TALES: TEXT ADVENTURE LEARNING ENVIRONMENT SUITE

Anonymous authorsPaper under double-blind review

ABSTRACT

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

1 Introduction

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

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

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

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



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 ECBD 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 a suite of 122 games. 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 games meant to be played by humans, slowly and iteratively over extended periods of time. Despite strong evidence of data contamination (See Appendix P) 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 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 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.

2 REASONING

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

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

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

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

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

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

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

We believe text-adventure games are an ideal environment to simulate the challenges of real-world tasks, i.e., evaluating an agent on all four core reasoning skills at the same time. A deductive failure may lead an agent to waste a large number of steps trying to accomplish a sub-goal that does not progress towards the objective. Inductive reasoning failures may cause an agent to repeatedly attempt the same action over and over without success. Figure 1 illustrates a simple task in a text-adventure game where multiple reasoning skills are required at each step and a single failure leads to later errors. We discuss explicit failure modes that result from the lack of these core reasoning skills in Section 5 and design choices meant to allow for better evaluation in Section 5.4

Broadly speaking, interactive text-game environments are ideal test-beds for exploring highly complex, interactive and grounded scenarios where previous work has shown the potential for cross-domain 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 length	13.70	33.06	5.83	41.67	87.15
Informative feedback	✓	✓	×	✓	✓
Intermediate rewards	✓	✓	X	✓	✓
Nearest-neighbor parser	✓	×	X	✓	✓
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.

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. Some environments use a nearest-neighbor parser which can understand similar action phrases to mean the same thing, e.g., take lamp, get lamp and pick up lamp. Most environments provide intermediate score where players receive scores throughout the game. 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 framework 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. We do not use all environments within each frameworks. See Appendix N 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 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 receive 1 point after completing sub-goals related to the task in the game.

https://competitions.codalab.org/competitions/21557

TEXTWORLDEXPRESS (TWX) (Jansen & Côté, 2022) is a highly optimized re-implementation of many TW game scenarios that runs approximately three orders of magnitudes faster compared to the TW counterparts. We opt to use TWX over TW for the performance improvement where applicable. While significantly faster, an arguable drawback of using TWX over TW is also in its stricter parser. TWX simplifies its parser for speed and thus does not allow for nearest-neighbor action phrases.

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

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

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

4 EVALUATION

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

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

For the main results, we do not provide any other instructions to the LLMs on how to play the game. We aim to measure LLMs' raw capabilities exempt of inductive bias coming from human expert with domain knowledge. In Section 5.4, we explore the effects of providing some domain knowledge to the agent. When calling the LLMs, the observation and feedback are provided as the *user* inputs while the LLM actions are recorded as the *assistant* outputs.

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

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

Table 2: Average scores per framework and total TALES score for the top 10 models. Almost uniformly, reasoning LLMs outperform non-reasoning LLMs in all frameworks. However, a higher thinking budget does not always lead to better overall results.

Model	TEXTWORLD	TEXTWORLDEXPRESS	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

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

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

5 ANALYSIS

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

Following our qualitative analysis of top LLM behaviors, we provide a short analysis of the strengths and weaknesses of the Claude-3.7-Sonnet thinking mode in contrast to the non-thinking mode. We focus on 3.7-Sonnet due to its performance in both modes as well as accessibility to the thinking traces. We then take a representative subset of the game logs from all Claude models, explicitly labeling reasoning failures in a per-step granularity and provide the aggregated results over 32 hand-annotated logs. Finally, we investigate whether the 100 steps limit for TALES poses an artificial limit through an experiment where the best performing models are allowed to run for up to 1000 steps.

5.1 REASONING FAILURE MODES IN ALFWORLD, SCIENCEWORLD, AND ZORK1

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

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

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

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

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

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

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

5.3 REASONING FAILURES

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

As shown in Table 3, **top LLMs still make reasoning errors but resolve them quickly enough to still succeed.** When explicitly labeling and enumerating reasoning errors made on a per-action basis, 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

Table 3: Average reasoning failures for the Claude family of models across 8 games (32 logs total).

Model	Spatial	Deductive	Inductive	Grounded	Total	TALES Score (†)
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

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.

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

5.4 STANDARDIZATION

(Liu et al., 2024) define an adaption module as "how test conditions are constructed" and note a need for fairness in evaluation condition across all objects of evaluation. In TALES, we motivate a minimal adaption module to evaluate the reasoning capabilities of LLM agents in complex, long-horizon, situated environments with text-adventure games. A number of other works explore individual frameworks within TALES, but all of these works, to our knowledge, introduce significant amounts of domain knowledge through their implementation of the adaption module(Chang et al., 2024; Paglieri et al., 2024; Lu et al., 2025; Feng et al., 2025).

] r

Table 4: Scores in TW for Qwen3-8B when different forms of domain knowledge are included in the prompt. Experiment settings are same as main results outside of changes to system prompt. Admissible Actions are the actions the agent can take at any step, directly queried from the environment. Action Templates are those otherwise provided by the **help** command. Environment Dynamics refer to information such as "The BBQ is for grilling things, the stove is for frying things, the oven is for roasting things" that the agent would otherwise need to discover on its own.

ļ	0	9
ļ	1	0
ļ	1	1
	4	0

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

Table 4 illustrates the most common modifications to the adaption module. For fairness in evaluation, we directly use the prompts from (Paglieri et al., 2024) and (Lu et al., 2025) and thus only show results from TW. All values are from the Qwen3-8B⁴ model with 1024 tokens of thinking budget and 100 total steps. The inclusion of this information, that the agent would otherwise need to discover on its own, can lead to drastically inflated score. Excess information can even harm performance, as we see with the inclusion of both Action Templates and Environment Dynamics.

6 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 the observation and other information is fed to the LLM as an input with the output taken as an action.

⁴The full TALES score for this model is not complete at submission time. They will be added when allowed.

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

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

7 CONCLUSION

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

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

We introduce baseline LLMs in our framework as canonical benchmarks for text-adventure environments without additional prompting or training. Future work can improve upon these through supervised fine-tuning, in-context learning, or reinforcement learning. Thinking LLMs effectively leverage thinking traces; future research could guide these traces toward specific reasoning types for improved contextual awareness.

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

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

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

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

8 REPRODUCIBILITY STATEMENT.

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

REFERENCES

- Prithviraj Ammanabrolu and Matthew J. Hausknecht. Graph constrained reinforcement learning for natural language action spaces. *ArXiv*, abs/2001.08837, 2020. URL https://api.semanticscholar.org/CorpusID:210911499.
- 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.
- 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.
- Prithviraj Ammanabrolu, Ethan Tien, Matthew J. Hausknecht, and Mark O. Riedl. How to avoid being eaten by a grue: Structured exploration strategies for textual worlds. *ArXiv*, abs/2006.07409, 2020. URL https://api.semanticscholar.org/CorpusID:219687812.
- Shreyas Basavatia, Keerthiram Murugesan, and Shivam Ratnakar. Starling: Self-supervised training of text-based reinforcement learning agent with large language models. *ArXiv*, abs/2406.05872, 2024. URL https://api.semanticscholar.org/CorpusID:270372057.
- Ruth MJ Byrne and Philip N Johnson-Laird. Spatial reasoning. *Journal of memory and language*, 28 (5):564–575, 1989.
- Ma Chang, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. Agentboard: An analytical evaluation board of multi-turn llm agents. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), Advances in Neural Information Processing Systems, volume 37, pp. 74325–74362. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/877b40688e330a0e2a3fc24084208dfa-Paper-Datasets_and_Benchmarks_Track.pdf.
- Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. Textworld: A learning environment for text-based games. *CoRR*, abs/1806.11532, 2018.
- Christopher Cui, Xiangyu Peng, and Mark Riedl. Thespian: Multi-character text role-playing game agents. 2023.
- Christopher Z Cui, Xiangyu Peng, and Mark O Riedl. A mixture-of-experts approach to few-shot task transfer in open-ended text worlds. *arXiv preprint arXiv:2405.06059*, 2024.
- Mica R Endsley, Daniel J Garland, et al. Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement*, 1(1):3–21, 2000.
- Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*, 2025.
- Hitesh Golchha, Sahil Yerawar, Dhruvesh Patel, Soham Dan, and Keerthiram Murugesan. Language guided exploration for rl agents in text environments, 2024. URL https://arxiv.org/abs/2403.03141.

- Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines, 2014. URL https://arxiv.org/abs/1410.5401.
- Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. Interactive fiction games: A colossal adventure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020.
 - Matthew J. Hausknecht, Ricky Loynd, Greg Yang, Adith Swaminathan, and J. Williams. Nail: A general interactive fiction agent. *ArXiv*, abs/1902.04259, 2019. URL https://api.semanticscholar.org/CorpusID:60441391.
 - Evan Heit. Properties of inductive reasoning. Psychonomic bulletin & review, 7:569–592, 2000.
 - Peter A. Jansen and Marc-Alexandre Côté. Textworldexpress: Simulating text games at one million steps per second. *arXiv*, 2022. URL https://arxiv.org/abs/2208.01174.
 - Philip N Johnson-Laird. Deductive reasoning. Annual review of psychology, 50(1):109–135, 1999.
 - Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.
 - Bill Yuchen Lin, Yicheng Fu, Karina Yang, Prithviraj Ammanabrolu, Faeze Brahman, Shiyu Huang, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *ArXiv*, abs/2305.17390, 2023. URL https://api.semanticscholar.org/CorpusID:258960143.
 - Yu Liu, Su Blodgett, Jackie Chi Kit Cheung, Vera Liao, Alexandra Olteanu, and Ziang Xiao. Ecbd: Evidence-centered benchmark design for nlp. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16349–16365, 2024.
 - Cong Lu, Shengran Hu, and Jeff Clune. Intelligent go-explore: Standing on the shoulders of giant foundation models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=apErWGzCAA.
 - Ida Momennejad, Hosein Hasanbeig, Felipe Vieira, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. Evaluating cognitive maps and planning in large language models with cogeval, 2023. URL https://arxiv.org/abs/2309.15129.
 - Keerthiram Murugesan, Mattia Atzeni, Pavan Kapanipathi, Kartik Talamadupula, Mrinmaya Sachan, and Murray Campbell. Efficient text-based reinforcement learning by jointly leveraging state and commonsense graph representations. In *Annual Meeting of the Association for Computational Linguistics*, 2021. URL https://api.semanticscholar.org/CorpusID:236459800.
 - Karthik Narasimhan, Tejas Kulkarni, and Regina Barzilay. Language understanding for text-based games using deep reinforcement learning. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.
 - Allen Newell. Reasoning, problem solving and decision processes: The problem space as a fundamental category. 1979.
 - Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. Balrog: Benchmarking agentic Ilm and vlm reasoning on games, 2024. URL https://arxiv.org/abs/2411.13543.
 - Xiangyu Peng, Christopher Cui, Wei Zhou, Renee Jia, and Mark Riedl. Story shaping: Teaching agents human-like behavior with stories. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 19, pp. 326–336, 2023.
 - Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. Adapt: As-needed decomposition and planning with language models, 2024. URL https://arxiv.org/abs/2311.05772.

- Shuofei Qiao, Runnan Fang, Ningyu Zhang, Yuqi Zhu, Xiang Chen, Shumin Deng, Yong Jiang,
 Pengjun Xie, Fei Huang, and Huajun Chen. Agent planning with world knowledge model.
 Advances in Neural Information Processing Systems, 37:114843–114871, 2025.
 - Dongwon Kelvin Ryu, Meng Fang, Shirui Pan, Gholamreza Haffari, and Ehsan Shareghi. A minimal approach for natural language action space in text-based games. *ArXiv*, abs/2305.04082, 2023. URL https://api.semanticscholar.org/CorpusID:258557517.
 - Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *ArXiv*, abs/2303.11366, 2023. URL https://api.semanticscholar.org/CorpusID:257636839.
 - Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.03768.
 - Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7584–7600, 2024.
 - H. Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Raj Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. *ArXiv*, abs/2407.18901, 2024. URL https://api.semanticscholar.org/CorpusID:271516633.
 - Chen Feng Tsai, Xiaochen Zhou, Sierra S. Liu, Jing Li, Mo Yu, and Hongyuan Mei. Can large language models play text games well? current state-of-the-art and open questions. *ArXiv*, abs/2304.02868, 2023. URL https://api.semanticscholar.org/CorpusID: 257985065.
 - Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Soft self-consistency improves language model agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2024.
 - Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader?, 2022. URL https://arxiv.org/abs/2203.07540.
 - Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, et al. Livebench: A challenging, contamination-free llm benchmark. In *The Thirteenth International Conference on Learning Representations*.
 - Zonghan Yang, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. React meets actre: When language agents enjoy training data autonomy. *ArXiv*, abs/2403.14589, 2024. URL https://api.semanticscholar.org/CorpusID:268856510.
 - Shunyu Yao, Rohan Rao, Matthew J. Hausknecht, and Karthik Narasimhan. Keep calm and explore: Language models for action generation in text-based games. *ArXiv*, abs/2010.02903, 2020. URL https://api.semanticscholar.org/CorpusID:222142129.
 - Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.
 - Xingdi Yuan, Marc-Alexandre Côté, Alessandro Sordoni, Romain Laroche, Rémi Tachet des Combes, 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.
 - Haiteng Zhao, Chang Ma, Guoyin Wang, Jing Su, Lingpeng Kong, Jingjing Xu, Zhi-Hong Deng, and Hongxia Yang. Empowering large language model agents through action learning. *CoRR*, 2024.

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

Yuqi Zhu, Shuofei Qiao, Yixin Ou, Shumin Deng, Ningyu Zhang, Shiwei Lyu, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Knowagent: Knowledge-augmented planning for llm-based agents. *CoRR*, 2024.

A ALL AGENT AVERAGE SCORES PER FRAMEWORK

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

Model	TEXTWORLD	TEXTWORLDEXPRESS	ALFWorld	SCIENCEWORLD	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
01	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
Owen2.5-7B-Instruct	27.7	45.6	0.0	12.6	0.7	11.7
Llama-3.2-3B-Instruct	21.4	42.0	0.0	10.0	1.5	10.4
phi-4	20.8	43.8	0.0	8.9	1.6	10.3
gpt-4.1-nano	12.8	38.7	0.0	9.4	3.6	10.0
Mistral-Small-24B-Instruct-2501	15.8	23.0	0.0	15.8	1.4	8.8
DeepSeek-R1-Distill-Llama-70B	8.7	39.8	0.0	7.7	1.3	8.4
Ministral-8B-Instruct-2410	10.9	22.8	0.0	2.3	0.4	4.6
Mistral-Small-3.1-24B-Instruct-2503	2.5	10.3	0.0	10.5	0.8	4.5
Mixtral-8x22B-Instruct-v0.1	17.1	8.4	0.0	4.0	0.4	3.7
Llama-3.2-1B-Instruct	0.0	19.0	0.0	2.4	0.6	3.3
Phi-3-mini-128k-instruct	2.7	9.4	0.0	2.4	0.3	2.2
Phi-3.5-MoE-instruct	0.0	7.0	0.0	2.3	0.3	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-3medium-128k-instruct	0.0	0.0	0.0	2.3	0.3	0.7

