscholar.org/CorpusID:60441391>.
 602
 603

604 Evan Heit. Properties of inductive reasoning. *Psychonomic bulletin & review*, 7:569–592, 2000.
 605
 606 Peter A. Jansen and Marc-Alexandre Côté. Textworldexpress: Simulating text games at one million
 607 steps per second. *arXiv*, 2022. URL <https://arxiv.org/abs/2208.01174>.
 608
 609

610 Philip N Johnson-Laird. Deductive reasoning. *Annual review of psychology*, 50(1):109–135, 1999.
 611 Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially
 612 observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.
 613
 614 Bill Yuchen Lin, Yicheng Fu, Karina Yang, Prithviraj Ammanabrolu, Faeze Brahman, Shiyu Huang,
 615 Chandra Bhagavatula, Yejin Choi, and Xiang Ren. Swiftsage: A generative agent with fast
 616 and slow thinking for complex interactive tasks. *ArXiv*, abs/2305.17390, 2023. URL <https://api.semanticscholar.org/CorpusID:258960143>.
 617
 618 Yu Liu, Su Blodgett, Jackie Chi Kit Cheung, Vera Liao, Alexandra Olteanu, and Ziang Xiao. Ecbl:
 619 Evidence-centered benchmark design for nlp. In *Proceedings of the 62nd Annual Meeting of the
 620 Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16349–16365, 2024.
 621
 622 Cong Lu, Shengran Hu, and Jeff Clune. Intelligent go-explore: Standing on the shoulders of giant
 623 foundation models. In *The Thirteenth International Conference on Learning Representations*,
 624 2025. URL <https://openreview.net/forum?id=apErWGzCAA>.
 625
 626 Ida Momennejad, Hosein Hasanbeig, Felipe Vieira, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa
 627 Jovic, Hamid Palangi, and Jonathan Larson. Evaluating cognitive maps and planning in large
 628 language models with cogeval, 2023. URL <https://arxiv.org/abs/2309.15129>.
 629
 630 Keerthiram Murugesan, Mattia Atzeni, Pavan Kapanipathi, Kartik Talamadupula, Mrinmaya Sachan,
 631 and Murray Campbell. Efficient text-based reinforcement learning by jointly leveraging state and
 632 commonsense graph representations. In *Annual Meeting of the Association for Computational Lin-
 633 guistics*, 2021. URL <https://api.semanticscholar.org/CorpusID:236459800>.
 634
 635 Karthik Narasimhan, Tejas Kulkarni, and Regina Barzilay. Language understanding for text-based
 636 games using deep reinforcement learning. In *Proceedings of the 2015 Conference on Empirical
 637 Methods in Natural Language Processing*, 2015.
 638
 639 Allen Newell. Reasoning, problem solving and decision processes: The problem space as a funda-
 640 mental category. 1979.
 641
 642 Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan,
 643 Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack
 644 Parker-Holder, and Tim Rocktäschel. Balrog: Benchmarking agentic llm and vlm reasoning on
 645 games, 2024. URL <https://arxiv.org/abs/2411.13543>.
 646
 647 Xiangyu Peng, Christopher Cui, Wei Zhou, Renee Jia, and Mark Riedl. Story shaping: Teaching
 648 agents human-like behavior with stories. In *Proceedings of the AAAI Conference on Artificial
 649 Intelligence and Interactive Digital Entertainment*, volume 19, pp. 326–336, 2023.
 650
 651 Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal,
 652 and Tushar Khot. Adapt: As-needed decomposition and planning with language models, 2024.
 653 URL <https://arxiv.org/abs/2311.05772>.

648 Shuofei Qiao, Runnan Fang, Ningyu Zhang, Yuqi Zhu, Xiang Chen, Shumin Deng, Yong Jiang,
 649 Pengjun Xie, Fei Huang, and Huajun Chen. Agent planning with world knowledge model.
 650 *Advances in Neural Information Processing Systems*, 37:114843–114871, 2025.

651

652 Dongwon Kelvin Ryu, Meng Fang, Shirui Pan, Gholamreza Haffari, and Ehsan Shareghi. A minimal
 653 approach for natural language action space in text-based games. *ArXiv*, abs/2305.04082, 2023.
 654 URL <https://api.semanticscholar.org/CorpusID:258557517>.

655 Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dy-
 656 namic memory and self-reflection. *ArXiv*, abs/2303.11366, 2023. URL <https://api.semanticscholar.org/CorpusID:257636839>.

657

658 Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew
 659 Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In
 660 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL
 661 <https://arxiv.org/abs/2010.03768>.

662

663 Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error:
 664 Exploration-based trajectory optimization of llm agents. In *Proceedings of the 62nd Annual
 665 Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7584–
 666 7600, 2024.

667 Qwen Team. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

668

669 H. Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Raj Manku, Vinty Dong, Edward Li, Shashank
 670 Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of
 671 apps and people for benchmarking interactive coding agents. *ArXiv*, abs/2407.18901, 2024. URL
 672 <https://api.semanticscholar.org/CorpusID:271516633>.

673 Chen Feng Tsai, Xiaochen Zhou, Sierra S. Liu, Jing Li, Mo Yu, and Hongyuan Mei. Can
 674 large language models play text games well? current state-of-the-art and open questions.
 675 *ArXiv*, abs/2304.02868, 2023. URL <https://api.semanticscholar.org/CorpusID:257985065>.

676

677 Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Soft self-consistency improves
 678 language model agents. In *Proceedings of the 62nd Annual Meeting of the Association for
 679 Computational Linguistics (Volume 2: Short Papers)*, 2024.

680

681 Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld:
 682 Is your agent smarter than a 5th grader?, 2022. URL <https://arxiv.org/abs/2203.07540>.

683

684 Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid
 685 Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, et al. Livebench: A challenging,
 686 contamination-free llm benchmark. In *The Thirteenth International Conference on Learning
 687 Representations*.

688

689 Zonghan Yang, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. React meets acrte: When
 690 language agents enjoy training data autonomy. *ArXiv*, abs/2403.14589, 2024. URL <https://api.semanticscholar.org/CorpusID:268856510>.

691

692 Shunyu Yao, Rohan Rao, Matthew J. Hausknecht, and Karthik Narasimhan. Keep calm and explore:
 693 Language models for action generation in text-based games. *ArXiv*, abs/2010.02903, 2020. URL
 694 <https://api.semanticscholar.org/CorpusID:222142129>.

695

696 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao.
 697 React: Synergizing reasoning and acting in language models. In *International Conference on
 698 Learning Representations (ICLR)*, 2023.

699

700 Xingdi Yuan, Marc-Alexandre Côté, Alessandro Sordoni, Romain Laroche, Rémi Tachet des Combes,
 701 Matthew J. Hausknecht, and Adam Trischler. Counting to explore and generalize in text-based
 games. *ArXiv*, abs/1806.11525, 2018. URL <https://api.semanticscholar.org/CorpusID:49547885>.

702 Haiteng Zhao, Chang Ma, Guoyin Wang, Jing Su, Lingpeng Kong, Jingjing Xu, Zhi-Hong Deng, and
703 Hongxia Yang. Empowering large language model agents through action learning. *CoRR*, 2024.
704

705 Victor Zhong, Dipendra Misra, Xingdi Yuan, and Marc-Alexandre Côté. Policy improvement using
706 language feedback models, 2024. URL <https://arxiv.org/abs/2402.07876>.
707

708 Yuqi Zhu, Shuofei Qiao, Yixin Ou, Shumin Deng, Ningyu Zhang, Shiwei Lyu, Yue Shen, Lei Liang,
709 Jinjie Gu, and Huajun Chen. Knowagent: Knowledge-augmented planning for llm-based agents.
710 *CoRR*, 2024.
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

756 **A ALL AGENT AVERAGE SCORES PER FRAMEWORK**

757

758 In Table 4, we include the average scores per framework and average-per-game score of all LLM
 759 agents.

760

Model	TEXTWORLD	TEXTWORLD EXPRESS	ALFWORLD	SCIENCEWORLD	JERICHO	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-3.3-70B-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
Ministrال-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

792 Table 4: Average scores per framework and total TALES score.

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810 B JERICHO WALKTHROUGH SCORES

812 Table 5 shows the percent of achievable score when using the walkthrough for all JERICHO for 50,
 813 100, 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
JerichoEnvDeephome	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
JerichoEnvHgwg	8.300	21.200	40.000	50.000	100.000	100.000	100.000
JerichoEnvHundark	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
JerichoEnvMurda	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
JerichoEnvOmnique	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

