

# TextAtari: 100K Frames Game Playing with Language Agents

Wenhao Li<sup>1</sup> Wenwu Li<sup>1</sup> Chuyun Shen<sup>2</sup> Junjie Sheng<sup>3</sup> Zixiao Huang<sup>2</sup> Di Wu<sup>1</sup>

Yun Hua<sup>4</sup> Wei Yin<sup>5</sup> Xiangfeng Wang<sup>2</sup> Hongyuan Zha<sup>6</sup> Bo Jin<sup>1</sup>

<sup>1</sup> Tongji University, Shanghai, China <sup>2</sup> East China Normal University, Shanghai, China

<sup>3</sup> Independent Researcher, Shanghai, China

<sup>4</sup> Shanghai Jiao Tong University, Shanghai, China <sup>5</sup> Bank of Communications, Shanghai, China

<sup>6</sup> The Chinese University of Hong Kong, Shenzhen, China

{whli, wenwu, wu2002, bjin}@tongji.edu.cn, zhahy@cuhk.edu.cn

{jarvis@stu, zxhuang@stu, cyshen@stu, xfwang@cs}.ecnu.edu.cn

yinw\_8@bankcomm.com, hyyh28@sjtu.edu.cn

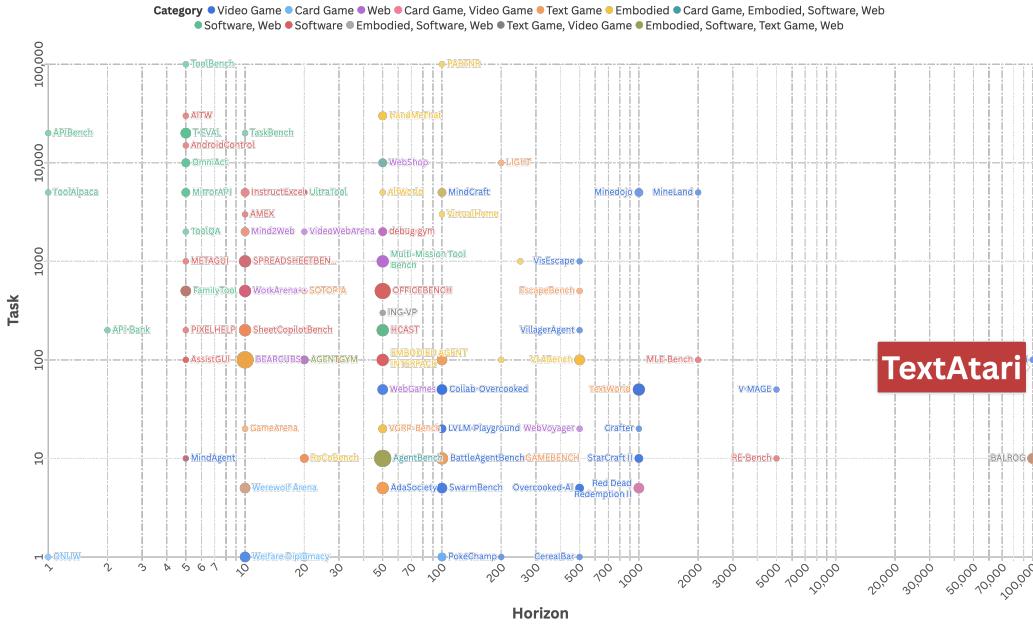


Figure 1: Statistics of tasks and horizons.

## Abstract

We present TextAtari, a comprehensive benchmark for evaluating language agents on very long-horizon decision-making tasks spanning up to 100,000 steps. By translating the visual state representations of classic Atari games into rich textual descriptions, TextAtari creates a challenging test bed that bridges sequential decision-making with natural language processing. Our benchmark encompasses nearly 100 distinct tasks with varying complexity, action spaces, and planning horizons, all rendered as text through an unsupervised representation learning

framework (AtariARI). We evaluate three open-source large language models (Qwen2.5-7B, Gemma-7B, and Llama3.1-8B) across three agent frameworks (zero-shot, few-shot chain-of-thought, and reflection reasoning) to systematically assess how different forms of prior knowledge affect performance on these unprecedented long-horizon challenges. The four distinct scenarios—Basic, Obscured, Manual Augmentation, and Reference-based—investigate the impact of semantic understanding, instruction comprehension, and expert demonstrations on agent decision-making. Our results reveal significant performance gaps between language agents and human players in these extensive planning tasks, highlighting challenges in sequential reasoning, state tracking, and strategic planning across tens of thousands of steps. TextAtari provides standardized evaluation protocols, baseline implementations, and a comprehensive framework for advancing research at the intersection of language models and planning. Our code is available at <https://github.com/Lww007/Text-Atari-Agents>.

## 1 Introduction

Sequential decision-making over extended time horizons represents one of the most fundamental challenges in artificial intelligence (Aghzal et al., 2025; Kang et al., 2024; Pignatelli et al., 2023). While humans naturally navigate complex, long-term planning scenarios—from playing strategic games to coordinating daily activities—AI systems have traditionally struggled with tasks requiring thousands of interdependent decisions. Recent analyses reveal a striking trend: the length of tasks that generalist autonomous AI agents can complete with 50% reliability has been doubling approximately every 7 months for the past 6 years (Kwa et al., 2025), as shown in Figure 2. This exponential growth suggests that within a decade, AI agents may independently complete tasks that currently take humans days or weeks—yet a critical gap remains in our ability to evaluate these systems on truly long-horizon challenges.

The AI community has developed numerous benchmarks for evaluating sequential decision-making, spanning web interfaces, desktop software, games, and embodied environments (Tan et al., 2024). However, our comprehensive analysis of 163 existing sequential decision benchmarks (see Appendix for the full list) reveals a critical limitation: most operate on remarkably short horizons. While web and software manipulation tasks typically involve fewer than 50 steps, text and video games represent the longest horizon challenges (75% approach 500 steps)—yet even these remain dramatically underexplored, with only 4 out of 163 benchmarks reaching the 100,000-step threshold\* needed to evaluate truly extended reasoning†.

This horizon gap is particularly concerning as digital games emerge as the most challenging sequential decision domains due to their unique combination of environmental complexity, non-linear decision paths, and partial observability—requiring agents to store and reason upon past experiences for effective decision-making (Ecoffet et al., 2019; Fan et al., 2022; Ma et al., 2024). As language models increasingly serve as the cognitive engine for autonomous systems, their capacity for sustained reasoning and decision-making across very long horizons becomes a critical frontier for research.

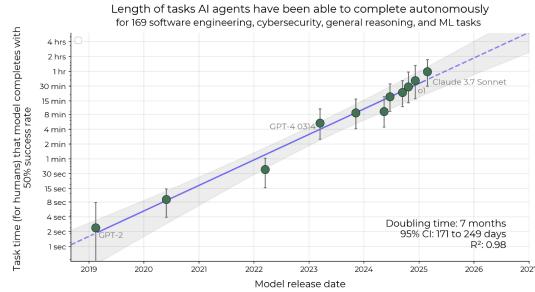


Figure 2: Screenshot from Kwa et al. (2025).

\*According to estimates from previous work (Kwa et al., 2025), current large language model agents can complete tasks that take humans less than 4 minutes with a success rate approaching 100%. Based on the statistics compiled in this paper (as shown in Figure 3), the median number of decision steps in most benchmarks is around 10 steps. Therefore, 100,000 steps would correspond to tasks that would take humans about a week to complete, this would require approximately 2-4 years of development time (Kwa et al., 2025), making the 100,000-step challenge an appropriately difficult goal that is unlikely to be achieved in the near future.

†These benchmarks with ultra-long decision-making sequences have their own limitations, which will be discussed shortly.

Recent advances in large language models (LLMs) (Anthropic, 2025; Guo et al., 2025; Jaech et al., 2024) have demonstrated remarkable capabilities in reasoning, planning, and decision-making within short contexts. These models can generate coherent multi-step plans and engage in complex reasoning tasks (Chen et al., 2025; Li et al., 2025). However, their ability to maintain consistent decision-making over very long horizons—spanning tens of thousands of steps—remains largely unexplored. This represents not merely a quantitative challenge but a qualitative shift in the nature of reasoning required: from short-term tactical decisions to long-term strategic planning with compounding consequences.

Consider the challenge of playing a video game. Human players naturally track game state, form strategic plans, adapt to changing conditions, and execute thousands of actions over extended gameplay sessions. This requires maintaining contextual awareness, making predictions about future states, and continuously adjusting strategies based on feedback—all capabilities essential for general-purpose AI systems (Lake et al., 2015, 2017; Rips & Hespos, 2015). Yet current language agents struggle with such long-horizon tasks, particularly when visual information must be processed through language descriptions rather than direct observation.

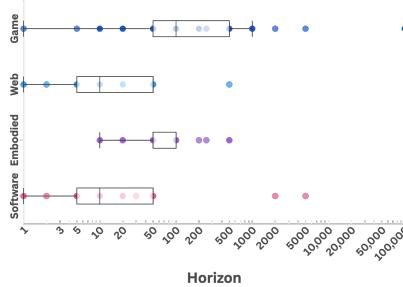


Figure 3: Horizon statistics.

In this work, we introduce TextAtari, a comprehensive benchmark for evaluating language agents on very long-horizon decision-making tasks spanning up to 100,000 steps. TextAtari transforms the visual states of classic Atari games into rich textual descriptions using an unsupervised representation learning framework (AtariARI) (Anand et al., 2019), creating a challenging testbed that bridges sequential decision-making with natural language processing. Our decision to transform visual game environments into textual representations is methodologically motivated. While visual-language models have advanced significantly, they remain limited in reasoning depth, context length handling, and decision coherence compared to text-only language models (Yang et al., 2022, 2024). This approach enables evaluation of pure reasoning

capabilities by eliminating confounding variables associated with visual processing, allows precise control over information representation, provides standardized inputs for enhanced experimental reproducibility, and aligns with the theoretical framework of language models as general reasoning engines interacting with environments through symbolic descriptions. TextAtari thus establishes a controlled experimental environment for evaluating language agents’ capacity to maintain coherent reasoning across extended horizons.