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

B JERICHO WALKTHROUGH SCORES

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

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

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

C ALL AGENT AVERAGE FINAL TOKENS USED PER FRAMEWORK

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

Model	TEXTWORLD	TEXTWORLDEXPRESS	ALFWorld	SCIENCEWORLD	JERICHO
o3 (medium)	41342.0	107026.3	128740.7	92422.7	378168.5
o3 (high)	32148.6	51361.7	68275.0	51323.4	251760.8
gpt-5 (thinking)	38774.8	377563.2	145110.7	197999.9	649059.0
o3 (low)	51609.6	84218.5	119657.4	88557.9	318262.7
claude-3.7-sonnet (thinking)	69138.9	63072.8	74516.3	128752.7	311684.2
claude-3.7-sonnet	72072.9	46948.4	65131.1	152130.6	298533.1
claude-3.5-sonnet-latest	60766.1	68812.7	78765.4	106749.2	291950.3
gpt-4.1	53378.9	46454.8	54107.6	86547.5	218123.2
gpt-5-mini (thinking)	151984.2	391444.6	508585.4	318636.6	878673.2
01	47765.6	113492.9	127460.2	74300.1	211958.7
gpt-4o	106863.0	49536.2	77946.0	107121.9	209712.8
claude-3.5-haiku	119839.9	85136.1	267643.6	205751.0	269458.9
Llama-3.1-405B-Instruct	66476.2	52624.6	106290.0	137657.2	226078.6
gemini-2.0-flash	142937.1	66075.5	138048.3	142883.3	230182.0
Qwen3-32B	198390.4	188065.3	190900.1	229708.1	374514.2
Llama-3.3-70B-Instruct	166373.8	70165.0	127348.7	128860.9	205362.8
Llama-3.1-70B-Instruct	133253.4	51885.9	106925.7	144615.6	210914.8
Qwen2.5-72B-Instruct	112658.0	52096.1	97211.2	168057.3	197628.3
Mistral-Large-Instruct-2407	107788.5	110228.6	118395.4	163232.0	243256.0
gpt-4.1-mini	184516.3	92775.4	130758.2	125310.5	188824.3
gpt-4o-mini	159840.9	60210.7	145236.3	172875.7	182620.7
Llama-4-Scout-17B-16E-Instruct	289709.8	120173.9	172633.9	222464.1	229947.2
gpt-5-nano	770352.7	623055.1	821024.1	730904.7	825754.3
Llama-4-Maverick-17B-128E-Instruct-FP8	287547.2	213139.5	354183.9	394875.5	372902.6
Mistral-Small-Instruct-2409	163334.9	304510.9	107549.3	150730.7	208261.8
Llama-3.1-8B-Instruct	222239.7	358837.4	96582.5	152293.0	165505.8
DeepSeek-R1	393654.5	398322.7	496328.4	431997.9	439399.3
Qwen2.5-7B-Instruct	143127.1	214926.3	91334.4	163021.2	171107.7
Llama-3.2-3B-Instruct	230950.5	79878.3	84620.4	195397.2	152544.7
phi-4	189031.6	100363.9	126068.1	153395.2	178713.4
gpt-4.1-nano	545577.7	171767.5	277643.8	201505.9	182678.8
Mistral-Small-24B-Instruct-2501	399093.8	500484.8	479125.0	418284.9	475649.7
DeepSeek-R1-Distill-Llama-70B	453695.7	637384.1	719404.2	482819.3	407401.8
Ministral-8B-Instruct-2410	220157.9	337447.5	112710.5	108916.4	118104.9
Mistral-Small-3.1-24B-Instruct-2503	448764.0	507986.4	477505.8	397054.6	514733.5
Mixtral-8x22B-Instruct-v0.1	158782.2	137583.5	92832.7	134827.6	156515.8
Llama-3.2-1B-Instruct	567691.8	279214.8	457857.3	138285.6	201648.5
Phi-3-mini-128k-instruct	245215.0	429993.4	257852.2	253989.5	237881.5
Phi-3.5-MoE-instruct	274848.9	295190.9	240007.5	252055.6	271680.0
Phi-4-mini-instruct	231947.3	199299.1	195407.4	190887.4	212508.9
Mixtral-8x7B-Instruct-v0.1	612791.9	555281.3	520434.6	560994.6	564967.6
Phi-3.5-mini-instruct	426125.5	476218.4	410459.6	327584.9	457434.4
Phi-3-medium-128k-instruct	620235.4	585925.5	581721.6	513787.5	595335.5

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

D AGENT SCORE STANDARD DEVIATIONS

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

Model	TEXTWORLD	TEXTWORLDEXPRESS	ALFWorld	SCIENCEWORLD	JERICHO
o3 (medium)	0.0	2.7	4.6	2.2	0.5
o3 (high)	0.0	3.2	9.1	1.1	1.3
gpt-5 (thinking)	0.0	5.5	7.0	2.5	1.1
o3 (low)	2.0	6.9	9.5	0.8	1.1
claude-3.7-sonnet (thinking)	2.8	4.7	10.2	2.9	0.9
claude-3.7-sonnet	0.0	1.4	3.7	3.7	1.1
claude-3.5-sonnet-latest	0.0	2.9	5.9	3.4	1.0
gpt-4.1	2.6	1.9	11.8	2.3	0.8
gpt-5-mini (thinking)	3.6	2.8	9.5	6.1	1.4
01	1.2	4.4	4.6	5.0	1.7
gpt-4o	6.1	0.4	14.9	2.8	0.6
claude-3.5-haiku	5.3	0.0	3.7	2.6	0.6
Llama-3.1-405B-Instruct	5.0	4.9	10.9	4.5	0.5
gemini-2.0-flash	8.6	1.3	4.6	3.4	0.4
Qwen3-32B	6.8	1.9	10.9	3.2	0.4
Llama-3.3-70B-Instruct	2.8	3.4	3.7	2.3	0.1
Llama-3.1-70B-Instruct	3.5	1.9	5.9	4.5	0.2
Qwen2.5-72B-Instruct	2.0	2.5	4.6	3.8	0.7
Mistral-Large-Instruct-2407	8.2	2.6	3.7	8.1	0.9
gpt-4.1-mini	6.1	1.7	7.5	3.6	0.3
gpt-4o-mini	5.4	1.7	0.0	1.5	0.2
Llama-4-Scout-17B-16E-Instruct	0.0	0.0	0.0	0.0	0.0
gpt-5-nano	7.7	5.1	3.7	4.1	0.3
Llama-4-Maverick-17B-128E-Instruct-FP8	1.3	0.0	0.0	0.1	0.3
Mistral-Small-Instruct-2409	5.1	0.0	0.0	2.2	0.0
Llama-3.1-8B-Instruct	4.7	2.9	0.0	0.9	0.1
DeepSeek-R1	3.9	0.0	0.0	2.2	0.1
Qwen2.5-7B-Instruct	0.0	0.0	0.0	0.7	0.1
Llama-3.2-3B-Instruct	2.6	2.9	0.0	1.6	0.3
phi-4	0.4	0.0	0.0	1.3	0.0
gpt-4.1-nano	2.1	4.5	0.0	1.0	2.4
Mistral-Small-24B-Instruct-2501	3.1	1.0	0.0	1.1	0.3
DeepSeek-R1-Distill-Llama-70B	2.8	0.3	0.0	0.4	0.1
Ministral-8B-Instruct-2410	4.2	0.0	0.0	0.0	0.0
Mistral-Small-3.1-24B-Instruct-2503	0.0	0.0	0.0	0.3	0.0
Mixtral-8x22B-Instruct-v0.1	3.0	2.3	0.0	1.7	0.1
Llama-3.2-1B-Instruct	0.0	0.0	0.0	0.0	0.0
Phi-3-mini-128k-instruct	2.0	0.0	0.0	0.3	0.0
Phi-3.5-MoE-instruct	0.0	2.7	0.0	0.0	0.1
Phi-4-mini-instruct	0.0	0.0	0.0	0.0	0.0
Mixtral-8x7B-Instruct-v0.1	0.0	0.0	0.0	0.0	0.0
Phi-3.5-mini-instruct	0.0	1.1	0.0	0.1	0.0
Phi-3-medium-128k-instruct	0.0	0.0	0.0	0.0	0.0

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

E ALL GAMES

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

921 922

918

919 920

Table 9: Games Organized by Framework. Part 1.

923 924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

Jericho 1.905 19. Huntdark 37. Reverb 2. Acorncourt 20. Infidel 38. Seastalker 3. Advent 21. Inhumane 39. Sherlock 4. Adventureland 22. Jewel 40. Snacktime 5. Afflicted 23. Karn 41. Sorcerer 6. Anchor 24. Library 42. Spellbrkr 25. Loose 7. Awaken 43. Spirit 26. Lostpig 44. Temple 8. Balances 9. Ballyhoo 27. Ludicorp 45. Theatre 10. Curses 28. Lurking 46. Trinity 29. Moonlit 11. Cutthroat 47. Tryst205 12. Deephome 30. Murdac 48. Weapon 13. Detective 31. Night 49. Wishbringer 14. Dragon 32. Omniquest 50. Yomomma 15. Enchanter 33. Partyfoul 51. Zenon 52. Zork1 16. Enter 34. Pentari 17. Gold 35. Planetfall 53. Zork2 18. Hhgg 36. Plundered 54. Zork3 55. Ztuu

940941942

ScienceWorld

943 944

946

947

948

949

951

952

953

954

955

I. Boll

2. ChangeTheStateOfMatterOf 3. ChemistryMix

- ${\bf 4.} \; {\tt Chemistry MixPaintSecondary Color} \\$
- 5. ChemistryMixPaintTertiaryColor

6. FindAnimal

- 7. FindLivingThing
- 8. FindNonLivingThing
- 9. FindPlant
- 10. Freeze
- 11. GrowFruit
- GrowPlant
 - 13. IdentifyLifeStages1
 - 14. IdentifyLifeStages2
 - 15. InclinedPlaneDetermineAngle

- 16. InclinedPlaneFrictionNamedSurfaces
- 17. InclinedPlaneFrictionUnnamedSurfaces
- 18. LifespanLongestLived
- $19. \ {\tt LifespanLongestLivedThenShortestLived}$
- 20. LifespanShortestLived
- 21. MeasureMeltingPointKnownSubstance
- 22. MeasureMeltingPointUnknownSubstance
- 23. Melt
- 24. MendelianGeneticsKnownPlant
- 25. MendelianGeneticsUnknownPlant
- 26. PowerComponent
- 27. PowerComponentRenewableVsNonrenewableEnergy
- 28. TestConductivity
- 29. TestConductivityOfUnknownSubstances
- 30. UseThermometer