858 Table 5: Max score percentage reached by following the provided walkthrough for each JERICHO
 859 game.

864 C ALL AGENT AVERAGE FINAL TOKENS USED PER FRAMEWORK
865866 In Table 6, we include the average final tokens used per game for each framework of all agents.
867

868 Model	869 TEXTWORLD	870 TEXTWORLDEXPRESS	871 ALFWORLD	872 SCIENCEWORLD	873 JERICHO
874 o3 (medium)	41342.0	107026.3	128740.7	92422.7	378168.5
875 o3 (high)	32148.6	51361.7	68275.0	51323.4	251760.8
876 gpt-5 (thinking)	38774.8	377563.2	145110.7	197999.9	649059.0
877 o3 (low)	51609.6	84218.5	119657.4	88557.9	318262.7
878 claude-3.7-sonnet (thinking)	69138.9	63072.8	74516.3	128752.7	311684.2
879 claude-3.7-sonnet	72072.9	46948.4	65131.1	152130.6	298533.1
880 claude-3.5-sonnet-latest	60766.1	68812.7	78765.4	106749.2	291950.3
881 gpt-4.1	53378.9	46454.8	54107.6	86547.5	218123.2
882 gpt-5-mini (thinking)	151984.2	391444.6	508585.4	318636.6	878673.2
883 o1	47765.6	113492.9	127460.2	74300.1	211958.7
884 gpt-4o	106863.0	49536.2	77946.0	107121.9	209712.8
885 claude-3.5-haiku	119839.9	85136.1	267643.6	205751.0	269458.9
886 Llama-3.1-405B-Instruct	66476.2	52624.6	106290.0	137657.2	226078.6
887 gemini-2.0-flash	142937.1	66075.5	138048.3	142883.3	230182.0
888 Qwen3-32B	198390.4	188065.3	190900.1	229708.1	374514.2
889 Llama-3.3-70B-Instruct	166373.8	70165.0	127348.7	128860.9	205362.8
890 Llama-3.1-70B-Instruct	133253.4	51885.9	106925.7	144615.6	210914.8
891 Qwen2.5-72B-Instruct	112658.0	52096.1	97211.2	168057.3	197628.3
892 Mistral-Large-Instruct-2407	107788.5	110228.6	118395.4	163232.0	243256.0
893 gpt-4.1-mini	184516.3	92775.4	130758.2	125310.5	188824.3
894 gpt-4o-mini	159840.9	60210.7	145236.3	172875.7	182620.7
895 Llama-4-Scout-17B-16E-Instruct	289709.8	120173.9	172633.9	222464.1	229947.2
896 gpt-5-nano	770352.7	623055.1	821024.1	730904.7	825754.3
897 Llama-4-Maverick-17B-128E-Instruct-FP8	287547.2	213139.5	354183.9	394875.5	372902.6
898 Mistral-Small-Instruct-2409	163334.9	304510.9	107549.3	150730.7	208261.8
899 Llama-3.1-8B-Instruct	222239.7	358837.4	96582.5	152293.0	165505.8
900 DeepSeek-R1	393654.5	398322.7	496328.4	431997.9	439399.3
901 Qwen2.5-7B-Instruct	143127.1	214926.3	91334.4	163021.2	171107.7
902 Llama-3.2-3B-Instruct	230950.5	79878.3	84620.4	195397.2	152544.7
903 phi-4	189031.6	100363.9	126068.1	153395.2	178713.4
904 gpt-4.1-nano	545577.7	171767.5	277643.8	201505.9	182678.8
905 Mistral-Small-24B-Instruct-2501	399093.8	500484.8	479125.0	418284.9	475649.7
906 DeepSeek-R1-Distill-Llama-70B	453695.7	637384.1	719404.2	482819.3	407401.8
907 Mistral-8B-Instruct-2410	220157.9	337447.5	112710.5	108916.4	118104.9
908 Mistral-Small-3.1-24B-Instruct-2503	448764.0	507986.4	477505.8	397054.6	514733.5
909 Mixtral-8x22B-Instruct-v0.1	158782.2	137583.5	92832.7	134827.6	156515.8
910 Llama-3.2-1B-Instruct	567691.8	279214.8	457857.3	138285.6	201648.5
911 Phi-3-mini-128k-instruct	245215.0	429993.4	257852.2	253989.5	237881.5
912 Phi-3.5-MoE-instruct	274848.9	295190.9	240007.5	252055.6	271680.0
913 Phi-4-mini-instruct	231947.3	199299.1	195407.4	190887.4	212508.9
914 Mixtral-8x7B-Instruct-v0.1	612791.9	555281.3	520434.6	560994.6	564967.6
915 Phi-3.5-mini-instruct	426125.5	476218.4	410459.6	327584.9	457434.4
916 Phi-3-medium-128k-instruct	620235.4	585925.5	581721.6	513787.5	595335.5

903 Table 6: Avg final tokens used per LLM per game for each framework. Ordering is based on the
904 agent's cumulative average score shown in Table 4.
905906
907
908
909
910
911
912
913
914
915
916
917

918 D AGENT SCORE STANDARD DEVIATIONS

919
920 In Table 7, we include the average standard deviation across seeds per framework of all LLM agents.
921

922 Model	923 TEXTWORLD	924 TEXTWORLDEXPRESS	925 ALFWORLD	926 SCIENCEWORLD	927 JERICHO
928 o3 (medium)	929 0.0	930 2.7	931 4.6	932 2.2	933 0.5
934 o3 (high)	935 0.0	936 3.2	937 9.1	938 1.1	939 1.3
940 gpt-5 (thinking)	941 0.0	942 5.5	943 7.0	944 2.5	945 1.1
946 o3 (low)	947 2.0	948 6.9	949 9.5	950 0.8	951 1.1
952 claude-3.7-sonnet (thinking)	953 2.8	954 4.7	955 10.2	956 2.9	957 0.9
958 claude-3.7-sonnet	959 0.0	960 1.4	961 3.7	962 3.7	963 1.1
964 claude-3.5-sonnet-latest	965 0.0	966 2.9	967 5.9	968 3.4	969 1.0
970 gpt-4.1	971 2.6	972 1.9	973 11.8	974 2.3	975 0.8
976 gpt-5-mini (thinking)	977 3.6	978 2.8	979 9.5	980 6.1	981 1.4
982 o1	983 1.2	984 4.4	985 4.6	986 5.0	987 1.7
988 gpt-4o	989 6.1	990 0.4	991 14.9	992 2.8	993 0.6
994 claude-3.5-haiku	995 5.3	996 0.0	997 3.7	998 2.6	999 0.6
999 Llama-3.1-405B-Instruct	1000 5.0	1001 4.9	1002 10.9	1003 4.5	1004 0.5
1005 gemini-2.0-flash	1006 8.6	1007 1.3	1008 4.6	1009 3.4	1010 0.4
1011 Qwen3-32B	1012 6.8	1013 1.9	1014 10.9	1015 3.2	1016 0.4
1017 Llama-3.3-70B-Instruct	1018 2.8	1019 3.4	1020 3.7	1021 2.3	1022 0.1
1023 Llama-3.1-70B-Instruct	1024 3.5	1025 1.9	1026 5.9	1027 4.5	1028 0.2
1029 Qwen2.5-72B-Instruct	1030 2.0	1031 2.5	1032 4.6	1033 3.8	1034 0.7
1036 Mistral-Large-Instruct-2407	1037 8.2	1038 2.6	1039 3.7	1040 8.1	1041 0.9
1043 gpt-4.1-mini	1044 6.1	1045 1.7	1046 7.5	1047 3.6	1048 0.3
1049 gpt-4o-mini	1050 5.4	1051 1.7	1052 0.0	1053 1.5	1054 0.2
1056 Llama-4-Scout-17B-16E-Instruct	1057 0.0	1058 0.0	1059 0.0	1060 0.0	1061 0.0
1063 gpt-5-nano	1064 7.7	1065 5.1	1066 3.7	1067 4.1	1068 0.3
1071 Llama-4-Maverick-17B-128E-Instruct-FP8	1072 1.3	1073 0.0	1074 0.0	1075 0.1	1076 0.3
1078 Mistral-Small-Instruct-2409	1079 5.1	1080 0.0	1081 0.0	1082 2.2	1083 0.0
1085 Llama-3.1-8B-Instruct	1086 4.7	1087 2.9	1088 0.0	1089 0.9	1090 0.1
1093 DeepSeek-R1	1094 3.9	1095 0.0	1096 0.0	1097 2.2	1098 0.1
1099 Qwen2.5-7B-Instruct	1100 0.0	1101 0.0	1102 0.0	1103 0.7	1104 0.1
1106 Llama-3.2-3B-Instruct	1107 2.6	1108 2.9	1109 0.0	1110 1.6	1111 0.3
1113 phi-4	1114 0.4	1115 0.0	1116 0.0	1117 1.3	1118 0.0
1119 gpt-4.1-nano	1120 2.1	1121 4.5	1122 0.0	1123 1.0	1124 2.4
1126 Mistral-Small-24B-Instruct-2501	1127 3.1	1128 1.0	1129 0.0	1130 1.1	1131 0.3
1134 DeepSeek-R1-Distill-Llama-70B	1135 2.8	1136 0.3	1137 0.0	1138 0.4	1139 0.1
1141 Minstral-8B-Instruct-2410	1142 4.2	1143 0.0	1144 0.0	1145 0.0	1146 0.0
1148 Mistral-Small-3.1-24B-Instruct-2503	1149 0.0	1150 0.0	1151 0.0	1152 0.3	1153 0.0
1156 Mixtral-8x22B-Instruct-v0.1	1157 3.0	1158 2.3	1159 0.0	1160 1.7	1161 0.1
1163 Llama-3.2-1B-Instruct	1164 0.0	1165 0.0	1166 0.0	1167 0.0	1168 0.0
1171 Phi-3-mini-128k-instruct	1172 2.0	1173 0.0	1174 0.0	1175 0.3	1176 0.0
1178 Phi-3.5-MoE-instruct	1179 0.0	1180 2.7	1181 0.0	1182 0.0	1183 0.1
1186 Phi-4-mini-instruct	1187 0.0	1188 0.0	1189 0.0	1190 0.0	1191 0.0
1194 Mixtral-8x7B-Instruct-v0.1	1195 0.0	1196 0.0	1197 0.0	1198 0.0	1199 0.0
1201 Phi-3.5-mini-instruct	1202 0.0	1203 1.1	1204 0.0	1205 0.1	1206 0.0
1208 Phi-3-medium-128k-instruct	1209 0.0	1210 0.0	1211 0.0	1212 0.0	1213 0.0