Our benchmark encompasses nearly 100 distinct tasks with varying complexity, action spaces, and planning horizons. TextAtari offers several key advantages: First, Atari games provide well-defined environments with clear objectives and measurable performance metrics. Second, by rendering these traditionally visual environments as text, we can directly evaluate language agents’ ability to process, reason about, and act upon textual information over extremely long horizons. Third, our four distinct scenario designs—Basic, Obscured, Manual Augmentation, and Reference-based—enable systematic investigation of how different forms of prior knowledge affect agent performance.

We hope TextAtari will serve as a valuable resource for the research community, providing standardized evaluation protocols, baseline implementations, and a comprehensive framework for advancing research at the intersection of language models and planning. Progress on this benchmark could lead to language agents capable of coherent decision-making across very long time horizons, bringing us closer to AI systems that can maintain consistent reasoning and planning at human-like scales.

## 2 Benchmark Design and Construction

TextAtari addresses the critical gap in existing sequential decision-making evaluations by transforming visual Atari environments into rich textual descriptions for language model processing. Our benchmark targets horizons of up to 100,000 steps—a threshold reached by fewer than 2.5% of existing benchmarks. We employed AtariARI, an unsupervised representation learning framework, to convert visual states into detailed textual descriptions that preserve essential gameplay features while creating a challenging testbed for language-based reasoning. The benchmark construction process

Detailed Description of Atari Games	
Game	Detailed Description
Venture	An exploration game where players control Winky, an adventurer navigating through a multi-room dungeon to collect treasures. Each room contains different monsters guarding treasure, requiring specific strategies to overcome. Players view the dungeon layout from an overhead perspective but transition to a zoomed-in view when entering a room. If players take too long in a room, the invincible "Hallmonster" appears, forcing swift action. The game features four different dungeons with increasing difficulty and unique monsters in each room, from snakes and trolls to giant spiders and the Grim Reaper.
VideoPinball	A digital recreation of pinball that simulates the physics and features of a traditional pinball machine. Players control left and right flippers to keep the ball in play, aiming to hit various targets to score points. The table includes bumpers, spinners, rollover targets, and bonus areas. Players can tilt the table (with limits) to influence ball direction. The game features realistic ball physics including momentum, ricochet angles, and speed changes. Special features include multiball play and bonus rounds that can be activated through specific target combinations. Scoring emphasizes both quick reflexes and strategic target selection.

Table 1: This table provides detailed descriptions of selected atari games., explaining their gameplay mechanics, objectives, and distinctive features.

involved selecting Atari games that represent diverse challenges in sequential decision-making. We prioritized games with varying complexity levels, action spaces, and planning horizons to evaluate different aspects of language agents' long-horizon reasoning capabilities.

## 2.1 Task Suite

Our task suite encompasses 23 classic Atari games spanning four major categories: Action Games, Puzzle and Strategy Games, Sports Games, and Arcade Classics. Each category presents distinct challenges for language models, testing different aspects of reasoning, planning, and decision-making over extended time horizons.

Action Games like Asteroids and Berzerk challenge models with spatial reasoning requirements and strategic target prioritization. Puzzle and Strategy Games such as Breakout and MontezumaRevenge emphasize planning and trajectory prediction, with MontezumaRevenge featuring extremely sparse rewards that require extensive planning. Sports Games including Boxing and Tennis present challenges in adversarial reasoning and anticipating opponent behaviors. Arcade Classics like MsPacman and Seaquest demand dynamic path planning and resource management in partially observable environments. Each game presents unique combinations of challenges across four key dimensions: spatial reasoning, planning and strategy, partial observability, and temporal reasoning. For instance, Seaquest requires sophisticated resource management (oxygen) while MsPacman demands strategic power pellet usage and ghost behavior modeling.

## 2.2 Environment Generation Pipeline

Our framework transforms standard Atari games from the Arcade Learning Environment (ALE) into a purely language-based interface for agents, as shown in Figure 4. Instead of pixel observations, the environment exposes a symbolic game state description at each timestep, allowing a large language model (LLM) to perceive the game through text alone. This is accomplished by extracting high-level state variables from the emulator's RAM (128 bytes) and converting them into natural language. In particular, we leverage the Atari Annotated RAM Interface (AtariARI) wrapper, which provides structured labels for key game entities and variables (e.g. player and object coordinates, scores, lives) by mapping RAM bytes to human-interpretable state information. These raw values are then annotated and linearized into textual observations using template-based descriptions.

Each game is supported by a dedicated translator module following a common design: a GameDescriber component provides static context (a brief overview of the game mechanics, objectives, and the action space in words), an ObsTranslator maps each current state vector to a sentence

Atari Games Classification and LLM Gaming Challenges (Summary)		
Game	Category	Challenges for LLM
<b>Action Games</b>		
Asteroids	Space Shooter	Spatial reasoning for circular movement Reaction-based gameplay timing Strategic target prioritization
Berzerk	Maze Shooter	Navigating complex maze layouts Dynamic obstacle avoidance Multitasking (walls, enemies, bullets)
<b>Puzzle and Strategy Games</b>		
Breakout	Brick-breaker	Geometry understanding Trajectory prediction Timing-sensitive paddle control
MontezumaRevenge	Puzzle Platformer	Extremely sparse rewards Complex dependency hierarchies Precise timing for trap avoidance
<b>Sports Games</b>		
Boxing	Sports	Adversarial reasoning Tactical positioning Timing attack and defense moves
Tennis	Sports	Court positioning strategy Opponent behavior anticipation Shot selection planning
<b>Arcade Classics</b>		
MsPacman	Maze	Dynamic path planning Ghost behavior modeling Strategic power pellet usage
Seaquest	Underwater Shooter	Resource management (oxygen) Multi-objective balancing Bidirectional threat assessment

Table 2: **Selected Atari Games and Their LLM Challenges (Summary).** This table presents key examples from each game category with their primary challenges for LLMs. Color coding indicates challenge types: spatial reasoning , planning & strategy , partial observability , and temporal reasoning .

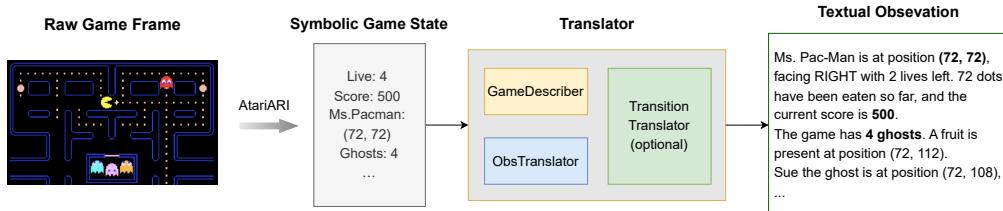


Figure 4: Environment verbalization with AtariARI.

(or set of sentences) describing the agent’s situation (for example, reporting positions of relevant objects and the current score), and a `TransitionTranslator` extends this by narrating state transitions (integrating the last state, the agent’s chosen action with a textual label, any reward obtained, and the resulting next state). This modular template structure generalizes across games – new Atari games can be integrated by implementing a similar translator with game-specific vocabulary and templates, while reusing the same interface methods for observations and transitions. For instance, the `Bowling` translator reports the ball and pin positions along with frame scores, whereas the `Boxing`

translator details both fighters’ locations and points; both adhere to the same class structure and output format.

To interface smoothly with an LLM-based agent, we throttle the observation generation to a fixed frequency (approximately 5 Hz), ensuring the agent has time to process each description and choose an action. We also enforce a limit on the description length (in tokens) so that each observation comfortably fits within the context window of the LLM. The result is a pixel-free, symbol-grounded interaction loop: the agent perceives an Atari game only through textual narratives of the game state and responds with actions accordingly, enabling direct application of language reasoning and prompting techniques to real-time game control.

To disentangle the contribution of different knowledge sources we define four textual conditions that vary only in the auxiliary information supplied to the language model, while holding the evaluation budget, sampling frequency, and backbone checkpoint fixed. Each condition introduces distinct types of prior knowledge into the prompt, allowing us to isolate their individual effects on agent performance. Below, we describe each setting in detail and provide the exact prompting format used during interaction.

**Basic Scenario.** At every step the agent receives only the live textual observation generated by the TextAtari interface. Apart from this stream and the immutable header (game synopsis, goals, legal actions), no external material is introduced, making Basic the reference against which all other conditions are measured. The prompt includes a static instruction and the latest observation with action choices.

**Obscured Scenario.** This condition tests reliance on lexical priors by replacing each domain-specific noun in the observation sentence—such as *ghost*, *paddle*, or *asteroid*—with the neutral token *item*. The transformation is executed dynamically at runtime using a fixed dictionary (e.g., “ghost” → “item”, “ball” → “item”). This forces the agent to interpret environment dynamics based on structure and positioning rather than familiar words. Numbers, coordinates, colours, scores, and the initial header remain untouched.

**Manual Augmentation Scenario.** To supply explicit rule knowledge, we prepend a concise manual excerpt to the prompt at the start of every episode. Manuals are harvested once per game from online repositories such as AtariAge. If scanned, the pages are passed through an OCR pipeline, and then summarised by a large language model (e.g., GPT-4) into  $\leq 300$  tokens capturing key information about controls, scoring mechanics, and game-end conditions. This summary is inserted after the game header and serves as grounding for the LLM during gameplay.