956957958

959 960

961962963

965966967

964

972 Table 10: Games Organized by Framework. Part 2. 973 ALFWorld 974 975 1. LookAtObjInLightSeen 7. PickCoolThenPlaceInRecepSeen 976 2. LookAtObjInLightUnseen 8. PickCoolThenPlaceInRecepUnseen 977 3. PickAndPlaceSimpleSeen 9. PickHeatThenPlaceInRecepSeen 978 4. PickAndPlaceSimpleUnseen 10. PickHeatThenPlaceInRecepUnseen 5. PickCleanThenPlaceInRecepSeen 11. PickTwoObjAndPlaceSeen 979 6. PickCleanThenPlaceInRecepUnseen 12. PickTwoObjAndPlaceUnseen 980 981 **TextWorld** 982 983 1. CookingLevel1 6. CookingLevel6 CookingLevel2 7. CookingLevel7 984 CookingLevel3 8. CookingLevel8 985 4. CookingLevel4 9. CookingLevel9 986 5. CookingLevel5 10. CookingLevel10 987 **TWX** 989 1. Arithmetic 9. SimonSaysWithMemory10 990 CoinCollector 10. SimonSaysWithMemory50 991 3. CookingWorld 11. SimonSaysWithMemory100 992 4. MapReader 12. SimonSaysWithMemory10Verbose 993 5. PeckingOrder 13. SimonSaysWithMemory50Verbose 6. SimonSays10 14. SimonSaysWithMemory100Verbose 994 7. SimonSays50 15. Sorting 995 8. SimonSays100 16. TextWorldCommonsense 996 997 998 ALL SCORES PER GAME: TEXTWORLD 999 1000 Table 11 shows the per-game scores of all models in TEXTWORLD across all seeds. 1001 1002 ALL SCORES PER GAME: TEXTWORLDEXPRESS G 1003 Table 12 shows the average per-game scores of all models in TEXTWORLDEXPRESS across all seeds. 1005 ALL SCORES PER GAME: ALFWORLD 1007 1008 1009 Table 13 shows the average per-game scores of all models in ALFWORLD across all seeds. 1010 1011 ALL SCORES PER GAME: SCIENCEWORLD 1012 1013 Tables 14 and 15 shows the per-task average scores of all models in SCIENCEWORLD across all 1014 seeds. 1015 1016 ALL SCORES PER GAME: JERICHO 1017 1018 Tables 16 and 17 shows the per-game scores of all models in JERICHO. * Indicates LLM has only 1019 been run on one seed. We will update the paper once all run seeds have been completed. 1020

1069 1070 1071 1072 1073 1074 1075 1076 1077 1078	1065 1066 1067 1068	1063 1064	1044 1045 1046 1047 1048 1049 1050 1051 1052 1053 1054 1055 1056 1057 1058 1059 1060 1061 1061	1052 1053 1054 1055 1056	1048 OM LX H L L L L L L L L L L L L L L L L L	1044 × 50 1045 × 60 1046 CT 1047 × 12	1040 1041 1042 1043	1035 1036 1037 1038 1039	1031 1032 1033 1034	1026 1027 1028 1029 1030
Models	$C_{ookingLevelI}$	CookingL	CookingLevel3	CookingLevel4	CookingLevels	CookingLevel6	CookingLevel7	CookingLevel8	Cookinglevel9	Cooking Level 10
(mı	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	0.001	100.0	100.0	100.0	0.001	0.001	100.0	100.0	0.001
claude-3.7-sonnet (thinking)	100.0	7.27	100.0	100.0	100.0	100.0	0.001	100.0	100.0	100.0
claude-3.7-sonnet	100.0	72.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
claude-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	90.9
gpt-5-mini (thinking)	100.0	47.3	0.001	100.0	100.0	100.0	0.001	100.0	100.0	0.001
01 mrt 10	100.0	7.87	0.007	100.0	0.001	0.001	0.001	100.0	100.0	78.7
gpt-+0 claude-3.5-haiku	100.0	58.2	100.0	100:0	100.0	100.0	100.0	100.0	100.0	90.9
Llama-3.1-405B-Instruct	100.0	54.5	85.0	100.0	100.0	100.0	100.0	100.0	100.0	69.1
gemini-2.0-flash	100.0	21.8	25.0	100.0	100.0	86.7	100.0	100.0	84.0	6.06
Qwen3-32B	100.0	27.3	85.0	100.0	100.0	46.7	86.7	100.0	0.89	81.8
Llama-3.3-70B-Instruct	100.0	0.0	25.0	100.0	100.0	0.0	100.0	88.0	92.0	90.9
Llama-3.1-70B-Instruct	100.0	0.0	25.0	100.0	100.0	0:0	0.001	100.0	84.0 50.0	47.3
Qwen2.3-72B-Instruct Mistral I aroe-Instruct-2407	100.0	32.7	1000	100.0	0.001 86.7	100.0	0.001	100.0	100.0	18.0
gpt-4.1-mini	0.09	5.5	70.0	100.0	46.7	86.7	100.0	40.0	76.0	36.4
gpt-40-mini	100.0	0.0	25.0	100.0	46.7	100.0	0.0	52.0	100.0	41.8
Llama-4-Scout-17B-16E-Instruct	33.3	27.3	25.0	100.0	100.0	0.0	0.0	40.0	40.0	45.5
gpt-5-nano	60.0	12.7	85.0	100.0	33.3	0.0	0.09	84.0 33.0	20.0	45.5
Liama-4-iMavefick-176-128E-insuret-FP8 Mistral-Small-Instruct-2409	100.0	0.0	25.0	100.0	100.0	0.0 20.0	0.00	32.0 100.0	16.0 76.0	41.8
Llama-3.1-8B-Instruct	73.3	1.8	25.0	0.0	100.0	0.0	0.0	28.0	0.09	9.1
DeepSeek-R1	2.99	0.0	25.0	75.0	100.0	0.0	0.0	28.0	40.0	36.4
Qwen2.5-7B-Instruct	33.3	18.2	25.0	100.0	0.0	0:0	0.0	100.0	0.0	0.0
Llama-3.2-3B-Instruct	0.04	0.0	0.00	0.00	0.0 33 3	33.3	0:0	24.0 40.0	16.0	1.6
gpt-4.1-nano	33.3	0:0	25.0	0.0	0.0	0.0	13.3	56.0	0.0	0:0
Mistral-Small-24B-Instruct-2501	46.7	0.0	25.0	25.0	33.3	0.0	0.0	20.0	8.0	0.0
DeepSeek-R1-Distill-Llama-70B	6.7	0.0	25.0	55.0	0.0	0.0	0.0	0.0	0.0	0.0
Ministral-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
IMIXITAL-8XZZB-INSTRUCT-VU.1	/ 0	0.0	0.0	0.00	0:0	0.0	0.0	0.450	0.0	0.0
Liania-3.2-1D-msuuct Phi-3-mini-128k-instruct	0.0	0.0	0.01	5.0	0.0	0.0	0.0	12.0	0.0	0.0
Phi-3.5-MoE-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Phi-4-mini-instruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mixtral-8x7B-Instruct-v0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0:0
Fin-5.5-min-instruct 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