957
958 Table 7: Standard deviation statistics for different LLMs Ordering is based on the agent's cumulative
959 average score shown in Table 4.
960

961
962
963
964
965
966
967
968
969
970
971

972 **E ALL GAMES**
973974 In Table 8 and Table 9 we list all tasks and games in their respective frameworks.
975976 **Table 8: Games Organized by Framework. Part 1.**
977

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

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

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

Table 9: Games Organized by Framework. Part 2.

ALFWORLD

1. LookAtObjInLightSeen	7. PickCoolThenPlaceInRecepSeen
2. LookAtObjInLightUnseen	8. PickCoolThenPlaceInRecepUnseen
3. PickAndPlaceSimpleSeen	9. PickHeatThenPlaceInRecepSeen
4. PickAndPlaceSimpleUnseen	10. PickHeatThenPlaceInRecepUnseen
5. PickCleanThenPlaceInRecepSeen	11. PickTwoObjAndPlaceSeen
6. PickCleanThenPlaceInRecepUnseen	12. PickTwoObjAndPlaceUnseen

TextWorld

1. CookingLevel1	6. CookingLevel6
2. CookingLevel2	7. CookingLevel7
3. CookingLevel3	8. CookingLevel8
4. CookingLevel4	9. CookingLevel9
5. CookingLevel5	10. CookingLevel10

TWX

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

G ALL SCORES PER GAME: TEXTWORLDEXPRESS

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

H ALL SCORES PER GAME: ALFWORLD

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

I ALL SCORES PER GAME: SCIENCEWORLD

Tables 13 and 14 shows the per-task average scores of all models in SCIENCEWORLD across all seeds.

J ALL SCORES PER GAME: JERICHO

Tables 15 and 16 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.

1080
 1081
 1082
 1083
 1084
 1085
 1086
 1087
 1088
 1089
 1090
 1091
 1092
 1093
 1094
 1095
 1096
 1097
 1098
 1099
 1100
 1101
 1102
 1103
 1104
 1105
 1106
 1107
 1108
 1109
 1110
 1111
 1112
 1113
 1114
 1115
 1116
 1117
 1118
 1119
 1120
 1121
 1122
 1123
 1124
 1125
 1126
 1127
 1128
 1129
 1130
 1131
 1132
 1133

Table 10: 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
claudie-3.7-sonnet (thinking)	100.0	72.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
claudie-3.7-sonnet	100.0	72.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
claudie-3.5-sonnet-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	100.0
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-4.0	86.7	14.5	70.0	100.0	86.7	100.0	100.0	100.0	100.0	78.2
claudie-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	100.0	100.0	100.0	100.0	90.9
Owen-3-32B	100.0	27.3	85.0	100.0	100.0	100.0	46.7	86.7	100.0	81.8
Llama-3.3-70B-Instruct	100.0	0.0	25.0	100.0	100.0	100.0	0.0	100.0	100.0	90.9
Llama-3.1-70B-Instruct	100.0	0.0	25.0	100.0	100.0	100.0	0.0	100.0	100.0	84.0
Owen-2.5-72B-Instruct	100.0	0.0	25.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
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-4.0-mini	100.0	0.0	25.0	100.0	100.0	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	100.0	0.0	0.0	40.0	45.5
gpt-5-nano	60.0	12.7	85.0	100.0	33.3	0.0	60.0	84.0	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	28.0	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
Owen-2.5-7B-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-2503	46.7	6.7	0.0	25.0	25.0	33.3	0.0	0.0	20.0	8.0
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	0.0	100.0	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	0.0	10.0	5.0	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 11: Model Performance on TEXTWORLDEXPRESS tasks.

Models	Arithmetic	CanCollector	CookingWorld	MapReader	PeckingOrder	Simons						TextWorldCommunication					
						Simons50	Simons100	Simons150	Simons200	Simons300	Simons400	Simons500	Simons700	Simons900	Simons1200	Simons1500	Simons1800
c3 (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	51.4	53.4	61.4	94.0	97.2
c3 (high)	80.0	100.0	53.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	92.0	92.0	93.2	100.0	100.0
ep-5 (thinking)	40.0	100.0	42.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	42.2	42.2	63.4	73.4	100.0
g3 (low)	100.0	100.0	45.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	64.4	64.4	64.4	63.4	60.0
g3 (medium)	70.0	100.0	42.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.6	90.6	100.0	100.0	100.0
g3 (high)	100.0	100.0	40.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
claus-e-7-source (thinking)	200.0	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	100.0
claus-e-5-source-latest	0.0	100.0	30.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	88.0	100.0	100.0	100.0	100.0
ep-4-1	60.0	100.0	42.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	14.2	24.2	32.2	32.2	56.6
ep-5-mini	0.0	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	100.0
claus-e-3.5-haiku	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	100.0	100.0
Liam-e-3.1-40B-Instruct	80.0	100.0	33.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	21.0	42.0	100.0	100.0	100.0
gemini-0.6-flash	30.0	100.0	33.6	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Qwen-3.2B	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	100.0	100.0	100.0	100.0
Liam-e-3.1-7B-Instruct	200.0	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	100.0	100.0
Qwen-e-3.1-7B-Instruct	300.0	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	100.0
Qwen-e-3.1-16B-Instruct	500.0	100.0	33.6	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	100.0
MinG-3.2B-Instruct-2407	200.0	100.0	22.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.0	11.0	11.2	19.2	44.0
ep-4-1-min	500.0	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	100.0	100.0	100.0
ep-4-1-min	0.0	100.0	30.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Liam-e-4-Scout-1TB-16E-Instruct	0.0	100.0	42.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	21.0	42.0	100.0	100.0	100.0
ep-5-Minimo	0.0	100.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	79.0	86.8	74.0	74.0	18.4
Liam-e-4-Maverick-1TB-128E-Instruct-FP8	0.0	100.0	42.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	13.0	5.0	10.0	10.0	30.0
MinG-3.2B-Instruct-2409	0.0	100.0	28.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	21.0	42.0	100.0	100.0	100.0
Liam-e-3.1-8B-Instruct	100.0	100.0	2.8	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
DeepSeq-RL	0.0	100.0	42.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Qwen-e-2.5-7B-Instruct	0.0	100.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Liam-e-3.2-16B-Instruct	0.0	100.0	80.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
phi-4-1-piano	0.0	100.0	14.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	50.0	48.0	50.0	50.0	50.0
MinG-Small-24B-Instruct-2501	0.0	100.0	28.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
DeepSeq-RL-DiffLiana-70B	0.0	0.0	28.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	34.0	68.0	92.0	4.0	4.0
MinG-Small-3.1-24B-Instruct-2503	0.0	100.0	20.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	2.0	2.0	20.0	20.0	20.0
Liam-e-3.2-128E-Instruct-v0.1	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17.0	34.0	100.0	0.0	0.0
DeepSeq-RL-DiffLiana-70B	0.0	100.0	80.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	20.0	40.0	100.0	0.0	0.0
Phi-3.5-Mod-instruct	0.0	100.0	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
Phi-4-1-min	0.0	100.0	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
Phi-4-8K-Mod-instruct-0.1	0.0	100.0	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
Phi-3.5-medium-128E-instruct	0.0	100.0	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
Phi-3.5-medium-128E-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

Table 12: Model Performance on ALFWORLD tasks.