**Reference-based Scenario.** Here the agent is primed with an expert demonstration prior to gameplay. For each title, we train a Proximal Policy Optimisation (PPO) controller using Stable-Baselines<sup>3</sup> until it reaches at least average human performance. A single full evaluation episode is then recorded and subsampled by extracting every 10th state–action pair. The state is converted into text via the Text-Atari encoder, and the action is rendered using a task-specific verb template. These state–action entries are then concatenated chronologically into a 400-token trajectory block injected once before live inference begins. This method offers the model an in-context exemplar of competent play without directly encoding future knowledge.

### 2.3 TextAtari Statistics

Our experiments revealed significant computational demands across the TextAtari benchmark, as shown in Figure 5. The left figure illustrates the considerable variation in action space complexity across the 23 Atari games, with games like Hero, Tennis, and BattleZone exhibiting action spaces approximately three times larger than simpler games such as Skiing and Freeway. This action space diversity creates varying degrees of decision complexity that challenge language models differently.

The computational resources required for these experiments were substantial, as shown in the upper right figure. Runtime performance varied dramatically across models and games, with LLaMA3.1-8B consistently requiring the most computation time—exceeding 18,000 minutes (300 hours) for multiple games. Qwen2.5-7B and Gemma-7B demonstrated more efficient processing, typically requiring 60-70% of LLaMA3.1’s runtime. The most computationally intensive games (Zaxxon, Seaquest, and Tennis) required over 15,000 minutes of processing time per model, highlighting the extraordinary computational demands of evaluating long-horizon reasoning.

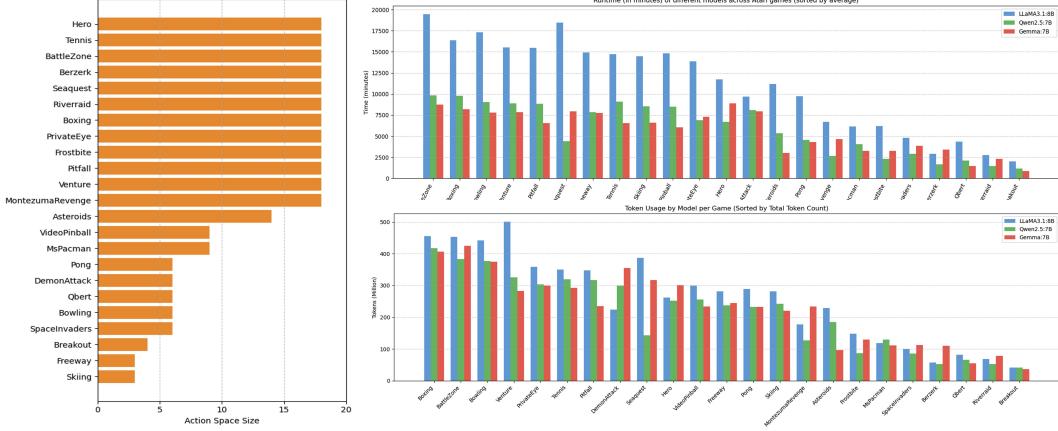


Figure 5: Computational costs.

Token consumption (lower right figure) further emphasizes the scale of these experiments. The most token-intensive games consumed between 40k-50k tokens per decision step across all models, with LLaMA3.1-8B consistently using more tokens than its counterparts. For context-heavy games like Boxing and BattleZone, this translated to billions of tokens processed throughout the full evaluation. This extreme token consumption approaches the practical limits of current context windows, particularly when agents must maintain coherent reasoning across millions steps.

In total, our comprehensive evaluation consumed approximately 820,000 GPU-minutes on A100 hardware, representing one of the most computationally intensive benchmarks for language agents to date. This substantial resource commitment underscores the challenge and importance of evaluating long-horizon reasoning capabilities.

### 3 Experiments

#### 3.1 Evaluation Protocol

All scenarios employ the identical inference agent, language model, and roll-out protocol: 23 Atari 2600 games, a horizon of 1000 interaction steps (for cost saving), and five random seeds. Because augmentation increases prompt length, a sliding-window policy discards the oldest system-level messages whenever the projected token count approaches the model’s context limit, ensuring that every query remains admissible. Consequently, observed performance differences can be attributed to the injected knowledge rather than disparities in prompt size or compute budget.

#### 3.2 Baselines

We examine three dialogue policies—Basic, Chain-of-Thought (CoT), and CoT with Reflection—that differ only in the structure and feedback they impose on the language model; all external augmentations (manual excerpts, expert trajectories, noun masking) are injected *before* prompt assembly and therefore affect the three agents identically. The discussion below focuses exclusively on each agent’s internal procedure for constructing, delivering, and updating prompts.

**Basic.** At every decision step the Basic agent issues the leanest prompt possible. It consists only of (i) a static system header that presents a one-sentence synopsis of the game, the win or termination clause, and a numbered list of legal actions, and (ii) a single user message that embeds the live Text-Atari observation. The agent neither requests a reasoning trace nor retains any form of memory; no few-shot examples or auxiliary knowledge are provided. The user instruction is fixed across all games and episodes, enforcing a strict zero-shot setting. The language model must respond with a single valid action identifier—in JSON form when required—without any additional commentary, making this template the shortest baseline against which richer prompting schemes are evaluated.

**Chain-of-Thought.** The CoT agent extends the Basic template by demanding explicit step-by-step reasoning. A leading system instruction frames the model as an expert Atari player and mandates a JSON reply with two keys—"thought process" and "action". Immediately thereafter the agent appends the game synopsis, the win or termination clause, and the latest observation, followed by a **fixed user instruction template** that drives the reasoning routine. This prompt asks the model to pause, articulate its internal deliberation, and emit a machine-readable decision. To stabilise style, a handful of demonstration dialogues illustrating ideal chain-of-thought reasoning are prefixed; these exemplars remain fixed across games. Optional long-term memory (archived reasoning traces and episode rewards) and short-term memory (the most recent state-action pairs) are inserted when token budget permits. A running counter tracks prompt length and discards the oldest background messages—first long-term memory, then exemplars—once the predicted total nears the context window, ensuring a deliverable query at every step.

**CoT with Reflection.** The Reflection agent carries out step-wise reasoning exactly as in CoT but inserts a self-critique phase at the end of every episode. When the terminal state is reached, the agent supplies the model with its complete thought trace, the realised reward trajectory, and the following system instruction. The model's JSON reply is stored as a *reflection block*. At the start of the next episode up to three of the most recent reflection blocks are inserted as additional system messages labelled *Recent Plans*, giving the model a concise self-generated briefing on past mistakes, lessons, and intended adjustments. Older reflections are discarded whenever their inclusion would push the prompt beyond the context window, ensuring consistent token budgets. During gameplay the per-step prompt and action-parsing logic remain identical to CoT, but every episode begins with a short, data-driven feedback loop that encourages iterative policy refinement without parameter updates.

### 3.3 Results

We evaluate three open-source large language models (Qwen2.5-7B, Gemma-7B, and Llama3.1-8B) across three agent frameworks (zero-shot, few-shot chain-of-thought, and reflection reasoning) on TextAtari. Our findings reveal that language models struggle significantly with very long-horizon tasks, with performance across over 90% of scenarios falling below 10% of human capability. Only in two specific tasks did the best agent configurations approach or marginally exceed human performance. Prior knowledge integration—specifically game manuals and expert demonstrations—emerged as the most consistent performance driver, yielding average improvements exceeding 100% across models and tasks. Surprisingly, reasoning-enhancement techniques such as chain-of-thought prompting and reflection mechanisms showed inconsistent benefits, with effectiveness varying dramatically across model architectures and game environments. This variability underscores the fundamental challenges in maintaining coherent state tracking, strategic planning, and decision consistency across extended time horizons. The substantial performance gap between even the best language agents and human players highlights the difficulty of maintaining coherent decision-making over tens of thousands of steps, suggesting that current language models, regardless of architecture or prompting technique, lack the cognitive mechanisms necessary for truly extended reasoning.

## 4 Conclusion

We introduced TextAtari, a comprehensive benchmark for evaluating language agents on very long-horizon decision-making tasks spanning up to 100,000 steps. By transforming visual Atari games into textual descriptions, we created a challenging testbed bridging sequential decision-making with natural language processing. Our experimental results revealed significant challenges in long-horizon planning for current language models. Even the best agent configurations achieved less than 10% of human performance across over 90% of tasks, with only two specific scenarios approaching human-level competence. Prior knowledge integration (game manuals and expert demonstrations)

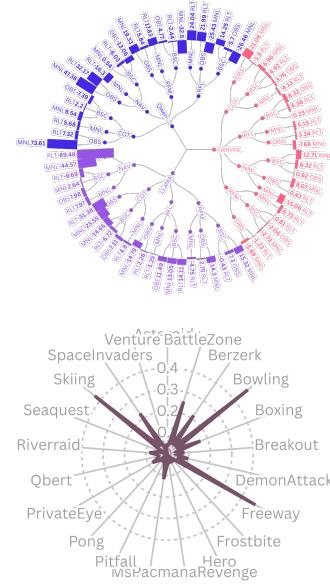


Figure 6: Relative performances.