1080	egge [©]																						
1081	Maria Capanagua de la companya de la	0.00.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.00	200	9 9	0.0	0.0	20	9 9	22	0.0	222
1082	a shorid	== ~ *	000	SC	·vo·v	n vo vo	n on o	none	n v	en en	es es	so e	n or o	n en e	u vo		4 ~	S.			-		
1083				_						_													
1084 sdi	ing	0.000	0.00	0.0	20.0	0.0	40.0	12.0	0.0	0.89	0.0	0.0	0.0	36.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	000
1085	no ^{se}																						
1086	ar for Meter																						
1087	(altabeted).	93.2	0.001	100.0	59.6	1000	1000	70.4	0.00	100.0	100.0	70.0	0.0	200	0.0	0.00	80.4	0.4	0.0	0.0	0.0	0.0	000
1088	and the state of t																						
1089 set	j.																						
1091	with their	94.0 92.0 63.4	0.00	0.00	27.7	0.0	0.00	73.6	90.0	00.0	0.00	0.0	0.9	200	0.0	0.0	3.2	0.4	2.0	0.0	0.0	0.0	0.0
1092	and the state of t	0,0,0						_ , , , ,	~ -			(- (401								
1093	` .																						
1094	GMeltose																						
1095	Mentoryle	0.0000	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	200	200	0.0	0.0	0.0	0.0	9.0	9 9	9.9	0.0	9.9
1096	as william.	2222	222	22	222	222	222	223	22	22	22	0.7	20%	529	22	22	2 ⊱	2 3	¥ % '	00	0 =	00	
1097	State of the state																						
1098	_e go.																						
1099 SS	and the state of t																						
1100	Menory	51.6 61.4 42.2 87.0	0.00	0.00	24.2	0000	000	8.8	0.00	000	9.6	000	5.0	203	3.0	000	0.1	0.0	2.0	9 9	0.0	0.0	0.0
1101	cal switter	,,,,,,	~			0	`	- ~ -			0,0,	_											
1102	gotto																						
1103	ilao																						
1104	para ship de la	4464	0	0,0	. 61 4	. 9.0	0.0	5.40	o o	9.0	2 0	0	. 0 -		. 0	. 0					~ -		
1105 X	Salahia.	53.4 4.6 4.6	8.8	88	4 ×	§ § §	88	9 8 8	38	≘ =	æ æ	100	13.	121	15.	7.00	==	===	12	ĕ ĕ	ŏö	ö ö	888
1106 H	gatu																						
1107	e de la companya de l																						
1108	ristrateric.	0000	0.0	0.0	000	0.0	0.0	000	0.0	000	0.0	0.0	000	200	0.0	0.0	0.0	5.0	0.0	99	9.8	9.9	99
1109	of Sales	2222	222	22	2 2	222	222	223	22	≥ 5	22	2 6	2 -	- 2 -	- 2	22	- 4	6 -		-	O <u>-</u>		,00
1110	,																						
1111	Sagara Sagara	0000	0.0	0.00	0.0	0.0	200	000	0.0	0.00	0.00	2.0	0.00	0.00	0.0	9 9	0.0	0.8	201	0.0	9 2	2.0	229
1112	,	2225	222	22	: ∞ ⊆	222	3 44.7	223	22	22	22	4.0	2 4	, ⊆	+ ≃		4-	0 2	3 4	÷ ≃	4 (1)	4.4	· e, O
1113	ar lab	0070		. 9	209		200		20	9,9	0,0	00		9.0	. 9.		0 -		2	. 9		0 -	
1114 × 3	ggr ^z si ^s idh	100.0 100.0 100.0	888	7.00	86 5	888	2.23	888	88	22	9 9	2, 5	8 8	1 2 7	10	4.0	27.	₹ 5	3 7	. 6	či či	2.2	00
1097 1098 1099 1100 1101 1102 1103 1104 1105 1106 1107 1108 1109 1111 1111 1112 1113 1114 1115 1116	.0																						
1116	gen Say!	9999	1000	100.0	0.00	1000	1000	100	9 0	0.00	1000	100.0	1000	1000	0.00	0.0	48.0	100.0	20.0	100.0	0.09	0.0	8.0
1117 [qa]																							
1118	girlig Oplast	0.000	0.00	0.001	9 9	0.00	0.001	0.00	0.00	0.00	30.0	0.001	0.00	000	25.0	0.0	100.0	50.0	0.0	0.0	0.0	25.0	0.0
1119	´																						
1120	Recite	0.0.0.0	0.00	0.00	0.00	0.00	000	0.00	0.00	0.00	0.08	0.00	0.00	888	0.0	0.0	0.0	0.0	0.0	0.0	88	0.0	888
1121							-		_		-,-,	_	_										
1122	king wind	76.8 53.6 42.0	0.0	2.8	9.6	2.5	9.0	0.80	9.8	2.0	0.0	0.0	0.0	2009	3 9 1	99	9.9	0.0	9 9	9 9	99	9.9	99
1123 cod	akiti.	K M 44 4	रिचेच	m m	ભે વ	7 76 1	ય લ્લુ	ri en e	71 K	⊱! 4 .	40	4.0	. 4. 5	201	j 0 .		0.2	22.0	-				
1124	xot	0000		00						00	00	00				. 0	00	0	. 0		۰.		
1125 care	gColi	9999	888	8 8	88	888	888	888	3 8	8 8	8 8	8 8	888	888	8 8	8 8	8 8	90 0	8 8	8 8	8 8	8 8	888
1126	aji.C	0 0 0 0					-			_	_			_									
1127 pdf	Indigit	0.08 0.09 0.09 0.09	9 8	20.0	8 9	8 8 8	80.6	00 8	30.0	20.0	200	000	8 8 8	900	88	88	0.0	000	88	88	88	0.0	88
1128													-FP8										
1129												5	nstruct				_	8	2503				
1130			(gu							107		-Instruc	128E-In	6			052-12	ma-701	15truct-	.0.1		7	uct
1131			(thinki	latest	ng)		nstruct		struct	truct ruct-24		B-16E	c-17B-;	truct	nct	truct	-Instru	till-Lla	24B-Ir	struct-v ruct	nstruct uct	t net-v0	uct k-instr
1132		ing)	claude-3.7-sonnet (thinking) claude-3.7-sonnet	claude-3.5-sonnet-latest grt-4.1	gpt-5-mini (thinking)	i si si	Llama-3.1-405B-Instruct	, leasn	Llama-5.5-70B-Instruct Llama-3.1-70B-Instruct	Qwen2.5-72B-Instruct Mistral-Large-Instruct-2407	·= -	out-17.	Llama 4-Maverick-17B-128E-Instruct-FP8	Llama-3.1-8B-Instruct	8-Instr	Llama-3.2-3B-Instruct phi-4	30 311-24B	R1-Dis	Mistral-Small-3,1-24B-Instruct-2503	Mixtral-8x22B-Instruct-v0. Llama-3.2-1B-Instruct	-128k-i E-instr.	instruc 'B-Instr	Phi-3.5-mini-instruct Phi-3-medium-128k-instruct
		o3 (medium) o3 (high) gpt-5 (thinking)	claude-3.7-sonnet	le-3.5-	-mini	gpt-4o	na-3.1-	Owen3-32B	na-5.5- na-3.1-î	n2.5-7. ral-Lan	f.1-min fo-mini	na-4-Sc	na-4-M.	na-3.1-8	Owen2.5-7B	na-3.2-: t	4.1-nan	Seek-I	ral-Sm	ral-8x2 va-3.2-1	3-mini- 3.5-Mo,	t-mini-	3.5-mir i-mediu
1133 Additional of the state of		o3 (medions) o3 (high) gpt-5 (thi	clauc	clauc gpt-4	g de c	gpt-40	Llam	Selli Selli		Owe Mistr	2 pg	Llam	Llam	Flan	% E	Phi-4	P-tdg Mistr	Deer	Mistr	Mixt	F F	Phi-4	Phi-3

1134	
1135	
1136 othina	88 88 88 89 8 8 8 8 8 8 8 8 8 8 8 8 8 8
1134 1135 1136 1137 Professor Profes	
1138 1139 1140 1141	
1139	100 000 000 000 000 000 000 000 000 000
1140 twomit	
1141	
1142	
1143 attRecent	
1144 Thentheeth	
Litteau	
1145	
1146	
1147	\$2.50 \$2.50
1148	
1146 1147 1148 1149 retteefteefteefteefteefteefteefteefteeft	
1150 1151 1152 1153 1154	>
1151 interest	0-0-000
1152	8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
1153 Z account	
1154	
1154 1155 1156 1157 1158	
1156 Since Inter	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
1157 📜 confinent	= 2 × × 4 = = = 6 × 4 + 0 = 0 × 4 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 = 0 =
1158	
1158 1159 1160 1161 1162 1162 1162	a
1160 0 angeografi	
1161 E Thenthuser	100 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1162 Ell graduation	
1162 July 45°	
1163 La	
1164	660 600 600 600 600 600 600 600 600 600
1165 OO Themstreet.	889889998080000000000000000000000000000
1163 1164 1165 1166 1167	
een	
1168 O	100 100 100 100 100 100 100 100 100 100
1169	
11/0	
1171 _{kseen}	
1171 1172 1173 1174	1000 1000 1000 1000 1000 1000 1000 100
1173	
1174	
1174 1175 1176 1177 Lask confinding to see the second seco	
1176 grajinte	
1177 Look A.	
1178 1179 1180 Last continues from 1	
1179	1000 1000
1180 Lankerous	
1181	8
1182	act-FP
1183	-Instr -OB
1184	king) t t t t t t t t t t t t t t t t t t t
1185	((thin for the property of th
	cing) some some some some control thank t
1186	o3 (medium) (chigh) (c
1187	o3 (medium) o3 (medium) gpt 5 (thinking) gpt 5 (thinking) claudes 27, sommer (thinking) claudes 27, sommer (thinking) claudes 27, sommer (thinking) claudes 27, sommer (thinking) gpt 44. gpt 43, sommer (thinking) gpt 46. gpt 46. gpt 45. gpt 46. gpt 46. gpt 47. gpt 46. gpt 47. gpt 46. gpt 47. gp
I	