Table 13: Model Performance on SCIENCEWORLD tasks. Part 1

Models	Boil	ChangeTheStateOfMatterOf	ChemistryMix	ChemistryMixPaintSecondaryColor	FindAnimal	FindLivingThing	FindNonLivingThing	GrowFruit	GrowPlant	IdentifyLifeStages1	IdentifyLifeStages2	InclinedPlaneDetermineAngle
o3 (medium)	48.4	100.0	100.0	100.0	85.0	100.0	100.0	48.0	100.0	89.0	100.0	100.0
o3 (high)	68.0	100.0	100.0	100.0	88.2	100.0	100.0	53.0	44.4	78.0	100.0	100.0
gpt-5 (drinking)	95.6	100.0	100.0	100.0	82.0	95.0	100.0	100.0	53.8	45.2	65.4	100.0
o3 (low)	5.8	100.0	100.0	100.0	94.0	85.0	100.0	100.0	40.0	37.4	77.0	94.0
claudie-3.7-sonnet (thinking)	5.8	70.0	88.4	100.0	100.0	76.0	66.6	100.0	48.4	45.0	48.4	54.6
claudie-3.7-sonnet-latest	5.8	43.2	100.0	100.0	100.0	100.0	100.0	100.0	48.4	45.0	40.4	90.0
gpt-4.1	5.0	89.4	100.0	84.0	60.0	45.0	100.0	100.0	21.2	48.8	62.6	100.0
gpt-5-mini (thinking)	52.4	100.0	100.0	100.0	52.2	65.0	100.0	81.6	48.0	72.4	78.2	45.2
o1	21.4	80.8	100.0	100.0	82.2	91.6	96.6	100.0	100.0	37.4	62.4	100.0
gpt-4.0	2.8	75.0	28.0	22.0	100.0	90.0	100.0	100.0	0.0	27.0	11.4	35.0
claudie-3.5-hakku	40.0	33.0	72.8	86.0	76.6	58.2	100.0	100.0	100.0	18.4	20.0	82.0
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	12.0	100.0
gemin-2.0-dash	6.6	80.8	42.0	60.0	9.4	25.0	50.0	85.0	40.0	40.4	45.6	81.0
Qwen-3.2B	5.0	16.6	88.4	48.0	45.4	25.0	70.0	40.0	20.0	34.6	13.0	19.4
Llama-3.3-70B-Instruct	9.2	1.8	75.0	62.0	53.4	50.0	70.0	100.0	25.0	0.0	41.2	37.8
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	35.0
Owen-2.5-72B-instruct	0.0	0.0	8.0	100.0	10.0	36.6	5.0	100.0	70.0	0.0	13.2	50.0
Mistral-Large-instruct-2407	2.4	16.6	73.2	90.0	64.6	40.0	70.0	85.0	30.2	10.0	20.0	60.0
gpt-4.1-mini	2.2	16.4	48.2	48.0	42.0	40.0	35.0	100.0	25.0	9.6	5.6	56.0
gpt-4.0-mini	0.8	0.8	29.8	28.0	9.4	16.8	25.0	63.4	10.0	20.8	6.6	48.6
Llama-4-Scout-17B-16E-Instruct	0.0	0.0	33.0	20.0	7.0	25.0	83.0	63.0	0.0	12.0	7.0	20.0
gpt-5-mano	1.8	0.8	33.0	88.0	60.0	23.4	21.8	78.2	21.8	0.0	32.0	15.4
Llama-4-Mateck-17B-128E-Instruct-FP8	0.0	0.0	33.0	20.0	7.0	23.4	8.0	0.0	0.0	21.0	7.0	30.2
Mistral-Small-Instruct-2409	0.0	0.4	43.2	72.0	8.8	25.0	25.0	95.0	36.6	0.0	13.4	13.0
Llama-3.1-3B-Instruct	0.4	0.0	8.0	24.0	19.6	20.0	18.4	48.2	10.0	0.0	7.6	0.0
DeepSeek-R1	0.0	0.0	33.0	20.0	8.8	10.2	21.8	53.4	23.4	0.0	29.2	14.4
Owen-2.5-7B-Instruct	0.0	0.0	29.8	22.0	8.8	1.6	4.8	0.0	0.0	12.0	7.0	8.0
Llama-3.2-3B-Instruct	0.0	0.0	23.2	22.0	7.0	0.0	0.0	25.0	0.0	0.0	6.0	0.0
phi-4	0.0	0.0	8.0	52.0	17.4	13.6	10.0	0.0	3.4	0.0	12.0	5.0
gpt-4.1-piano	0.0	0.0	28.0	20.0	7.6	10.2	15.2	0.0	13.0	6.4	4.0	0.0
Mistral-Small-24B-Instruct-2501	0.0	0.0	33.0	20.0	7.0	5.0	17.0	17.0	3.4	0.0	12.0	14.4
DeepSeek-R1-Distill-Llama-70B	0.0	0.0	8.0	20.0	8.2	10.2	13.6	10.2	8.4	0.0	12.0	6.0
Mistral-Small-1.1-24B-Instruct-2410	0.0	0.0	8.0	20.0	7.0	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	11.4	36.0	7.0	10.0	0.0	0.0	15.0	0.0	11.0	6.4
Llama-3.2-1B-Instruct-v0.1	0.0	0.0	8.0	20.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3.5-mini-instruct	0.0	0.0	11.4	20.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3.5-mini-instruct-v0.1	0.0	0.0	8.0	20.0	7.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3.5-medium-128k-instruct	0.0	0.0	8.0	20.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-3.5-medium-128k-instruct	0.0	0.0	8.0	20.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 14: Model Performance on SCIENCE WORLD tasks. Part 2.

1350
 1351
 1352
 1353
 1354
 1355
 1356
 1357
 1358
 1359
 1360
 1361
 1362
 1363
 1364
 1365
 1366
 1367
 1368
 1369
 1370
 1371
 1372
 1373
 1374
 1375
 1376
 1377
 1378
 1379
 1380
 1381
 1382
 1383
 1384
 1385
 1386
 1387
 1388
 1389
 1390
 1391
 1392
 1393
 1394
 1395
 1396
 1397
 1398
 1399
 1400
 1401
 1402
 1403

Table 15: Model Performance on JERICHO games(tasks), part 1.