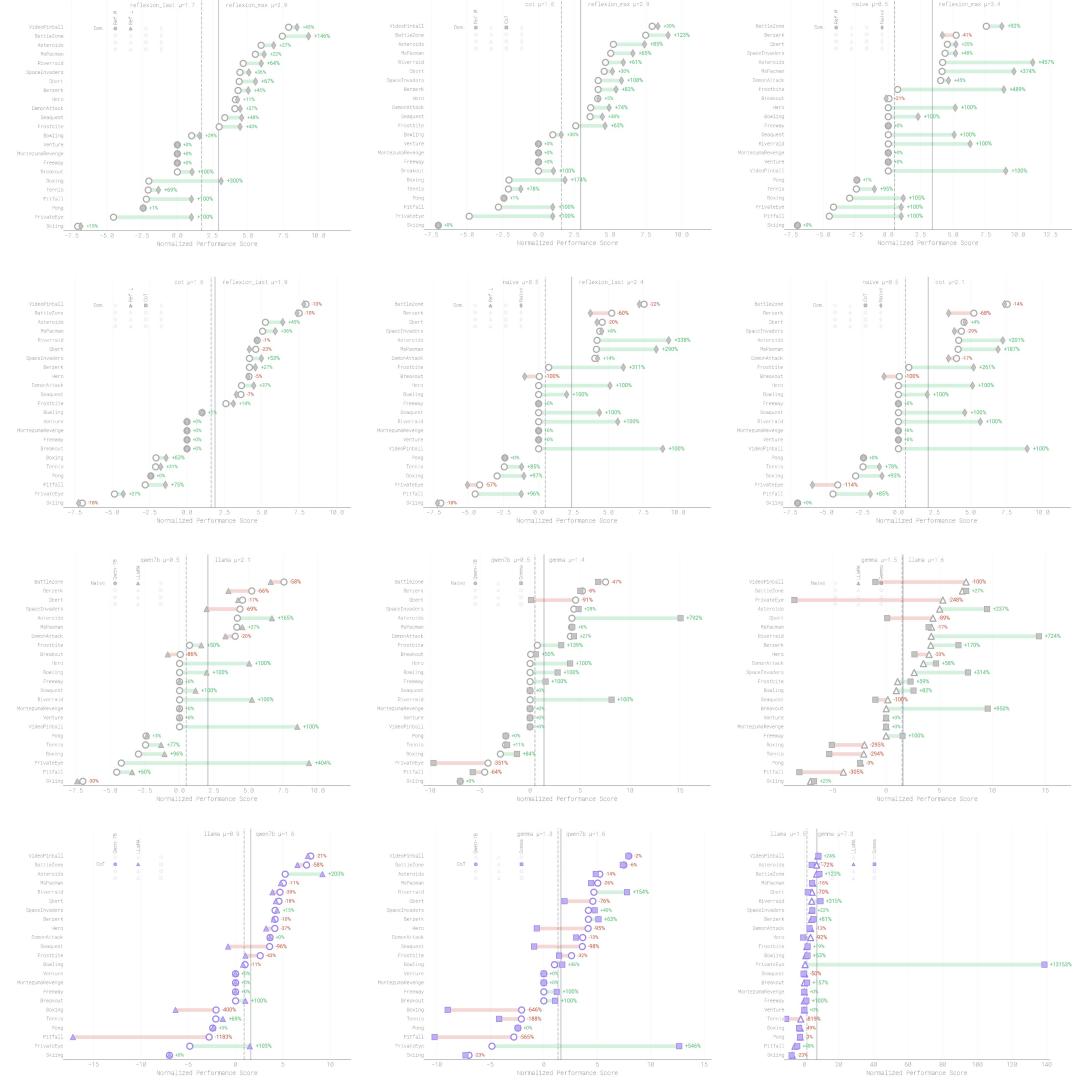


Figure 7: Selected performance comparison. See Appendix for more details.

yielded the most consistent improvements, averaging over 100% performance gains, while reasoning enhancement techniques showed inconsistent benefits across model architectures and environments. TextAtari addresses a critical gap in existing benchmarks by specifically targeting language-based reasoning over extended temporal scales. Progress on this benchmark could advance autonomous agent development toward systems capable of maintaining consistent reasoning and planning at human-like scales.

**Limitations and future work** TextAtari has several limitations suggesting directions for future research. The benchmark’s substantial computational demands—approximately 820,000 GPU-minutes with some games requiring over 300 hours per model and up to 50k tokens per decision step—may limit accessibility. Future work should explore more efficient evaluation protocols without sacrificing benchmark validity. The transformation of visual environments into text, while methodologically sound, introduces potential information loss. Future iterations could explore multimodal extensions and evaluate how different textual representation granularities affect performance. Additionally, as new architectures emerge specifically targeting sequential reasoning and memory management, TextAtari should evolve accordingly. Finally, while our benchmark demonstrates the significant gap between current AI systems and human capabilities, it does not fully diagnose the specific cognitive mechanisms responsible for this discrepancy. Future work could incorporate more detailed error analysis and causal interventions to identify specific reasoning bottlenecks, guiding targeted architectural improvements for long-horizon reasoning in language models.

## References

- Aghzal, M., Plaku, E., Stein, G. J., and Yao, Z. A survey on large language models for automated planning. *arXiv preprint arXiv:2502.12435*, 2025.
- Anand, A., Racah, E., Ozair, S., Bengio, Y., Côté, M.-A., and Hjelm, R. D. Unsupervised state representation learning in atari. In *NeurIPS*, 2019.
- Anthropic. Claude 3.7 sonnet, 2025. URL <https://www.anthropic.com/clause/sonnet>.
- Chen, Q., Qin, L., Liu, J., Peng, D., Guan, J., Wang, P., Hu, M., Zhou, Y., Gao, T., and Che, W. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.
- Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O., and Clune, J. Go-explore: a new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.
- Fan, L., Wang, G., Jiang, Y., Mandlekar, A., Yang, Y., Zhu, H., Tang, A., Huang, D.-A., Zhu, Y., and Anandkumar, A. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *NeurIPS*, 2022.
- Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- Kang, L., Zhao, Z., Hsu, D., and Lee, W. S. On the empirical complexity of reasoning and planning in llms. *arXiv preprint arXiv:2404.11041*, 2024.
- Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Von Arx, S., et al. Measuring ai ability to complete long tasks. *arXiv preprint arXiv:2503.14499*, 2025.
- Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.
- Li, Z.-Z., Zhang, D., Zhang, M.-L., Zhang, J., Liu, Z., Yao, Y., Xu, H., Zheng, J., Wang, P.-J., Chen, X., et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025.
- Ma, W., Mi, Q., Zeng, Y., Yan, X., Lin, R., Wu, Y., Wang, J., and Zhang, H. Large language models play starcraft ii: Benchmarks and a chain of summarization approach. In *NeurIPS*, 2024.
- Pignatelli, E., Ferret, J., Geist, M., Mesnard, T., van Hasselt, H., Pietquin, O., and Toni, L. A survey of temporal credit assignment in deep reinforcement learning. *arXiv preprint arXiv:2312.01072*, 2023.
- Rips, L. J. and Hespos, S. J. Divisions of the physical world: Concepts of objects and substances. *Psychological bulletin*, 141(4):786, 2015.
- Tan, W., Zhang, W., Xu, X., Xia, H., Ding, Z., Li, B., Zhou, B., Yue, J., Jiang, J., Li, Y., et al. Cradle: Empowering foundation agents towards general computer control. *arXiv preprint arXiv:2403.03186*, 2024.
- Yang, M., Schuurmans, D., Abbeel, P., and Nachum, O. Dichotomy of control: Separating what you can control from what you cannot. *arXiv preprint arXiv:2210.13435*, 2022.
- Yang, S., Walker, J., Parker-Holder, J., Du, Y., Bruce, J., Barreto, A., Abbeel, P., and Schuurmans, D. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024.

## A TextAtari Supplementary Material

- Section **B**: Related Work
- Section **C**: Task Details
- Section **D**: Prompt Engineering
- Section **E**: Missing Results
- Section **F**: Border Impact

## B Related Work

This section provides a comprehensive analysis of existing sequential decision-making benchmarks to contextualize TextAtari’s contribution. We surveyed 163 benchmarks across multiple domains—Video Games, Web Interactions, Software Operations, Text Games, Card Games, and Embodied Tasks—to evaluate the current landscape of agent evaluation frameworks.

Our analysis reveals a critical limitation in existing benchmarks: most operate on remarkably short horizons, with the median decision step count remaining below 50 for web and software manipulation tasks. Even among text and video games, which represent the longest horizon challenges, 75% approach only 500 steps. As illustrated in Figure 3 of the main paper, fewer than 2.5% of existing benchmarks reach the 100,000-step threshold that TextAtari addresses.

Table 3-5 catalog these benchmarks, highlighting key characteristics including domain category, decision horizon (number of steps required for task completion), number of distinct tasks, and agent type (single-agent or multi-agent). This collection provides empirical evidence for the “horizon gap” discussed in the introduction—the absence of standardized evaluations for truly long-horizon sequential decision-making tasks that would take humans days or weeks to complete.

TextAtari addresses this gap by providing a standardized framework for evaluating language agents’ capacity to maintain coherent reasoning and decision-making across extended horizons of up to 100,000 steps. Unlike most existing benchmarks that focus on short-term tactical decisions, TextAtari challenges agents with long-term strategic planning where consequences compound over tens of thousands of steps—a qualitative shift in reasoning requirements that better approximates real-world extended planning challenges.

### B.1 Compare to Existing 100K-Step Benchmarks

NetHack Learning Environment ([Küttler et al., 2020; Paglieri et al., 2024](#)), while technically supporting long gameplay sessions of tens of thousands of steps like TextAtari, suffers from several critical limitations. The complexity of NetHack creates an extremely sparse reward landscape that makes systematic evaluation challenging—agents often fail catastrophically before meaningful learning can occur. Additionally, NetHack’s representation combines ASCII symbols with complex game mechanics that require substantial domain expertise to interpret properly. This creates an artificial hurdle of game-specific knowledge rather than testing general reasoning capabilities. Unlike TextAtari, which provides rich natural language descriptions accessible to any language model, NetHack’s symbolic representation obscures the underlying state, making it difficult to disentangle reasoning failures from representation understanding issues.