1100	ſ	
1188 1189	Angle	
1190	IndirectPlanetDeeringerteethe	
1190	APlane De	10000 100000 100000 100000 100000 100000 100000 100000 100000 100000 100000 100000 100000 100000 100000 100000 100000 100000 100000 100000 1000000
1192	Inclined	
1192	-ಬ	
1194	, if EStage	1100.0 100.0 1000.0 1000.0
1194	Herity lite Shape 2	000093396458948889898888989809777808800000000000000
1196	,	
1197	Hernity Life Stages 1	004040001004004000000400000000000000000
1198	i antifyl ite	88.08.08.08.08.08.08.08.08.08.08.08.08.0
1199	Ide	
1200	GrowPlant	885.2 92.2 92.2 92.2 92.2 92.2 92.2 92.2 92.2 92.2 93.4 94.4 95.4 96.6
1201		2001.01.01.01.01.01.01.01.01.01.01.01.01.
1202	Growfruit	78.45.45.45.45.45.45.45.45.45.45.45.45.45.
1203		284474 647 C C C C C C C C C C C C C C C C C C C
1204	Freede	448.0 449.0 449.0 449.0 449.0 449.0 449.0 449.0 449.0 449.0 449.0 449.0 449.0 449.0 600.0 60
		000004000000000000000000000000000000000
1206 d	FindPlant	100.00 10
1205 1206 1207 1208 1209 1210 1211 1212 1213 1214 1215 1216 1217 1218 1219 1220 1221 1222 1223 1224 1224 1225	ing	
1208 gs	Find Hood Lines Think	10000 100000 100000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 100000 10000 10000 10000 10000 10000 10000 10000 10000 10000 100000 10
1209	FindNone	
1210	,	
1211	Find Living thing	1100.0 1100.0 1100.0 1100.0 10
1212	Find in	000008800880888888888888888888888888888
1213		
1214	Find Animal	88.00 88
1215	ζ,	
1216	ChemieryMixPaintCetiaryColor	
1217 E	DainTert	1000.0 10
1218	. stryMixt	5558865888469848559485885886566666666666
1219 Ja	Chemit	
1220	Cheroleter Mest Parint Secondary Color	
1221	andaryCo.	
1222	Saint Sector	100.0 100.0
1223	atryMixt	
1225 E	-Ni ^t	
1226 1227	Chemistryhtix	100.00 10
1228	Cite	
1229	Charge The State Official reds	
1230	StateOffite	10000 10000
1231	nge The 3.	3333744884846481-44911999999999999999999
1232	Chia	
1233	Boil	484 484 488 68 68 68 68 68 68 68 68 68 68 68 68 6
1234		
1235		nect-F
1236		
1237		sing) 2407 2407 3-Inst amar-128E 3-Inst amar-108 instruct 6 instruct 6 instruct 7 instruct 6 instruct 7 instruct 7 instruct 7 instruct 7 instruct 7 instruct 7 instruct 8 instruct 8 instruct 9 instruct 1 instr
1238		(thin) laberst ng) ng) ng) ng) nstruc struct struct-2 ruct-2 ruct-1 ruct
1239		ng) ng) ng) ng) ng) ng) ng) ng) ngiku ngik
1240		o3 (medium) o3 (nigh) o3 (fuigh) o3 (fuigh) o3 (fuigh) o3 (town) gra-5 (funking) e3 (town) e4 (town) e4 (town) e4 (town) e4 (town) e4 (town) e4 (town) e5 (t
1241	Models	o3 (modium) o3 (ingth) gar 5 (thinki) gar 5 (thinki) gar 6 (thinki) gar 6 (thinki) gar 6 (thinki) gar 7 (thinki) gar 7 (thinki) gar 8 (thinki) gar 8 (thinki) gar 9 (thinki
	2	

1242	000000000000000000000000000000000000000
1243	10000 100000 100000 100000 100000 10000 10000 10000 10000 10000 10000 10000 100000 1
	,
1245 1246	
1246	11000000000000000000000000000000000000
1247 Teccomment	
1248	
1249 Test Coule of Test.	100 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1250	
1250 1251 1252 1253 1254	an and a second
1252 yazuntuk	
1253 georgebet	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
1254 CARRAGERE	
1255	
1256 (marketer	\$0001044*0100100000\$~01.00.00000
- sweet	7100 7100
1257	
1258 1259 ¥ approximated	00000000000000000000000000000000000000
1259	\$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$
1260 Q _{Menter}	
1261	
1262 E	1000 1000 1000 1000 1000 1000 1000 100
1263	
1264	62 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
1265	5 = 1 = 1 = 2 = 2 = 2 = 2 = 2 = 2 = 2 = 2
1266 1267 1268 1269	
1267	100.0 100.0
1268 S agreetering	
1269 UO NO.	
1270 90 1271 1272 Longitude the state of the	
1271 E grant and a	1000 1000
1272 Lightenite	
1273 £1 Head	
1274	1000 10
1275 Q	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
1276	\$
1276 <u>S</u> 1277 <u>S</u> 1278 <u>e</u> gesturende de d	
1278 D attraction	10000000000000000000000000000000000000
1279 Ligginger	
1280	
1281	1000 (
1 ice pair	
1283 author	
1282 1283 1284 1285 1286	000000000000000000000000000000000000000
1285	1000 1000 1000 1000 1000 1000 1000 100
1286	
1007	
1287	000000000000000000000000000000000000000
1288	00000000000000000000000000000000000000
1286 1287 1288 1289 1290	
	₩ ₩
1291	- Instruct
1292	at a
1293	(thing) king) king) history
1294	(in)
1295	of (readtum) geo G (figh) geo G
^	

1296	ı	
1297	Ludicorp	112.0 112.3
1298	Lostpie	25.25.25.25.25.25.25.25.25.25.25.25.25.2
1299		 %
1300	1,00 ⁵⁸	2 5 7 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
1301	Library	5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
1302 1303		
1304	Karn	$\left \begin{array}{c} \\ 47.48 \\ 28.80 \\ 20.000 \\ 20.$
1305	Jewel	$\left[\begin{smallmatrix} & & & & & & & & & & & & & & & & & & &$
1306	Inhumane	\$3.8.9.8.8.8.9.8.8.9.9.9.9.9.9.9.9.9.9.9.
1307		
1308	Infidel	[221889888888888888888888888888888888888
1309 1310	Hundark	
1311		
1312	Hhgg	
1313	Gold	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1314 ed	Enter	7420 7420
1313 1314 1315 1316 1317 1318 1318 1319	Enchanter	7 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
1316 (tas		
1317)səm	Dragon	$\begin{smallmatrix} 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 $
	Detective	33.3.3.3.3.3.4.4.4.4.4.4.4.4.4.7.2.2.2.2.2.2.2.2.2.2
1320 OH		
1321 OI	Deephome	770 10.00 10
1322		
1323	Cuthroat	13.6 6 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1324 auce	Curses	
1326		
1327 Ja	Ballyhoo	
1320 OHDIAIS O	Balances	22,22,23,24,25,25,24,25,25,24,25,25,24,25,25,25,25,25,25,25,25,25,25,25,25,25,
1329		
1330	Awaken	
1331 9I 9	Anchor	7.1.7.7.7.7.7.7.7.7.7.7.7.7.7.7.7.7.7.7
1332 Taple 1		\$ # 4 2 9 8 7 8 - 1 - 1 0 8 0 8 0 8 0 9 0 0 0 0 0 0 0 0 0 0 0 0
1334	Afflicted	18.95 28.43 38.62
1335	Adventureland	1117 100 100 100 100 100 100 100
1336	Advent	
1337	Advent	2828
1338 1339		
1340	Acomeount	1000 1000 1000 1000 1000 1000 1000 100
1341	905	10001010101010101010101010101010101010
1342	,	
1343		nct-FF
1344		E-Insu E-Insu 708 106-25
1345		est ruct t. 1. 2.407 (6E-In. 18-12-8 Struct: -1. Lhma 2.410 t. t
1346 1347)) neet (the neet the neet the neet the neet the neet and neet as a second neet as a second neet the n
1348		ininking (17, 500 minking (17, 50 minking
1349	Models	o3 (neglum) o3 (neglum) o3 (high) o3 (high) o3 (high) o3 (huh) o4 (high) o5 (huh) o5 (huh) o5 (huh) o5 (huh) o5 (huh) o6
	×	\$