Models	905	Advent	Adventureland	Afflicted	Awaken	Balances	Ballyhoo	Curses	Cuthroat	Deepphone	Detective	Dragon	Enchanter	Enter	Gold	Higg	Hundark	Infiel	Inhumane	Jewel	Kam	Library	Loose	Louping	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	60	25.7	12.0					
o3 (high)	100.0	100.0	20.7	14.0	17.3	2.0	0.0	27.5	6.0	1.1	12.0	10.3	54.4	6.4	6.0	74.0	9.6	3.2	0.0	1.2	28.9	0.4	4.7	7.6	31.4	12.3						
gp4-5 (thinking)	100.0	100.0	27.7	9.8	26.4	2.4	0.0	27.5	0.5	16.8	36.7	13.8	38.3	4.8	12.8	65.0	12.0	2.5	0.0	1.2	35.6	1.1	5.3	39.3	60	40.0	11.2					
o3 (low)	80.0	40.0	23.2	19.2	1.2	29.4	0.0	29.5	2.0	0.3	16.8	40.0	13.8	72.0	6.0	25.0	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 (atest)	40.0	6.7	33.3	19.4	35.5	1.6	4.0	39.2	0.0	0.0	12.0	90	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					
gp4-1	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					
gp4-5-mini (thinking)	40.0	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					
gp4-5-mini (high)	40.0	26.0	10.3	5.1	0.4	0.0	19.6	2.0	0.2	5.6	9.3	31.1	5.6	9.8	58.6	1.2	2.5	0.0	0.0	11.1	0.9	1.8	11.3	40.0	40.0	25.7						
gp4-6	80.0	66.7	13.9	1.4	1.1	0.0	0.0	23.9	0.0	0.2	9.6	6.7	27.2	4.8	12.5	22.0	0.0	2.5	0.0	1.0	4.4	0.0	1.8	34.7	0.0	4.3	7.9					
gp4-6	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	7.2	3.2	38.0	7.8	2.5	0.0	1.2	4.4	0.0	1.2	13.3	0.8	5.7	2.5					
claude-3.5-hakku	0.0	0.0	20.0	10.3	0.0	0.0	0.0	19.6	0.0	0.0	2.4	5.7	25.0	0.8	3.2	3.0	0.0	2.5	0.0	0.0	0.4	0.0	0.0	0.0	0.0	11.4	2.5					
Llama-3.1-40B-Instruction	0.0	0.0	11.7	0.0	4.5	0.0	0.0	19.6	0.0	0.0	5.6	5.6	50.0	4.0	4.0	15.0	0.0	1.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	14.3	2.5					
gemin-2.0-flash	0.0	0.0	10.3	0.0	0.0	0.8	0.0	11.8	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.1	28.3	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.9				
Qwen-3.2B	0.0	6.7	10.3	0.0	3.5	0.4	0.0	5.9	0.0	0.0	1.6	5.9	31.1	3.2	3.8	3.0	0.0	2.5	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	11.4	3.2				
Llama-3.3-70B-Instruction	0.0	0.0	10.3	0.0	0.0	1.2	0.0	11.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.3	36.1	4.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.3			
Llama-3.1-70B-Instruction	0.0	0.0	10.3	0.0	0.8	0.0	0.0	19.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.6	28.9	4.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.3			
Llama-3.1-70B-Instruction	0.0	0.0	10.3	0.0	0.0	0.0	0.0	11.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.5	11.1	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.3			
Qwen-2.0-Scout	0.0	0.0	10.3	0.0	0.0	0.0	0.0	13.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.2	41.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.3			
gp4-mini	0.0	0.0	10.3	0.0	0.0	0.0	0.0	9.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.2	3.4	8.3	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
Llama-3.4-Scout-17B-16E-Instruction	0.0	0.0	10.3	0.0	0.0	0.0	0.0	2.7	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	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
Qwen-4-Maverick-17B-128E-Instruction	0.0	0.0	10.3	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	0.0	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	
Qwen-2.0-Scout-2409	0.0	0.0	10.3	0.0	0.5	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	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-Large-Instruction-2407	60.0	0.0	10.3	0.0	1.6	0.0	0.0	11.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.5	33.3	8.0	0.0	0.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
gp4-4-mini	0.0	0.0	10.3	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	0.0	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
Llama-3.4-Scout-17B-16E-Instruction	0.0	0.0	10.3	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	0.0	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
gp4-piano	0.0	0.0	10.3	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	0.0	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
Mistral-Small-Instruction-2501	0.0	0.0	10.3	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	0.0	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
DeepSeek-RL-Distill-Llama-70B	0.0	0.0	10.3	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	0.0	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
Mistral-8B-Instruction-2410	0.0	0.0	10.3	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	0.0	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
Mistral-5.5-7B-Instruction-2503	0.0	0.0	10.3	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	0.0	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
Mistral-5.5-22B-Instruction-v0.1	0.0	0.0	10.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	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
Llama-3.2-1B-Instruction	0.0	0.0	10.3	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	0.0	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
Phi-3-mini-128E-instruct	0.0	0.0	10.3	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	0.0	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
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	0.0	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
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	0.0	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
Phi-4.5-8X-B-Instruction-0.1	0.0	0.0	10.3	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	0.0	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
Phi-3.5-mini-128E-instruct	0.0	0.0	10.3	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	0.0	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
Phi-3.5-medium-128E-instruct	0.0	0.0	10.3	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	0.0	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

Table 16: Model Performance on JERICHO games(tasks), part 2.

Models	Lurking	Moonlit	Murdac	Night	Omniguest	Partyfoul	Planetfall	Plundered	Reverb	Seastalker	Sherlock	Sorceror	Spellbktr	Temple	Trinity	Tryst205	Weapon	Whabinger	Yomomma	Zenon	Zork1	Zork2	Zork3	Zuuu								
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	2.4	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					
o3 (low)	5.0	0.0	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	17.8	1.0	0.0	34.3	6.0	48.6	3.0					
clade-3:7-sonnet (thinking)	5.0	0.0	6.8	24.0	14.0	0.0	12.9	7.5	3.2	24.8	11.2	9.2	40.0	7.0	10.7	1.3	16.0	14.2	2.9	0.0	13.4	2.9	0.0	14.4	0.0	40.0	5.0					
clade-3:7-sonnet (thinking)	5.0	0.0	11.3	20.0	10.0	0.0	7.1	4.5	4.0	24.0	18.6	10.6	36.0	5.3	11.3	1.0	14.3	11.8	0.6	0.0	11.6	0.0	0.0	12.2	0.0	40.0	5.0					
clade-3:7-sonnet (thinking)	5.0	0.0	10.8	20.0	10.0	0.0	7.1	3.8	2.4	4.0	20.6	3.6	32.0	9.2	11.7	1.4	14.3	13.8	0.0	0.0	13.0	0.0	0.0	13.5	0.0	42.9	5.0					
clade-3:5-sonnet-latest	2.0	0.0	8.7	0.0	10.0	0.0	10.0	6.8	6.4	0.0	11.8	3.2	12.0	9.2	11.7	1.4	2.9	6.0	0.3	0.0	6.8	0.0	0.0	11.7	0.0	20.0	6.0					
gp-4.1	5.0	0.0	5.6	0.0	10.0	0.0	4.3	6.0	1.6	0.0	1.8	3.2	12.0	7.2	24.0	7.0	9.7	0.5	40.0	4.2	1.0	0.0	12.5	1.2	0.0	12.5	1.0	31.4	1.0			
gp-5-mini (thinking)	5.0	0.0	8.7	8.0	16.0	0.0	7.1	2.2	2.4	6.8	10.4	7.6	8.8	7.2	40.0	4.2	6.7	1.0	16.6	6.4	0.0	0.0	13.0	2.3	0.0	7.9	1.0	42.9	4.0			
ol	7.0	0.0	8.6	0.0	12.0	0.0	8.6	3.8	4.0	8.8	3.8	3.2	8.8	20.0	3.2	8.0	0.8	20.0	7.8	9.0	1.1	11.4	9.4	1.7	0.0	13.8	0.6	0.0	12.5	1.0	42.9	9.6
gp-4.0	5.0	0.0	5.6	0.0	10.0	0.0	8.6	7.5	3.2	24.8	11.2	9.2	40.0	7.0	10.7	1.3	16.0	14.2	0.9	0.0	9.6	0.0	0.0	14.5	0.0	31.4	0.0					
clade-3:5-haraku	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	0.8	0.0	9.0	0.0	0.0	11.8	1.7	0.0	13.1	0.0	31.4	0.0					
Liana-3:1-405B-Instinct	5.0	0.0	5.6	20.0	10.0	0.0	7.1	3.8	0.0	0.0	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					
gemin-2:0-flash	5.0	0.0	5.2	20.0	10.0	0.0	4.3	3.8	1.6	0.0	8.2	10.4	40.0	1.2	6.7	1.3	0.0	4.0	0.0	0.0	11.6	0.0	0.0	25.7	1.0	0.0	0.0	0.0				
Qwenz-3:2B	3.0	0.0	4.4	0.0	8.0	0.0	5.7	0.0	4.0	0.0	17.4	2.4	0.0	1.2	4.2	0.0	4.8	0.0	0.0	7.4	0.0	0.0	42.9	4.0	0.0	0.0	0.0					
Liana-3:3-70B-Instuct	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	4.4	0.0	0.0	11.6	0.0	0.0	4.3	0.0	31.4	3.0					
Liana-3:1-70B-Instuct	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	16.0	8.4	0.0	0.0	6.0	0.0	0.0	3.1	0.0	42.9	13.0					
Qwenz-3:7B-Instuct	5.0	0.0	4.6	4.0	10.0	0.0	3.8	0.0	0.0	9.8	3.6	16.0	1.2	0.0	0.8	5.0	0.0	2.8	0.0	0.0	6.0	0.0	0.0	12.8	1.0	28.6	0.0					
Misra-Large-Instuct-2407	5.0	0.0	5.6	0.0	10.0	0.0	2.9	0.8	0.0	0.0	3.2	8.8	24.0	2.0	7.2	1.6	0.0	5.0	0.0	0.0	6.4	0.0	0.0	6.9	0.0	28.6	1.0					
gp-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	3.0	0.0	0.0	7.4	0.0	0.0	3.1	2.0	5.7	0.0	0.0	0.0				
gp-4.0-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	4.0	0.0	0.0	6.0	0.0	0.0	1.4	0.0	42.9	0.0					
Liana-4:Scout-17B-16E-Instuct	5.0	0.0	5.0	0.0	10.0	0.0	0.0	3.8	0.0	0.0	3.0	1.0	0.0	0.0	1.2	4.2	0.0	0.0	6.0	0.0	0.0	6.0	0.0	0.0	28.6	0.0	0.0	0.0				
gp-5-mini	4.0	0.0	5.0	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	6.0	0.0	0.0	28.6	0.0				
Liana-4:Maverick-17B-128E-Instuct-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	1.4	0.0	0.0	28.6	0.0				
Misra-Small-Instuct-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	0.0	1.2	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	2.0	0.0	0.0	17.1	2.0				
DeePS-Seek-R1	3.0	0.0	3.4	0.0	10.0	0.0	3.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	2.9	0.0	0.0	0.0	0.0				
Liana-3:1-3B-Instuct	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.8	0.0	0.0	2.6	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0				
Qwenz-3:5-7B-Instuct	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	0.0	1.2	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	1.4	0.0	0.0	0.0	0.0				
Liana-3:2-3B-Instuct	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	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0				
phi-4	1.0	0.0	0.0	0.0	10.0	0.0	0.0	3.0	0.0	0.0	2.6	0.0	0.0	0.0	1.2	0.0	0.0	0.2	0.0	0.0	6.0	0.0	0.0	2.0	0.0	0.0	22.9	0.0				
Misra-Small-Instuct-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	0.0	1.2	0.0	0.0	0.2	5.7	1.6	0.0	0.0	4.0	0.0	0.0	14.3	0.0	0.0	0.0			
DeePS-Seek-R1-Distill-14:18-7B	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	0.0	1.2	0.0	0.0	0.0	2.9	0.0	0.0	0.0	4.0	0.0	0.0	14.3	0.0	0.0	0.0			
Misra-Small-3:1-24B-Instuct-2503	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0				
Misra-1:8x2B-Instuct-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	0.0	0.3	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				
Liana-3:2-1B-Instuct	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	0.0	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	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	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	10.0	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				
Phi-3:mini-128k-Instuct	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	0.0	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-3:4-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	0.0	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-MoE-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	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-3:3-medium-128k-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	0.0	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				