SmartPlay ([Wu et al., 2024b](#)), while built on Minecraft and theoretically supporting extended gameplay, fundamentally compromises on measuring true long-horizon planning by relying heavily on predefined macro actions. These high-level abstractions dramatically reduce the actual decision space—agents make far fewer genuine decisions than the step count suggests. This abstraction masks the fundamental challenges of maintaining coherent reasoning across extended sequences. In contrast, TextAtari preserves the raw granularity of decision-making, requiring agents to make individual primitive actions that compound over tens of thousands of steps, thus providing a more genuine measure of extended reasoning capabilities without artificial simplifications.

### B.2 Compare to Existing 1K-Step Benchmarks

Many existing benchmarks that approach the thousand-step range suffer from various simplifications that reduce their effectiveness for evaluating true long-horizon reasoning. Game-based benchmarks

Table 3: Comprehensive language agent benchmark collection for sequential decision making (Part I). This table catalogs benchmarks for evaluating language agents across diverse domains. Each benchmark is characterized by its domain category (e.g., Video Game, Web, Software, Text Game, Card Game, Embodied), decision horizon (number of steps required for task completion), number of distinct tasks, and agent type (single-agent or multi-agent). This collection contextualizes TextAtari’s contribution to the benchmark landscape, particularly in addressing the challenge of very long-horizon decision-making tasks (up to 100,000 steps) for language agents.

ID	Name	Category	Horizon	Tasks	Agent Type
1	StarCraft II (Ma et al., 2024)	Video Game	1000	10	multi-agent
2	Red Dead Redemption II (Tan et al., 2024)	Video Game	1000	5	single-agent
3	V-MAGE (Zheng et al., 2025)	Video Game	100, 1000, 5000	50	single-agent
4	ONUW (Jin et al., 2024)	Card Game	1	1	multi-agent
5	WebGames (Thomas et al., 2025)	Web	50	50	single-agent
6	BEARCUBS (Song et al., 2025)	Web	10	100	single-agent
7	BattleAgentBench (Wang et al., 2024c)	Video Game	100	10	multi-agent
8	Collab-Overcooked (Sun et al., 2025)	Video Game	10, 100	50	multi-agent
9	SwarmBench (Ruan et al., 2025)	Video Game	100	5	multi-agent
10	DSGBench (Tang et al., 2025)	Card Game, Video Game	10, 100, 1000	5	multi-agent
11	PokéChamp (Karten et al., 2025)	Video Game	10, 100	1	multi-agent
12	Minedojo (Fan et al., 2022)	Video Game	1000	5000	single-agent
13	TextWorld (Côté et al., 2019)	Text Game	1000	50	single-agent
14	AlfWorld (Shridhar et al., 2020b)	Embodied	50	5000	single-agent
15	BabyAI-Text (Carta et al., 2023)	Video Game	100, 1000	50	single-agent
16	AgentBench (Liu et al., 2024a)	Card Game, Embodied, Software, Web	10, 50	10	single-agent
17	GameTraversalBenchmark (Nasir et al., 2024)	Video Game	10, 100	100	single-agent
18	WorkArena++(Boisvert et al., 2024)	Web	1, 10	500	single-agent
19	EMBODIED AGENT INTERFACE(Li et al., 2024b)	Embodied	50	100	single-agent
20	VLABench (Zhang et al., 2024c)	Embodied	500	100	single-agent
21	MindCraft (White et al., 2025)	Video Game	10, 100	5000	multi-agent
22	Multi-Mission Tool Bench (Yu et al., 2025)	Software, Web	50	1000	single-agent
23	SPREADSHEETBENCH (Ma et al., 2024c)	Software	10	1000	single-agent
24	OSWorld (Xie et al., 2024b)	Software	10	500	single-agent
25	WebCanvas (Pan et al., 2024)	Web	10	1000	single-agent
26	TUR[K]INGBENCH (Xu et al., 2024b)	Web	50	100	single-agent
27	Windows Agent Arena (Bonatti et al., 2024)	Software, Web	10	100	single-agent
28	VisualAgentBench (Liu et al., 2024b)	Embodied, Software, Web	10	1000	single-agent
29	OFFICEBENCH (Wang et al., 2024f)	Software	50	500	single-agent
30	WONDERBREAD (Wornow et al., 2024)	Web	10	500	single-agent
31	EscapeBench (Qian et al., 2024)	Text Game	100, 1000, 500	500	single-agent
32	VisEscape (Lim et al., 2025)	Video Game	100, 500	1000	single-agent
33	B-MoCA (Lee et al., 2024)	Software	10	100	single-agent
34	Spider2-V (Cao et al., 2024)	Software	10, 50	500	single-agent
35	ELT-Bench (Jin et al., 2025)	Software	10, 100, 50	100	single-agent
36	VideoGUI (Lin et al., 2024b)	Software	10, 50	500	single-agent
37	TaskBench (Shen et al., 2024b)	Software, Web	10	20000	single-agent
38	AQA-Bench (Yang et al., 2024)	Text Game	10, 50	10	single-agent
39	FamilyTool (Wang et al., 2025b)	Software, Web	1, 5	500	single-agent
40	MirrorAPI (Guo et al., 2025)	Software, Web	1, 5	5000	single-agent
41	WhodunitBench (Xie et al., 2024a)	Card Game	10, 50	50	multi-agent
42	GTA (Wang et al., 2024a)	Software, Web	1, 5	500	single-agent
43	m&m’s (Ma et al., 2024b)	Software, Web	1, 5	5000	single-agent
44	DISCOVERYWORLD (Jansen et al., 2024)	Video Game	1000	50	single-agent
45	debug-gym (Yuan et al., 2025)	Software	50	2000	single-agent
46	HCAST (Rein et al., 2025)	Software	5, 50	200	single-agent
47	AdaSociety (Huang et al., 2024c)	Video Game	50	5	multi-agent
48	Mars (Tang et al., 2024)	Video Game	1000, 50	10	single-agent
49	AndroidControl (Li et al., 2024c)	Software	5	15000	single-agent
50	A3 (Chai et al., 2025)	Software	10, 5, 50	200	single-agent
51	ROBOTOUILLE (Gonzalez-Pumariega et al., 2025)	Video Game	10, 50	50	multi-agent
52	MindAgent (Gong et al., 2024)	Video Game	10, 5	10	multi-agent
53	PlanBench (Valmeekam et al., 2023)	Text Game	10, 50	1000	single-agent
54	$\tau$ -bench (Yao et al., 2024)	Software, Web	5, 50	200	single-agent
55	TheAgentCompany (Xu et al., 2024a)	Software, Web	10, 50	200	single-agent

like StarCraft II (Ma et al., 2024), Red Dead Redemption II (Tan et al., 2024), and Minecraft variants (MineDojo (Fan et al., 2022), Mars (Tang et al., 2024), MineLand (Yu et al., 2024), Crafter (Hafner, 2022)) all rely on predefined macro-actions or action abstractions that dramatically simplify the actual planning required. These abstract actions encapsulate complex sequences of decisions into single steps, creating an illusion of long-horizon planning while actually testing much shorter decision sequences.

V-MAGE (Zheng et al., 2025) remains limited to just two specific environments (SuperMario and FlappyBird), providing insufficient diversity to evaluate general reasoning capabilities across different domains. Its narrow focus on platforming mechanics fails to test the breadth of reasoning types

Table 4: Comprehensive language agent benchmark collection for sequential decision making (Part II). This table catalogs benchmarks for evaluating language agents across diverse domains. Each benchmark is characterized by its domain category (e.g., Video Game, Web, Software, Text Game, Card Game, Embodied), decision horizon (number of steps required for task completion), number of distinct tasks, and agent type (single-agent or multi-agent). This collection contextualizes TextAtari’s contribution to the benchmark landscape, particularly in addressing the challenge of very long-horizon decision-making tasks (up to 100,000 steps) for language agents.