1350	Ι,	
1351	Zuu	0.0000000000000000000000000000000000000
1352	101K3	440.0 440.0
1353	2012	
1354		
1355	Tork!	15.8 1.15.3 1.15
1356	Lenon	
1357		
1358	Yomonima	000000000000000000000000000000000000000
1359		
1360	Wishbringer	13.2 13.2 13.2 13.3 13.4 13.8
1361		
1362	Weapon	
1363	205	40769080976000000000000000000000000000000
1364	Trystals	# 4 # W 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1365	Tinity	113.46 113.47 113.47 113.48 113.88 113.88 113.88 113.88 113.88 113.88 114.00 110.00
1366	100	
1367	Temple	160 1143 1143 1143 1143 1143 1143 1143 114
1368 ä	Spirit	11.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1
1369 (SX)	7	
1370 stas	Spellbrikt	100.0000000000000000000000000000000000
1371 See E	atet	088828888888888888888888888888888888888
1373 Eg	Sorceter	
1374 OH	Snacktime	\$50.00000000000000000000000000000000000
1375	Sna	
1376	Sherlock	11.4 11.6 12.6 13.6 13.6 13.6 13.6 13.6 13.6 13.6 13
1377 50	, net	0.0000000000000000000000000000000000000
1378	Seastalker	4 + 0
1379 Eu	Reverb	33.00 2.44.85.88.88.88.88.88.88.88.88.88.88.88.88.
1380	, A	w 4 w 4 4 4
1381 Each	Plundered	47, 47, 47, 47, 47, 47, 47, 47, 47, 47,
1382	(all	
1383	Planetfall	$\begin{array}{c} 7.777 \\ \times 0.000000000000000000000000000000000$
1384	Pentari	7-1-1 7-1 7
1367 1368 1369 1370 1371 1372 1373 1374 1375 1376 1377 1378 1377 1378 1380 1381 1382 1383 1384 1385 1386	Ac.	
1386 eq. [387]	Partyfoul	
1387 <u>E</u>		
1389	Omniquest	250 14.00 10.0
1390	Might	24,0 24,0 25,0 25,0 25,0 25,0 25,0 25,0 25,0 25
1391		
1392	Murdac	9.88
1393	Moonlii	
1394		
1395	Linking	8, 8, 8, 8, 8, 8, 7, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8,
1396	,	80
1397		nct-Fl
1398		E-Insti
1399		st s
1400		tr (thir at the cing) in the cing at the
1401		king) seomt control (thinh control
1402	els	o3 (medium) o5 (medium) o5 (medium) o5 (medium) o3 (migh) o3 (now) o3 (now) o3 (now) o4 (now) o5 (now) o4 (now) o5 (now)
1403	Models	o3 (medition) o3
	•	

K AVERAGE SIMON SAYS SCORE VERSUS OVERALL TALES SCORE

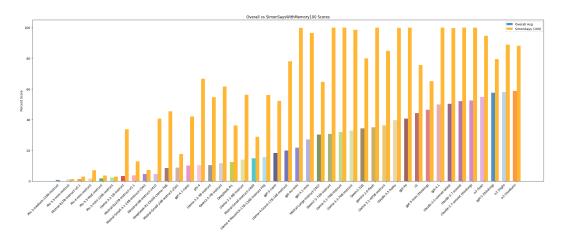


Figure 2: All TALES and average SIMON SAYS scores for each model, sorted by TALES performance. We see that an increase in performance in SIMON SAYS typically correlates with an increase in performance for TALES overall.

L COMPUTE

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

M 1000 STEPS OF ZORK1

While even the best performing LLMs make reasoning mistakes, we find what allows them to still find success in TALES is the ability to both avoid making an excessive number of these mistakes and the ability to self-correct. We argue that 100 steps are sufficient to evaluate the performance of current state-of-the-art LLMs because even the best overall LLM, o3, fails to approach the maximum possible score within 100 steps for Zork1 (13.8% vs 29.1%). In this experiment, we explore whether any of the top models can achieve a score comparable to the walkthrough after 100 steps in Zork1 while allowed to run for 1000 steps with the entire history kept in the context. If so, we examine the behaviors that enable this and determine the required number of steps. We select the overall top 3 performing models, o3 and both the thinking and non-thinking modes of Claude-3.7-Sonnet.

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

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

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

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

N FRAMEWORK ENVIRONMENT SUBSELECTION

N.1 TEXTWORLD

1458

1459

1460

1461 1462

1463 1464

1465 1466

1467

1488

1489

1490

For TEXTWORLD, we use the following environments:

```
1468
     test/difficulty_level_1/tw-cooking-recipe1+take1+open-0nQyHWbvh6d|
1469

→ XFPmhLKX.z8

1470
     test/difficulty_level_2/tw-cooking-recipe1+take1+cook+open-0nQyHW|
1471

→ bvh6dXFPmhLKX.z8

1472
     test/difficulty_level_3/tw-cooking-recipe1+take1+cut+open-0nQyHWb|
1473

→ vh6dXFPmhLKX.z8

     test/difficulty_level_4/tw-cooking-recipe1+take1+open+go6-0nQyHWb|
1474

→ vh6dXFPmhLKX.z8

1475
     test/difficulty_level_5/tw-cooking-recipe1+take1+open+go9-0nQyHWb|
1476
      1477
     test/difficulty_level_6/tw-cooking-recipe1+take1+open+go12-0nQyHW|
1478

→ bvh6dXFPmhLKX.z8

1479
     test/difficulty_level_7/tw-cooking-recipe1+take1+cook+cut+open-0n_
1480

→ QyHWbvh6dXFPmhLKX.z8

1481
     {\tt test/difficulty\_level\_8/tw-cooking-recipe3+take3+open+go6-0nQyHWb}_{\tt J}
1482
      1483
     test/difficulty_level_9/tw-cooking-recipe3+take3+cook+cut+open+go|
1484

→ 6-0nQyHWbvh6dXFPmhLKX.z8

     test/difficulty_level_10/tw-cooking-recipe3+take3+cook+cut+open+g_1
1485

→ o12-0nQyHWbvh6dXFPmhLKX.z8

1486
1487
```

N.2 TEXTWORLDEXPRESS

For TEXTWORLDEXPRESS, we use the game parameters:

```
1491
      TASKS = [
1492
           (
1493
               "CookingWorld",
1494
               "cookingworld",
1495
               "numLocations=1, numIngredients=2, numDistractorItems=5,
1496
                   includeDoors=0, limitInventorySize=0",
          ),
1497
1498
               "TextWorldCommonsense",
1499
               "twc",
1500
               "numLocations=1, numItemsToPutAway=1, includeDoors=0, limitI
1501
                → nventorySize=0",
1502
           ),
1503
1504
               "CoinCollector",
1505
               "coin",
1506
               "numLocations=1, numDistractorItems=5,
1507
                → limitInventorySize=0",
1508
           ),
           ("Arithmetic", "arithmetic", ""),
1510
               "MapReader",
1511
               "mapreader",
```

```
1512
                "numLocations=2, maxDistanceApart=1,
1513
                → maxDistractorItemsPerLocation=2, includeDoors=0,
                   limitInventorySize=0",
1515
           ),
1516
           ("Sorting", "sorting", ""),
           ("SimonSays10", "simonsays", "gameLength=10, numDistractors=4,
1517
           → memorization=0"),
1518
           ("SimonSays50", "simonsays", "gameLength=50, numDistractors=4,
1519

→ memorization=0"),
1520
           ("SimonSays100", "simonsays", "gameLength=100,
1521
               numDistractors=4, memorization=0"),
1522
           (
1523
               "SimonSaysWithMemory10",
1524
               "simonsays",
1525
               "gameLength=10, numDistractors=4, memorization=1,
1526

    verbose=0",

1527
           ),
1528
           (
               "SimonSaysWithMemory50",
1529
               "simonsays",
1530
               "gameLength=50, numDistractors=4, memorization=1,
1531

    verbose=0",

1532
           ),
1534
               "SimonSaysWithMemory100",
1535
               "simonsays",
1536
               "gameLength=100, numDistractors=4, memorization=1,
1537

    verbose=0",

1538
           ),
1539
               "SimonSaysWithMemory10Verbose",
1540
               "simonsays",
1541
               "gameLength=10, numDistractors=4, memorization=1,
1542
                \rightarrow verbose=1",
1543
           ),
1544
1545
               "SimonSaysWithMemory50Verbose",
1546
               "simonsays",
1547
               "gameLength=50, numDistractors=4, memorization=1,

    verbose=1",

1549
           ),
1550
               "SimonSaysWithMemory100Verbose",
1551
               "simonsays",
1552
               "gameLength=100, numDistractors=4, memorization=1,
1553
                \rightarrow verbose=1",
1554
1555
           ("PeckingOrder", "peckingorder", ""),
1556
      1
1557
1558
      N.3 ALFWORLD
1559
1560
      The 12 games for ALFWORLD. Note that these are from when the "-game-seed" is not set. Changing
1561
      this value would cause the games to change.
1562
1563
      valid_seen/pick_and_place_simple-Book-None-SideTable-329/trial_T2|
1564
       \rightarrow 0190908_050633_745514
      valid_seen/look_at_obj_in_light-AlarmClock-None-DeskLamp-323/tria;
1565

→ 1_T20190909_044715_250790
```

```
valid_seen/pick_clean_then_place_in_recep-ButterKnife-None-Counte_
1567

→ rTop-8/trial_T20190909_105559_983897

1568
     valid_seen/pick_heat_then_place_in_recep-Apple-None-DiningTable-2
1569
      → 6/trial_T20190907_060234_011675
1570
     valid_seen/pick_cool_then_place_in_recep-Apple-None-CounterTop-14|
      → /trial_T20190909_044933_815840
1571
     valid_seen/pick_two_obj_and_place-AlarmClock-None-Dresser-305/tri
1572
      → al_T20190907_165826_194855
     valid_unseen/pick_and_place_simple-Muq-None-Desk-308/trial_T20190|
1574

→ 908 125200 737896

1575
     valid_unseen/look_at_obj_in_light-AlarmClock-None-DeskLamp-308/tr
1576

→ ial_T20190908_222917_366542

1577
     valid_unseen/pick_clean_then_place_in_recep-Bowl-None-Cabinet-10/
1578
      → trial_T20190909_061130_844814
1579
     valid_unseen/pick_heat_then_place_in_recep-Apple-None-Fridge-10/t |
1580
      → rial_T20190906_182259_116320
1581
     valid_unseen/pick_cool_then_place_in_recep-Bread-None-CounterTop-
      → 10/trial_T20190908_091747_866951
1582
     valid_unseen/pick_two_obj_and_place-CD-None-Safe-308/trial_T20190|
1583
      → 907_050942_897916
1584
1585
```