1458

1459 Table 17: Average reasoning failures for the Claude family of models across 8 games (32 logs total).

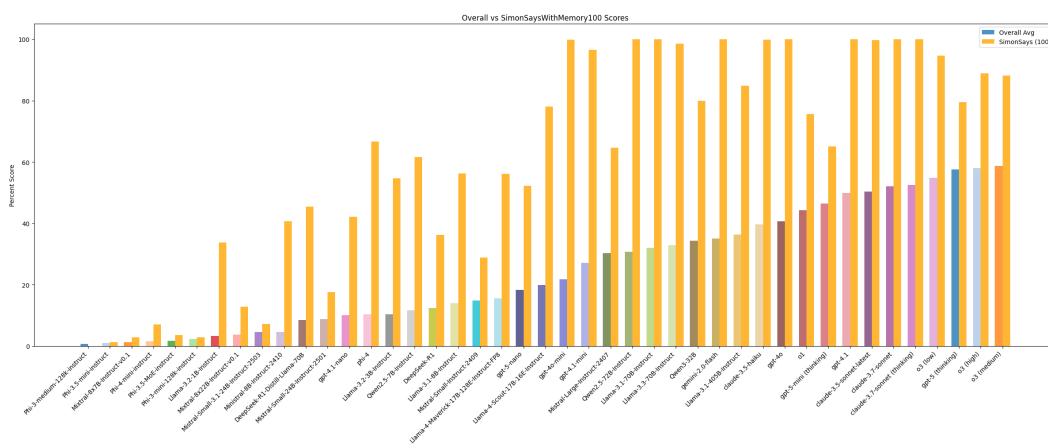
Model	Failures(↓)					TALES Score (↑)
	Spatial	Deductive	Inductive	Grounded	Total	
3.7-Sonnet (Thinking)	1.29	1.29	2.14	1.0	5.71	52.5
3.7-Sonnet	0.86	3.57	5.43	1.0	10.86	52.1
3.5-Sonnet	1.14	1.43	9.43	0.86	12.86	50.4
3.5-Haiku	3.86	7.14	25.29	2.86	39.14	39.6

1466

1467

K AVERAGE SIMON SAYS SCORE VERSUS OVERALL TALES SCORE

1469



1484

1485 Figure 3: 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.
1486

1487

1488

1489

L COMPUTE

1490

1491 All experiments were run intermittently over a course of roughly six months. Open-weight models
1492 were run on a combination of a four node cluster of 8xMI300s and one node of 8xA100s. Anthropic
1493 API experiments accrued a cost of 1,562.15 U.S. dollars. OpenAI API experiments accrued an
1494 estimated cost of 6,870.76 U.S. dollars.
1495

1496

1497

M REASONING FAILURES

1498

1499 In this section, we provide explicit reasoning failure numbers from a smaller subset of the top
1500 performing LLMs, similar to what is shown in Figure 1. We focus again on the Claude family
1501 of models due to having access to a range of model performances as well as both thinking and
1502 non-thinking modes for their top model. We outline the specific criteria we used to classify actions as
1503 a specific type of reasoning failure. From a selection of 32 game logs of the most difficult games in
1504 the frameworks discussed in Section 5, we exhaustively label all actions that exhibit some reasoning
1505 failure, discuss our criteria for designating a chosen action as a reasoning failure, and present the
1506 cumulative, average failures per LLM, finding that **inductive reasoning failures** where the agent
1507 fails to account for explicit or implicit feedback are by far the most common.

1508

1509 As shown in Table 17, **top LLMs still make reasoning errors but resolve them quickly enough to**
1510 **still succeed.** When explicitly labeling and enumerating reasoning errors made on a per-action basis,
1511 we see a clear trend with the worse performing LLMs also making a larger number of reasoning
errors. These reasoning errors ultimately waste a large portion of the weaker LLM’s allotted steps
for the environments, leading to a decreased performance. The stronger LLMs, while still making
reasoning errors, make few enough errors to still have sufficient time to complete the tasks.

1512 Claude-3.5-Haiku frequently encounters inductive reasoning failures as weaker models tend to get
 1513 trapped in cyclical action loops that do not result in any meaningful change in state. We consider
 1514 each repeated step in this cycle as a distinct failure because each step represents an instance where
 1515 the agent fails to incorporate feedback. See Appendix Q for our labeling criteria and examples.
 1516

1517 N 1000 STEPS OF ZORK1

1518 While even the best performing LLMs make reasoning mistakes, we find what allows them to still
 1519 find success in TALES is the ability to both avoid making an excessive number of these mistakes
 1520 and the ability to self-correct. We argue that 100 steps are sufficient to evaluate the performance of
 1521 current state-of-the-art LLMs because even the best overall LLM, o3, fails to approach the maximum
 1522 possible score within 100 steps for ZORK1 (13.8% vs 29.1%). In this experiment, we explore whether
 1523 any of the top models can achieve a score comparable to the walkthrough after 100 steps in ZORK1
 1524 while allowed to run for 1000 steps with the entire history kept in the context. If so, we examine the
 1525 behaviors that enable this and determine the required number of steps. We select the overall top 3
 1526 performing models, o3 and both the thinking and non-thinking modes of Claude-3.7-Sonnet.
 1527

1528 **Scores improve slightly, but the best LLMs are still far from the walkthrough score even with
 1529 10 times more steps.** That is with 1000 steps, the best LLMs fail to reach 29.3% of the total score.
 1530 o3 manages to achieve a score of 20.9%, a performance increase of only 7.6% over its original
 1531 score of 13.3% for 10 times the steps. Claude-3.7 non-thinking⁵ and thinking achieve 16.9% and
 1532 15.3% respectively. The key behavior pattern we see in both thinking models is a slightly directed,
 1533 random exploration of the area of the game past the bottleneck that stops other, weaker LLMs. This
 1534 exploration is far less focused than the iterative search we see agents perform early in the game and
 1535 in simpler environments such as AW.