ID	Name	Category	Horizon	Tasks	Agent Type
56	WebArena (Zhou et al., 2024a)	Web	10, 50	1000	single-agent
57	WebShop (Yao et al., 2022)	Web	5, 50	10000	single-agent
58	SHORTCUTBENCH (Shen et al., 2024a)	Software, Web	10, 50	10000	single-agent
59	BALROG (Paglieri et al., 2024)	Text Game, Video Game	10, 100000	10	single-agent
60	MLGym (Nathani et al., 2025)	Software, Web	50	10	single-agent
61	VGRP-Bench (Ren et al., 2025)	Text Game	10, 50	20	single-agent
62	INVESTORBENCH (Li et al., 2024a)	Software	250, 50	5	single-agent
63	AndroidWorld (Rawles et al., 2024)	Software	5, 50	100	single-agent
64	MobileAgentBench (Wang et al., 2024b)	Software	10, 5	100	single-agent
65	SPA-Bench (Chen et al., 2024b)	Software	5, 50	500	single-agent
66	PARTNR (Chang et al., 2024)	Embodyied	10, 100	100000	multi-agent
67	LVLM-Playground (Wang et al., 2025a)	Video Game	100, 5	20	multi-agent
68	GameArena (Hu et al., 2024)	Text Game	10, 5	20	single-agent
69	TEXTARENA (Guertler et al., 2025)	Text Game	10, 500	100	multi-agent
70	MinePlanner (Hill et al., 2023)	Video Game	100, 5	50	single-agent
71	GAMEBENCH (Costarelli et al., 2024)	Text Game	250	10	single-agent
72	GTBENCH (Duan et al., 2024)	Text Game	50	10	multi-agent
73	ING-VP (Zhang et al., 2024a)	Text Game, Video Game	50	300	single-agent
74	RoCoBench (Mandi et al., 2024)	Embodyied	20, 5	10	multi-agent
75	VillagerAgent (Dong et al., 2024)	Video Game	500	200	multi-agent
76	LLMARENA (Chen et al., 2024a)	Text Game	10, 200	10	multi-agent
77	CivRealm (Qi et al., 2024)	Video Game	1000	10	multi-agent
78	SmartPlay (Wu et al., 2024b)	Video Game	100, 100000, 200, 5	10	single-agent
79	MAgIC (Xu et al., 2024c)	Card Game	20, 5	5	multi-agent
80	AgentBoard (Ma et al., 2024a)	Embodied, Software, Text Game, Web	10, 20, 50	10	single-agent
81	AGENTGYM (Xi et al., 2024)	Embodied, Software, Text Game, Web	20	100	single-agent
82	Welfare Diplomacy (Mukobi et al., 2022)	Card Game	1, 10	1	multi-agent
83	Werewolf Arena (Bailis et al., 2024)	Card Game	10	5	multi-agent
84	MiniWoB (Shi et al., 2017)	Web	1, 10	100	single-agent
85	MiniWoB++(Liu et al., 2018)	Web	1, 10	100	single-agent
86	WorkArena(Drouin et al., 2024)	Web	10, 20	100	single-agent
87	ManiSkill (Mu et al., 2021)	Embodied	50	20	single-agent
88	LIBERO (Liu et al., 2023)	Embodied	100	100	single-agent
89	RoboCasa (Nasiriany et al., 2024)	Embodied	500	100	single-agent
90	ARNOLD (Gong et al., 2023)	Embodied	100	10	single-agent
91	Rlbench (James et al., 2020)	Embodied	200	100	single-agent
92	Overcooked-AI (Carroll et al., 2019)	Video Game	500	5	multi-agent
93	CerealBar (Pérez-Rodríguez et al., 2023)	Video Game	500	1	multi-agent
94	LLM-Coordination (Agashe et al., 2023)	Video Game	50, 500	5	multi-agent
95	MineLand (Yu et al., 2024)	Video Game	2000	5000	multi-agent
96	APIBench (Peng et al., 2022)	Software, Web	1	20000	single-agent
97	ToolBench (Xu et al., 2023b)	Software, Web	5	100000	single-agent
98	API-Bank (Li et al., 2023b)	Software, Web	2	200	single-agent
99	ToolAlpaca (Tang et al., 2023)	Software, Web	1	5000	single-agent
100	T-EVAL (Chen et al., 2024e)	Software, Web	5	20000	single-agent
101	UltraTool (Huang et al., 2024b)	Software, Web	20	5000	single-agent
102	SheetCopilotBench (Li et al., 2023a)	Software	10	200	single-agent
103	InstructExcel (Payan et al., 2023)	Software	10	5000	single-agent
104	SheetRM (Chen et al., 2024c)	Software	10	500	single-agent
105	GAIA (Mialon et al., 2024)	Software, Web	50	500	single-agent
106	Mind2Web (Deng et al., 2023)	Web	10	2000	single-agent
107	WEBLINX (Lu et al., 2024)	Web	50	2000	single-agent
108	METAGUI (Sun et al., 2022)	Software	5	1000	single-agent
109	AITW (Rawles et al., 2023)	Software	5	30000	single-agent
110	PIXELHELP (Li et al., 2020)	Software	5	200	single-agent

(strategic planning, resource management, spatial understanding) that TextAtari’s diverse game selection enables.

Text-based environments like TextWorld (Côté et al., 2019) and EscapeBench (Qian et al., 2024) offer simplified worlds with limited state spaces and highly constrained action possibilities. Their deliberately simplified environments lack the complexity, state space size, and causal depth of Atari games. BabyAI-Text (Carta et al., 2023) similarly restricts itself to basic grid-world scenarios with limited objects and interactions, creating artificially simplified planning problems.

Benchmarks like MLE-Bench (Chan et al., 2024), RE-Bench (Wijk et al., 2024), and DISCOVERYWORLD (Jansen et al., 2024) limit themselves to specialized domains (machine learning experiments and scientific discovery) that contain substantial human annotation and guidance. These embedded hints and structured exploration spaces implicitly simplify the planning challenge compared to TextAtari’s more general game environments where agents must discover effective strategies independently.

CivRealm (Qi et al., 2024), while offering complex strategic gameplay, suffers from an extremely restricted action space at each decision point combined with extensive domain-specific knowledge requirements. The game’s built-in advisors and infrastructure for managing civilization development significantly reduce the actual reasoning burden on agents. Moreover, its specialized domain of civilization building lacks the diversity of reasoning types required by TextAtari’s varied environments.

TextAtari overcomes these limitations by providing: (1) diverse environments spanning multiple reasoning types without domain specialization; (2) primitive action spaces that require genuine sequential decision-making without macro-action shortcuts; (3) natural language descriptions that eliminate the need for specialized visual processing while preserving state information; (4) standardized evaluation protocols across games with varying difficulty levels; and (5) true long-horizon challenges where decisions have compounding consequences over tens of thousands of steps. These advantages make TextAtari uniquely positioned to evaluate language agents’ capacity for extended reasoning in a way that existing benchmarks cannot match.

### B.3 Future Direction

Despite TextAtari’s strengths in evaluating long-horizon reasoning, we acknowledge that existing benchmarks offer valuable perspectives our work can benefit from. Complex video games like StarCraft II and Red Dead Redemption II present rich, visually grounded environments with emergent dynamics that more closely mirror real-world complexity. While their macro-action approach simplifies decision horizons, these games capture strategic depth and environmental richness that complement TextAtari’s focus on extended reasoning. Similarly, realistic task environments like DISCOVERYWORLD and MLE-Bench provide domain-specific challenges that better represent specialized professional reasoning in scientific discovery and machine learning experimentation—contexts where language agents may ultimately provide significant value.

We view TextAtari not as a replacement for these specialized benchmarks but as addressing a specific gap in evaluating **pure long-horizon reasoning capacity**. Our future work aims to develop a more comprehensive *meta-benchmark* that integrates these complementary strengths—combining TextAtari’s extended planning horizons with the visual grounding of complex video games, the specialized reasoning of professional domains, and the embodied interaction of Minecraft-style environments. This integrated approach would enable more holistic evaluation of language agents across multiple dimensions of intelligence: sustained reasoning over time, multimodal understanding, specialized domain knowledge application, and adaptive real-time control—potentially revealing capabilities and limitations that no single benchmark could identify in isolation.

## C More Task Details

This appendix provides comprehensive information about the Atari games used in our TextAtari benchmark. The following tables present detailed classifications and descriptions of all 23 Atari games included in our evaluation, organized by game category with specific challenges they pose for language models.

Tables 6 and 7 classify the games into four major categories (Action Games, Puzzle and Strategy Games, Sports Games, and Arcade Classics) and highlight the specific cognitive challenges each game presents for language models. These challenges are color-coded to indicate their primary nature: spatial reasoning, planning and strategy, partial observability, and temporal reasoning. This classification helps illuminate why certain games prove more difficult for language agents than others, and which cognitive capabilities are most critical for success across different game environments.

Tables 8-11 provide comprehensive descriptions of each game’s mechanics, objectives, and distinctive features. These detailed descriptions offer context for understanding the complexity of each environment and the specific demands they place on decision-making agents. The descriptions cover

aspects such as control mechanisms, scoring systems, hazards, progression structures, and unique gameplay elements that distinguish each title.

Together, these tables provide a thorough understanding of the task space encompassed by TextAtari, highlighting the diverse challenges that make this benchmark particularly valuable for evaluating long-horizon reasoning in language agents. The wide variety of game mechanics, objectives, and difficulty levels ensures that TextAtari tests a broad spectrum of reasoning capabilities essential for advanced sequential decision-making.

## D Prompt Engineering

This section provides detailed information about the prompt templates used across our TextAtari experimental framework. These prompts determine how information is presented to the language models and how reasoning is structured during gameplay.

### D.1 Scenario-Based Prompts

We developed four distinct scenario-based prompts to systematically investigate how different forms of prior knowledge affect language agent performance in long-horizon game environments.

#### Basic

```
You are an expert-level game player. Your whole response should be in
JSON format.
You are in a game. {game_description} {goal_description}

Currently, {state_description}. {action_description}
Please suggest one valid action identifier. Your Suggested Action is:
```

The Basic scenario provides minimal information, representing our control condition. The prompt includes only essential game elements: a brief game description, goal statement, current state description, and available actions. This lean template forces the language model to rely entirely on its intrinsic knowledge and reasoning capabilities without external guidance. By intentionally omitting additional context, demonstrations, or strategic hints, we establish a baseline measurement of the model’s inherent game-playing abilities.

#### Obscured

```
You are an expert-level game player. Your whole response should be in
JSON format.
You are in a game. {game_description} {goal_description}

Currently, {masked_state_description}. {action_description}
Please suggest one valid action identifier. Your Suggested Action is:
```

The Obscured scenario tests the model’s reliance on domain-specific terminology and semantic priors. This prompt systematically replaces all game-specific nouns (e.g., “ghost,” “paddle,” “asteroid”) with the generic token “item” while preserving structural information like coordinates, colors, and scores. This transformation forces the model to reason about game dynamics based purely on spatial relationships and numeric patterns rather than leveraging pre-trained associations with familiar game elements. The contrast between Basic and Obscured performance reveals how much models depend on semantic understanding versus structural reasoning.