N.4 SCIENCEWORLD

1586 1587

1588 1589

1591 1592

1593 1594

1595

1596

1597

1598

1599

1603 1604 We use the first variation of the test set for each of the 30 tasks.

O ANNOTATED LOGS: CRITERIA AND EXAMPLES

O.1 LABELING CRITERIA

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

O.2 SPATIAL REASONING FAILURE EXAMPLE

```
1605
1606
      Task Description: Your task is to melt lead. First, focus on the
      → substance. Then, take actions that will cause it to change its
1608
         state of matter...
1609
1610
      This room is called the kitchen. In it, you see:
1611
              the agent
1612
              a substance called air
1613
              a chair. On the chair is: nothing.
1614
              a counter. On the counter is: a bowl (containing a red
1615
              → apple, a banana, an orange, a potato), a drawer.
1616
              a cupboard. The cupboard door is closed.
1617
              a freezer. The freezer door is closed.
              a fridge. The fridge door is closed.
1618
              a glass jar (containing a substance called sodium
1619
                 chloride)
```

```
a lighter
1621
               a oven, which is turned off. The oven door is closed.
               a painting
1623
               a sink, which is turned off. In the sink is: nothing.
               a substance called soap
               a stopwatch, which is deactivated.
1625
               a stove, which is turned off. On the stove is: nothing.
1626
               a table. On the table is: a glass cup (containing
1627
                \rightarrow nothing).
1628
               a thermometer, currently reading a temperature of 10
1629

→ degrees celsius

1630
      You also see:
1631
               A door to the bathroom (that is open)
1632
               A door to the hallway (that is closed)
1633
               A door to the outside (that is closed)
1634
1635
      Action: open oven
1636
1637
1638
      Action: go hallway
1639
1640
1641
1642
      Action: go workshop
1643
1644
1645
      This room is called the workshop. In it, you see:
1646
               the agent
               a substance called air
1647
               a table. On the table is: a battery, a green wire, a metal
1648
                   pot (containing a substance called lead), a orange
1649
                   light bulb, which is off, a orange wire, a red wire, a
1650
                  switch, which is off, a violet light bulb, which is
1651
                → off, a yellow light bulb, which is off.
1652
               a ultra low temperature freezer. The ultra low temperature
1653

    freezer door is closed.

      You also see:
1655
               A door to the hallway (that is open)
1656
1657
      Action: pick up metal pot
1658
      You move the metal pot to the inventory.
1659
1660
      Action: go kitchen
1661
1662
1663
      Above we provide an example of a spatial reasoning failure from Claude-3.5-Haiku in ScienceWorld-
1664
      Melt. The agent navigates to the workshop, through the hallway and thus should realize the path is
1665
      kitchen to hallway to workshop. However, despite this, the agent attempts the action go kitchen.
1666
1667
      O.3 DEDUCTIVE REASONING FAILURE EXAMPLE
1668
1669
      Task Description: Your task is to freeze lead. First, focus on the
1670
      → substance. Then, take actions that will cause it to change its
         state of matter.
1671
1672
```

```
1674
      Action: look around
1675
1676
      This room is called the kitchen. In it, you see:
1677
               the agent
1678
               a substance called air
               a chair. On the chair is: nothing.
1679
               a counter. On the counter is: a bowl (containing a red
1680
               → apple, a banana, an orange, a potato), a drawer.
1681
               a cupboard. The cupboard door is closed.
1682
               a freezer. The freezer door is closed.
1683
               a fridge. The fridge door is closed.
1684
               a glass jar (containing a substance called sodium
1685

→ chloride)

1686
               a lighter
1687
               a oven, which is turned off. The oven door is closed.
1688
               a painting
               a sink, which is turned off. In the sink is: nothing.
1689
               a substance called soap
1690
               a stopwatch, which is deactivated.
               a stove, which is turned off. On the stove is: nothing.
1692
               a table. On the table is: a glass cup (containing
1693
               \rightarrow nothing).
1694
               a thermometer, currently reading a temperature of 10
1695

→ degrees celsius

      You also see:
1697
               A door to the bathroom (that is open)
1698
               A door to the hallway (that is closed)
1699
               A door to the outside (that is closed)
1700
      Action: open freezer
1701
1702
1703
1704
      Above we provide an example of a spatial reasoning failure from Claude-3.7-Sonnet (zero-shot)
      in ScienceWorldFreeze. While looking for lead, the agent checks a kitchen freezer, an extremely
1705
      unlikely place to find lead.
1706
1707
1708
      O.4 INDUCTIVE REASONING FAILURE EXAMPLE
1709
1710
1711
1712
      If you pray enough, your prayers may be answered.
1713
1714
      Action: pray
1715
      If you pray enough, your prayers may be answered.
1716
1717
      Action: pray
1718
1719
      If you pray enough, your prayers may be answered.
1720
1721
      Action: pray
1722
1723
      If you pray enough, your prayers may be answered.
1724
1725
      Action: pray
1726
```

. . .

```
1728
      Above we provide an example of an inductive reasoning failure from Claude-3.5-Haiku in ZORK1.
1729
      While the game environment returns a signal that praying might result in some change, Claude-3.5-
1730
      Haiku continues to do this for 36 steps, far past when most reasonable players would have moved
1731
1732
1733
      O.5 GROUNDED REASONING FAILURE EXAMPLE
1734
1735
      This room is called the workshop. In it, you see:
1736
                the agent
                a substance called air
1737
                a table. On the table is: a battery, a green wire, a
1738
                \hookrightarrow orange light bulb, which is off, a orange wire, a red
1739

→ wire, a switch, which is off, a violet light bulb,

1740
                → which is off, a yellow light bulb, which is off.
1741
                a ultra low temperature freezer. The ultra low temperature
1742
                → freezer door is closed.
1743
      You also see:
1744
                A door to the hallway (that is open)
1745
1746
1747
      Action: go to hallway
1748
1749
      You move to the hallway.
1750
1751
1752
1753
      Action: go to hallway
1754
1755
      You move to the hallway.
1756
1757
      Action: go to kitchen
1758
      You move to the kitchen.
1759
1760
      Action: use lighter on lead
1761
1762
      The lighter heats up the lead a small amount.
1763
1764
      Action: take red wire
1765
1766
1767
1768
      Above we provide an example of an inductive reasoning failure from Claude-3.7-Sonnet (Thinking)
      in ScienceWorldBoil. We see the agent attempt to interact with an item in its context despite having
1769
      moved to another location where the item is no longer present.
1770
1771
1772
          DATA CONTAMINATION EVIDENCE FROM THINKING TRACES: EXAMPLES
1773
1774
      The blue glow of my sword suggests there's a danger nearby, likely
1775
           something from Zork lore (like a grue). Let me explore the
1776
           available paths carefully. I have the lantern on, which should
1777
           protect me from grues in dark areas. Let me try going north
          first.
1778
```

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

Q HYPERPARAMETERS

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