1536 We tried allowing Claude-3.7-Sonnet to think for up to 4096 tokens. However, the model never uses
 1537 more than 700 tokens for its thinking, a similar value seen for the rest of the benchmark where the
 1538 thinking effort is capped to 1024. This is a significantly smaller thinking effort than o3 which uses
 1539 up to 5000 thinking tokens throughout its 1000 steps ZORK1 playthrough. This suggests that o3's
 1540 performance is due to a willingness to leverage many more thinking tokens at any particular step.
 1541 However, the highest thinking efforts do not appear to occur at any significant points during gameplay
 1542 and we are unable to verify the actual contents since we do not have access to the thinking traces.
 1543

1544 O RL FINE-TUNING HYPERPARAMETERS

1545 We use Qwen3 rather than Qwen2.5 due to its inherit reasoning training. Following guidance located
 1546 in the appendix of their publication, for verl-agent we use a learning rate of 1e-6 for the actor and
 1547 1e-5 for the critic. We train over a total of 150 'epochs' which verl-agent represents as one full rollout
 1548 of the environment. To more closely adhere to our zero-shot evaluations, we keep the entire history in
 1549 the trajectory up to 50 steps. This necessitates a far smaller batch size of 8. During evaluation, we
 1550 use a batch size of 32 to sufficiently sample all evaluation environments. We use the collection of
 1551 all existing game seeds and formats not evaluated on in our zero-shot experiments as a training set
 1552 and the specific environments from the zero-shot experiments as the test set. We allocate a thinking
 1553 budget of 512 tokens. In their ALFWORLD implementation, (Feng et al., 2025) use a terminal reward
 1554 of 10. In our RL experiments, we use the native reward from the environment, as our motivation is
 1555 not to contest (Feng et al., 2025)'s results but to evaluate the effects of including privileged or domain
 1556 knowledge directly in the prompt.
 1557

1559 P FRAMEWORK ENVIRONMENT SUBSELECTION

1561 P.1 TEXTWORLD

1563 For TEXTWORLD, we use the following environments:

1564
 1565 ⁵Despite scoring higher than its thinking variant, zero-shot Claude-3.7-Sonnet suffers a catastrophic inductive
 1566 reasoning failure by repeatedly issuing the quitting commands after step 479.

```

1566 test/difficulty_level_1/tw-cooking-recipe1+take1+open-0nQyHWbvh6d_]
1567   ↳ XFPmhLKX.z8
1568 test/difficulty_level_2/tw-cooking-recipe1+take1+cook+open-0nQyHW_]
1569   ↳ bvh6dXFPmhLKX.z8
1570 test/difficulty_level_3/tw-cooking-recipe1+take1+cut+open-0nQyHWb_]
1571   ↳ vh6dXFPmhLKX.z8
1572 test/difficulty_level_4/tw-cooking-recipe1+take1+open+go6-0nQyHWb_]
1573   ↳ vh6dXFPmhLKX.z8
1574 test/difficulty_level_5/tw-cooking-recipe1+take1+open+go9-0nQyHWb_]
1575   ↳ vh6dXFPmhLKX.z8
1576 test/difficulty_level_6/tw-cooking-recipe1+take1+open+go12-0nQyHW_]
1577   ↳ bvh6dXFPmhLKX.z8
1578 test/difficulty_level_7/tw-cooking-recipe1+take1+cook+cut+open-0n_]
1579   ↳ QyHWbvh6dXFPmhLKX.z8
1580 test/difficulty_level_8/tw-cooking-recipe3+take3+open+go6-0nQyHWb_]
1581   ↳ vh6dXFPmhLKX.z8
1582 test/difficulty_level_9/tw-cooking-recipe3+take3+cook+cut+open+go_]
1583   ↳ 6-0nQyHWbvh6dXFPmhLKX.z8
1584 test/difficulty_level_10/tw-cooking-recipe3+take3+cook+cut+open+g_]
1585   ↳ o12-0nQyHWbvh6dXFPmhLKX.z8
1586
1587 P.2 TEXTWORLDEXPRESS
1588
1589 For TEXTWORLDEXPRESS, we use the game parameters:
1590
1591 TASKS = [
1592   (
1593     "CookingWorld",
1594     "cookingworld",
1595     "numLocations=1, numIngredients=2, numDistractorItems=5,
1596     ↳ includeDoors=0, limitInventorySize=0",
1597   ),
1598   (
1599     "TextWorldCommonsense",
1600     "twc",
1601     "numLocations=1, numItemsToPutAway=1, includeDoors=0, limitI_]
1602     ↳ nventorySize=0",
1603   ),
1604   (
1605     "CoinCollector",
1606     "coin",
1607     "numLocations=1, numDistractorItems=5,
1608     ↳ limitInventorySize=0",
1609   ),
1610   ("Arithmetic", "arithmetic", ""),
1611   (
1612     "MapReader",
1613     "mapreader",
1614     "numLocations=2, maxDistanceApart=1,
1615     ↳ maxDistractorItemsPerLocation=2, includeDoors=0,
1616     ↳ limitInventorySize=0",
1617   ),
1618   ("Sorting", "sorting", ""),
1619   ("SimonSays10", "simonsays", "gameLength=10, numDistractors=4,
1620     ↳ memorization=0"),
1621   ("SimonSays50", "simonsays", "gameLength=50, numDistractors=4,
1622     ↳ memorization=0"),
1623   ("SimonSays100", "simonsays", "gameLength=100,
1624     ↳ numDistractors=4, memorization=0"),

```

```

1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
  (
    "SimonSaysWithMemory10",
    "simonsays",
    "gameLength=10, numDistractors=4, memorization=1,
     ↳ verbose=0",
  ),
  (
    "SimonSaysWithMemory50",
    "simonsays",
    "gameLength=50, numDistractors=4, memorization=1,
     ↳ verbose=0",
  ),
  (
    "SimonSaysWithMemory100",
    "simonsays",
    "gameLength=100, numDistractors=4, memorization=1,
     ↳ verbose=0",
  ),
  (
    "SimonSaysWithMemory10Verbose",
    "simonsays",
    "gameLength=10, numDistractors=4, memorization=1,
     ↳ verbose=1",
  ),
  (
    "SimonSaysWithMemory50Verbose",
    "simonsays",
    "gameLength=50, numDistractors=4, memorization=1,
     ↳ verbose=1",
  ),
  (
    "SimonSaysWithMemory100Verbose",
    "simonsays",
    "gameLength=100, numDistractors=4, memorization=1,
     ↳ verbose=1",
  ),
  ("PeckingOrder", "peckingorder", ""),
]

```

P.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.

```

1662 valid_seen/pick_and_place_simple-Book-None-SideTable-329/trial_T2 ]
1663   ↳ 0190908_050633_745514
1664 valid_seen/look_at_obj_in_light-AlarmClock-None-DeskLamp-323/tria ]
1665   ↳ l_T20190909_044715_250790
1666 valid_seen/pick_clean_then_place_in_recep-ButterKnife-None-Counte ]
1667   ↳ rTop-8/trial_T20190909_105559_983897
1668 valid_seen/pick_heat_then_place_in_recep-Apple-None-DiningTable-2 ]
1669   ↳ 6/trial_T20190907_060234_011675
1670 valid_seen/pick_cool_then_place_in_recep-Apple-None-CounterTop-14 ]
1671   ↳ /trial_T20190909_044933_815840
1672 valid_seen/pick_two_obj_and_place-AlarmClock-None-Dresser-305/tri ]
1673   ↳ al_T20190907_165826_194855
1674 valid_unseen/pick_and_place_simple-Mug-None-Desk-308/trial_T20190 ]
1675   ↳ 908_125200_737896

```

```

1674 valid_unseen/look_at_obj_in_light-AlarmClock-None-DeskLamp-308/tr_
1675   ↳ ial_T20190908_222917_366542
1676 valid_unseen/pick_clean_then_place_in_recep-Bowl-None-Cabinet-10/_
1677   ↳ trial_T20190909_061130_844814
1678 valid_unseen/pick_heat_then_place_in_recep-Apple-None-Fridge-10/t_
1679   ↳ rial_T20190906_182259_116320
1680 valid_unseen/pick_cool_then_place_in_recep-Bread-None-CounterTop-_
1681   ↳ 10/trial_T20190908_091747_866951
1682 valid_unseen/pick_two_obj_and_place-CD-None-Safe-308/trial_T20190_
1683   ↳ 907_050942_897916
1684

```

1685 P.4 SCIENCEWORLD

1686 We use the first variation of the test set for each of the 30 tasks.

1689 Q ANNOTATED LOGS: CRITERIA AND EXAMPLES

1691 Q.1 LABELING CRITERIA

1693 For spatial reasoning failures, we primarily label those actions that involve failing navigation through
 1694 path finding, backtracking, or iteratively searching spaces for necessary elements for completing the
 1695 assigned task. Deductive reasoning failures were primarily actions that involved interaction with
 1696 game elements that cannot be reasonably linked back to the target objective or failures to solve
 1697 in-game puzzles. Inductive reasoning failures were comprised mostly of actions that the agent had
 1698 previously attempted in a near-identical state to negative feedback. If there was any non-trivial change
 1699 in game state compared to when the action was first attempted, we did not label the action as an
 1700 inductive reasoning failure if the action did not fail due to a parser issue. Grounded reasoning failures
 1701 encompassed what is commonly referred to as hallucination: attempts to interact with game elements
 1702 not mentioned in observations or feedback.