The Manual Augmentation scenario supplements the basic prompt with explicit game knowledge through official game manual excerpts. We curate these manual snippets from authentic Atari documentation (primarily sourced from AtariAge archives), limiting them to approximately 300

### Manual Augmentation

You are in a game. {game\_description} {goal\_description}

This is the game manual for this game. You need to read it carefully and understand the content and play strategies of the game:

{manual\_excerpt}

tokens to ensure consistency across games. These excerpts provide formal explanations of game mechanics, scoring systems, strategic considerations, and win conditions. This condition tests whether explicit domain knowledge injection enhances performance and to what degree different language models can operationalize written instructions into effective gameplay strategies.

### Reference-based

You are in a game. {game\_description} {goal\_description}

This is the trajectory of playing this game using the RL algorithm.  
Please read these trajectories carefully and refer to them to make decisions during gameplay:

[{state\_1} -> {action\_1}; {state\_2} -> {action\_2}; ...]

The Reference-based scenario provides the language model with exemplar gameplay through expert demonstrations. For each game, we include trajectory samples from a trained Proximal Policy Optimization (PPO) agent that achieves at least average human-level performance. These trajectories are sampled at regular intervals (every 10th state-action pair) and formatted as state-action sequences, giving the language model concrete examples of successful gameplay without directly encoding future knowledge. This condition examines how effectively language models can learn from demonstrations and adapt observed strategies to novel game states.

## D.2 Agent Reasoning Frameworks

Beyond the knowledge variation in scenarios, we implemented three distinct reasoning frameworks that structure how language models approach decision-making.

### Basic Agent Prompt

{state\_description}.{action\_description}  
Please suggest an action based on the current game state and the information you get.  
You must select the appropriate action from the given action descriptions and cannot refrain from taking action or perform any prohibited actions.  
Your Suggested Action is:

The Basic Agent employs the simplest decision-making process, requesting a direct action selection without intermediate reasoning steps. The prompt instructs the model to suggest a valid action based on the current game state, emphasizing the need to select from available actions without abstention. This framework tests the model's ability to make intuitive decisions without explicit reasoning prompts, relying on the model's internal processing to map observations directly to actions.

### Chain-of-Thought Agent Prompt

```
Currently, {state_description}. {action_description}
Now select your action. You should first take a deep breath. Then you
should think step by step about the action selection and lay out your
thought process explicitly. After that you should decide an action
based on the thought. For the whole response, you should use JSON
format with two keys "thought process" and "action".
```

The Chain-of-Thought (CoT) Agent introduces explicit reasoning into the decision process by requiring the model to articulate its thought process before selecting an action. The prompt instructs the model to “take a deep breath” and “think step by step,” encouraging deliberate consideration of the current state, available actions, and potential outcomes. The response must follow a structured JSON format with separate fields for the thought process and final action selection. This framework tests whether explicit reasoning improves decision quality in long-horizon gaming environments.

### Reflection Agent Prompt

```
You are an analytic and game coach. You need to analyse the game and
summarize the current strategy step by step.
You will be given the history of a past experience in which you were
placed in an environment and given a task to complete. You were
unsuccessful in completing the task.
Do not summarize your environment, but rather think about the strategy
and path you took to attempt to complete the task. Think step by step
what mistakes you made leading the failure.
Then devise a concise, new plan of action that accounts for your mistake
with reference to specific actions that you should have taken.
For example, if you tried A and B but forgot C, then you should reason
that forgetting C is the key mistake.
After that, devise a plan to achieve C with environment-specific actions.
Remind yourself of the plan you will take in the next trial and give your
plan after "Plan".
Respond in JSON with four keys: "Strategy", "Knowledge", "Reflexion", and
"New Plan".
```

The Reflection Agent extends the CoT framework by incorporating episodic self-critique and strategic adaptation. After each game episode concludes, this agent prompts the model to analyze its performance, identify mistakes, extract lessons, and formulate an improved strategy for subsequent attempts. The reflection must follow a structured JSON format covering strategy analysis, knowledge updates, specific reflection on failures, and a concrete new plan. These reflections are preserved and presented to the model at the beginning of subsequent episodes, creating a feedback loop that enables iterative improvement without parameter updates. This framework tests whether meta-cognitive processes enhance performance over extended interactions.

The combination of these scenario-based knowledge variations and reasoning frameworks creates a comprehensive experimental matrix for evaluating language models’ capabilities in long-horizon decision-making tasks. By systematically controlling these variables while maintaining consistent evaluation protocols, we can isolate the specific contributions of different knowledge sources and reasoning strategies to overall performance.

## E Missing Results

In this section, we provide a comprehensive set of additional results and visualizations that extend the analysis presented in the main paper. These figures offer a detailed performance comparison across

the three dimensions of our experimental framework: model architectures (Qwen2.5-7B, Llama3.1-8B, and Gemma-7B), scenario types (Basic, Obscured, Game Manual, and Reference-based), and agent frameworks (Naive, Chain-of-Thought, Reflexion\_last, and Reflexion\_max).

**Nomenclature Note.** For clarity, we should note that throughout this appendix, the visualization labels use “RL Trajectory” to refer to what we call the “Reference-based” scenario in the main text. This naming discrepancy reflects implementation details, as the Reference-based scenario was implemented using trajectories generated by reinforcement learning (PPO) agents. The underlying experimental condition remains consistent with the Reference-based scenario described in the methodology section.

## E.1 Experimental Framework and Visualization Structure

Our experimental framework explores three critical dimensions. First, we evaluate three state-of-the-art open-source small-scale LLMs - Qwen2.5-7B, Llama3.1-8B, and Gemma-7B. These models represent different architectural approaches, training methodologies, and capabilities, allowing us to examine how fundamental model design affects performance on long-horizon sequential decision-making tasks.

Second, we investigate four distinct knowledge conditions. The Basic scenario provides only the essential game description and current observation. The Obscured scenario replaces domain-specific nouns with neutral tokens to test reliance on lexical priors. The Game Manual scenario supplements the Basic scenario with concise game manual excerpts. The Reference-based scenario (labeled as “RL Trajectory” in figures) primes the agent with expert demonstrations generated by reinforcement learning algorithms.

Third, we implement four prompting strategies. The Naive agent employs a zero-shot approach with minimal prompting. The Chain-of-Thought (CoT) agent encourages step-by-step reasoning before making decisions. The Reflexion\_last agent incorporates the most recent episode reflection to guide current gameplay. The Reflexion\_max agent utilizes the best-performing reflection from previous episodes to optimize decision-making.

Each figure in this appendix presents a systematic comparison while holding two dimensions fixed and varying the third. The visualizations use a consistent format wherein bar charts represent the relative performance differences (percentage change) across all 23 Atari environments, while vertical lines indicate the normalized average performance scores for each configuration. Each row contains multiple pairwise comparisons, allowing for direct assessment of performance trends across different experimental conditions.

The figures are organized into three major categories. The first 12 figures examine how different knowledge conditions affect performance while keeping the model architecture and agent framework constant. The second 12 figures explore how different prompting strategies affect performance while keeping the model architecture and scenario type constant. The left figures investigate how different model architectures perform while keeping the scenario type and agent framework constant.

## E.2 Key Observations

The visualization structure allows us to highlight several important experimental design elements. The comparative analysis between Basic, Obscured, Game Manual, and Reference-based scenarios reveals how different forms of prior knowledge impact performance. By comparing these scenarios across models and agent types, we can isolate the relative importance of domain-specific vocabulary, explicit rule knowledge, and expert demonstrations in guiding language model decision-making.

The comparison between Naive, CoT, Reflexion\_last, and Reflexion\_max agents helps determine whether explicit reasoning strategies and reflection mechanisms provide consistent benefits across different games and knowledge conditions. This analysis is particularly important for understanding how different forms of prompted reasoning affect long-horizon decision-making capabilities.

By comparing Qwen2.5-7B, Llama3.1-8B, and Gemma-7B under identical conditions, we can examine whether architectural differences lead to systematic performance variations in long-horizon sequential decision-making. These comparisons allow us to assess whether certain model architectures are inherently better suited to particular types of reasoning required by different Atari games.

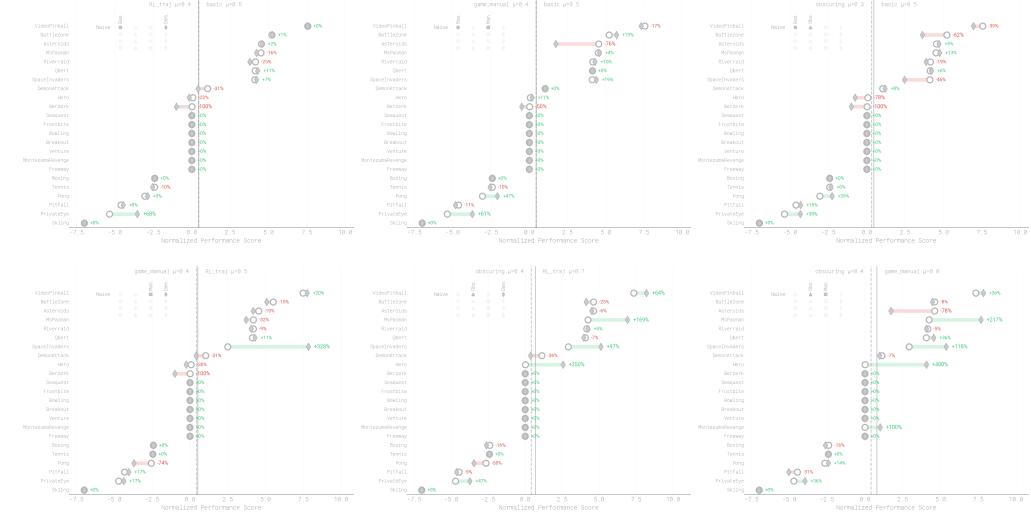


Figure 8: Performance comparison of Qwen2.5-7B with Naive agent across different scenarios. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Basic vs. RL Trajectory, (middle) Basic vs. Game Manual, and (right) Basic vs. Obscured scenarios. Bottom row: Comparison between (left) Game Manual vs. RL Trajectory, (middle) Obscured vs. RL Trajectory, and (right) Obscured vs. Game Manual scenarios.