1704 Q.2 SPATIAL REASONING FAILURE EXAMPLE

1706 Task Description:Your task is to melt lead. First, focus on the
 1707 ↳ substance. Then, take actions that will cause it to change its
 1708 ↳ state of matter...

1709 ...

1710 This room is called the kitchen. In it, you see:

```

1711   the agent
1712   a substance called air
1713   a chair. On the chair is: nothing.
1714   a counter. On the counter is: a bowl (containing a red
1715   ↳ apple, a banana, an orange, a potato), a drawer.
1716   a cupboard. The cupboard door is closed.
1717   a freezer. The freezer door is closed.
1718   a fridge. The fridge door is closed.
1719   a glass jar (containing a substance called sodium
1720   ↳ chloride)
1721   a lighter
1722   a oven, which is turned off. The oven door is closed.
1723   a painting
1724   a sink, which is turned off. In the sink is: nothing.
1725   a substance called soap
1726   a stopwatch, which is deactivated.
1727   a stove, which is turned off. On the stove is: nothing.
1728   a table. On the table is: a glass cup (containing
1729   ↳ nothing).

```

```

1728     a thermometer, currently reading a temperature of 10
1729     ↳ degrees celsius
1730 You also see:
1731     A door to the bathroom (that is open)
1732     A door to the hallway (that is closed)
1733     A door to the outside (that is closed)
1734 >
1735 Action: open oven
1736
1737 ...
1738 Action: go hallway
1739
1740 ...
1741 Action: go workshop
1742
1743 ...
1744 This room is called the workshop. In it, you see:
1745     the agent
1746     a substance called air
1747     a table. On the table is: a battery, a green wire, a metal
1748     ↳ pot (containing a substance called lead), a orange
1749     ↳ light bulb, which is off, a orange wire, a red wire, a
1750     ↳ switch, which is off, a violet light bulb, which is
1751     ↳ off, a yellow light bulb, which is off.
1752     a ultra low temperature freezer. The ultra low temperature
1753     ↳ freezer door is closed.
1754 You also see:
1755     A door to the hallway (that is open)
1756 >
1757 Action: pick up metal pot
1758
1759 You move the metal pot to the inventory.
1760
1761 Action: go kitchen
1762
1763 Above we provide an example of a spatial reasoning failure from Claude-3.5-Haiku in ScienceWorld-
1764 Melt. The agent navigates to the workshop, through the hallway and thus should realize the path is
1765 kitchen to hallway to workshop. However, despite this, the agent attempts the action go kitchen.
1766
1767 Q.3 DEDUCTIVE REASONING FAILURE EXAMPLE
1768
1769 Task Description:Your task is to freeze lead. First, focus on the
1770     ↳ substance. Then, take actions that will cause it to change its
1771     ↳ state of matter.
1772
1773 ...
1774 Action: look around
1775
1776 This room is called the kitchen. In it, you see:
1777     the agent
1778     a substance called air
1779     a chair. On the chair is: nothing.
1780     a counter. On the counter is: a bowl (containing a red
1781     ↳ apple, a banana, an orange, a potato), a drawer.
1782     a cupboard. The cupboard door is closed.

```

1782 a freezer. The freezer door is closed.
 1783 a fridge. The fridge door is closed.
 1784 a glass jar (containing a substance called sodium
 1785 → chloride)
 1786 a lighter
 1787 a oven, which is turned off. The oven door is closed.
 1788 a painting
 1789 a sink, which is turned off. In the sink is: nothing.
 1790 a substance called soap
 1791 a stopwatch, which is deactivated.
 1792 a stove, which is turned off. On the stove is: nothing.
 1793 a table. On the table is: a glass cup (containing
 1794 → nothing).
 1795 a thermometer, currently reading a temperature of 10
 1796 → degrees celsius
 1797 You also see:
 1798 A door to the bathroom (that is open)
 1799 A door to the hallway (that is closed)
 1800 A door to the outside (that is closed)

1801 Action: open freezer

1802
 1803 Above we provide an example of a spatial reasoning failure from Claude-3.7-Sonnet (zero-shot)
 1804 in ScienceWorldFreeze. While looking for lead, the agent checks a kitchen freezer, an extremely
 1805 unlikely place to find lead.
 1806

1807 Q.4 INDUCTIVE REASONING FAILURE EXAMPLE

1808 ...

1810 If you pray enough, your prayers may be answered.

1812 Action: pray

1814 If you pray enough, your prayers may be answered.

1816 Action: pray

1818 If you pray enough, your prayers may be answered.

1820 Action: pray

1822 If you pray enough, your prayers may be answered.

1824 Action: pray

1825 ...

1827 Above we provide an example of an inductive reasoning failure from Claude-3.5-Haiku in ZORK1.
 1829 While the game environment returns a signal that praying might result in some change, Claude-3.5-
 1830 Haiku continues to do this for 36 steps, far past when most reasonable players would have moved
 1831 on.
 1832

1833 Q.5 GROUNDED REASONING FAILURE EXAMPLE

1834
 1835 This room is called the workshop. In it, you see:
 the agent

1836 a substance called air
 1837 a table. On the table is: a battery, a green wire, a
 1838 → orange light bulb, which is off, a orange wire, a red
 1839 → wire, a switch, which is off, a violet light bulb,
 1840 → which is off, a yellow light bulb, which is off.
 1841 a ultra low temperature freezer. The ultra low temperature
 1842 → freezer door is closed.

1843 You also see:

1844 A door to the hallway (that is open)

1845

1846 ...

1847 Action: go to hallway

1849 You move to the hallway.

1850

1851 ...

1853 Action: go to hallway

1855 You move to the hallway.

1856 Action: go to kitchen

1858 You move to the kitchen.

1860 Action: use lighter on lead

1862 The lighter heats up the lead a small amount.

1863

1864 Action: take red wire

1865

1866

1867

1868 Above we provide an example of an inductive reasoning failure from Claude-3.7-Sonnet (Thinking
 1869 in ScienceWorldBoil. We see the agent attempt to interact with an item in its context despite having
 1870 moved to another location where the item is no longer present.

1871

1872 R DATA CONTAMINATION EVIDENCE FROM THINKING TRACES: EXAMPLES

1873

1874 The blue glow of my sword suggests there's a danger nearby, likely
 1875 → something from Zork lore (like a grue). Let me explore the
 1876 → available paths carefully. I have the lantern on, which should
 1877 → protect me from grues in dark areas. Let me try going north
 1878 → first.

1879

1880 An example of a thinking trace from the Claude-3.7-Sonnet thinking mode that leads us to strongly
 1881 suspect data contamination. Most notably, the thinking traces references a *grue*, a creature from
 1882 ZORK1 mythology that had not at all been previously mentioned in the game's history.

1883

1884 S HYPERPARAMETERS

1885

1886 The following is the command used to run experiments across 5 seeds. All arguments not included
 1887 are default values.

1888

1889

for i in {{1..5}}; do

```
1890 python benchmark.py --agent agents/llm.py zero-shot --env
1891     ↳ jericho scienceworld textworld textworld_express alfworld
1892     ↳ --llm $LLM --context 100 --nb-steps 100 --conversation
1893     ↳ --wandb --seed "20241106$((i))" &
1894     sleep 60
1895 done
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
```