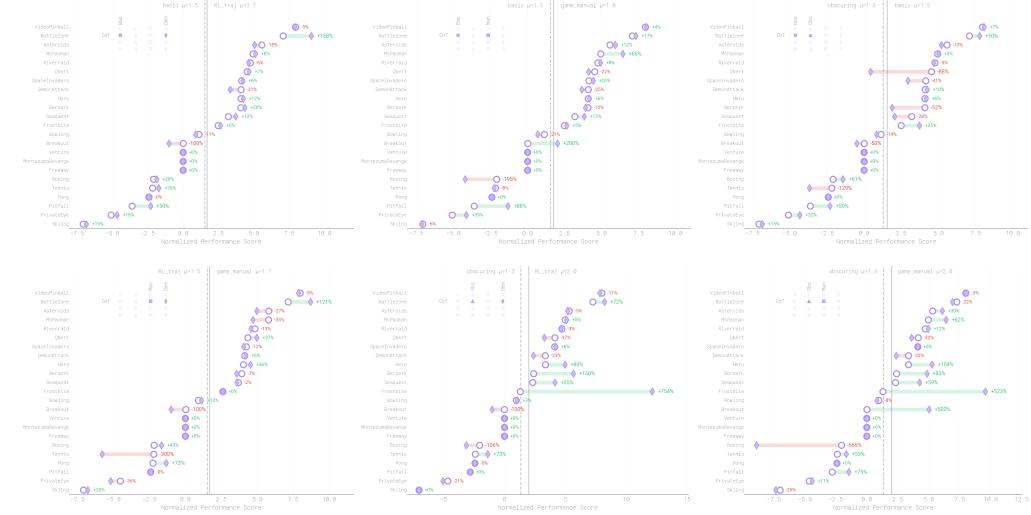


Figure 9: Performance comparison of Qwen2.5-7B with Chain-of-Thought (CoT) agent across different scenarios. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Basic vs. RL Trajectory, (middle) Basic vs. Game Manual, and (right) Basic vs. Obscured scenarios. Bottom row: Comparison between (left) Game Manual vs. RL Trajectory, (middle) Obscured vs. RL Trajectory, and (right) Obscured vs. Game Manual scenarios.

The visualization of performance across all 23 Atari environments highlights game-specific challenges and reveals which combinations of models, scenarios, and agent frameworks are most effective for particular types of games. This fine-grained analysis helps identify specific strengths and weaknesses across different experimental configurations.

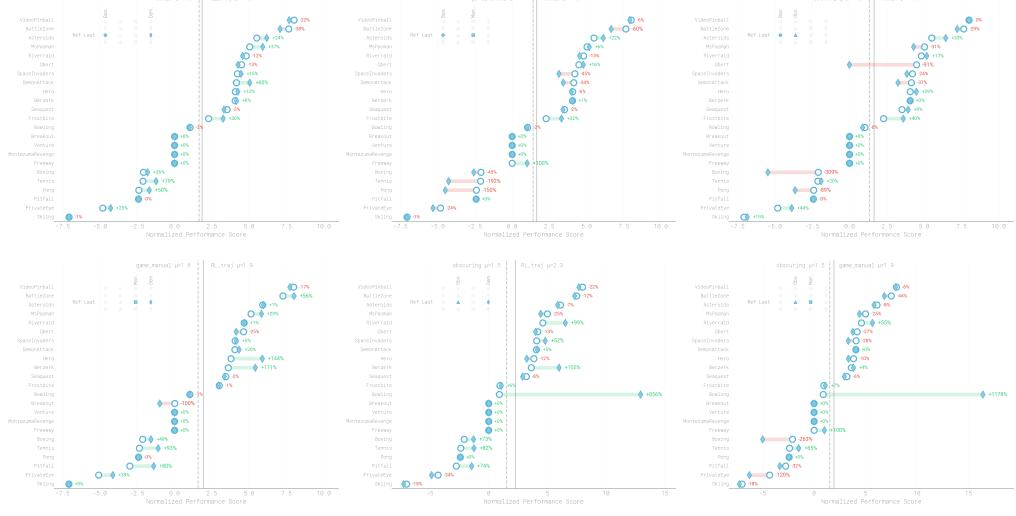


Figure 10: Performance comparison of Qwen2.5-7B with Reflexion\_last agent (using the most recent reflection) across different scenarios. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Basic vs. RL Trajectory, (middle) Basic vs. Game Manual, and (right) Basic vs. Obscured scenarios. Bottom row: Comparison between (left) Game Manual vs. RL Trajectory, (middle) Obscured vs. RL Trajectory, and (right) Obscured vs. Game Manual scenarios.

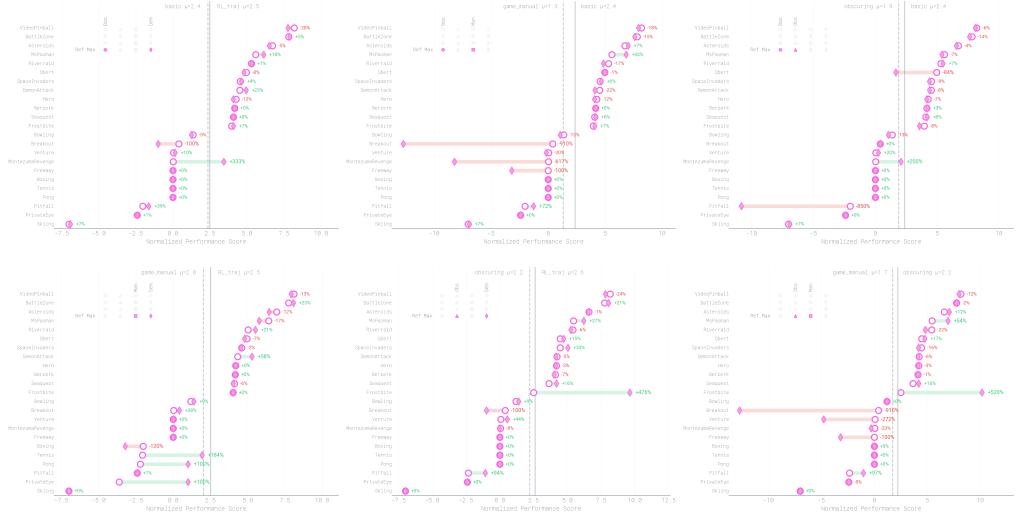


Figure 11: Performance comparison of Qwen2.5-7B with Reflexion\_max agent (using the best-performing reflection) across different scenarios. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Basic vs. RL Trajectory, (middle) Basic vs. Game Manual, and (right) Basic vs. Obscured scenarios. Bottom row: Comparison between (left) Game Manual vs. RL Trajectory, (middle) Obscured vs. RL Trajectory, and (right) Obscured vs. Game Manual scenarios.

### E.3 Methodological Considerations

These extended results should be interpreted with several methodological considerations in mind. The relative performance differences across individual games demonstrate the inherent variability in sequential decision-making tasks, suggesting that no single approach is universally optimal across all game environments. This variability underscores the importance of evaluating language agent performance across diverse task types.

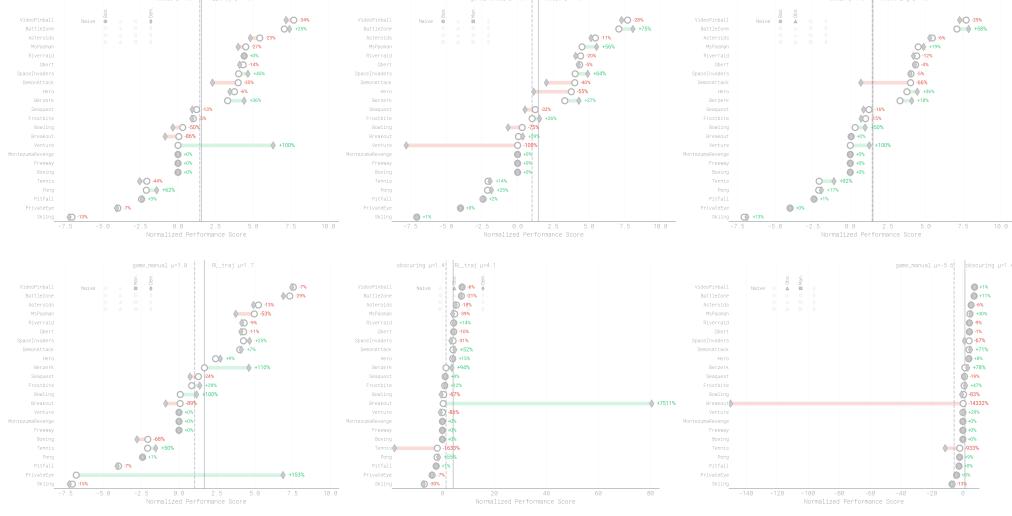


Figure 12: Performance comparison of Llama3.1-8B with Naive agent across different scenarios. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Basic vs. RL Trajectory, (middle) Basic vs. Game Manual, and (right) Basic vs. Obscured scenarios. Bottom row: Comparison between (left) Game Manual vs. RL Trajectory, (middle) Obscured vs. RL Trajectory, and (right) Obscured vs. Game Manual scenarios.

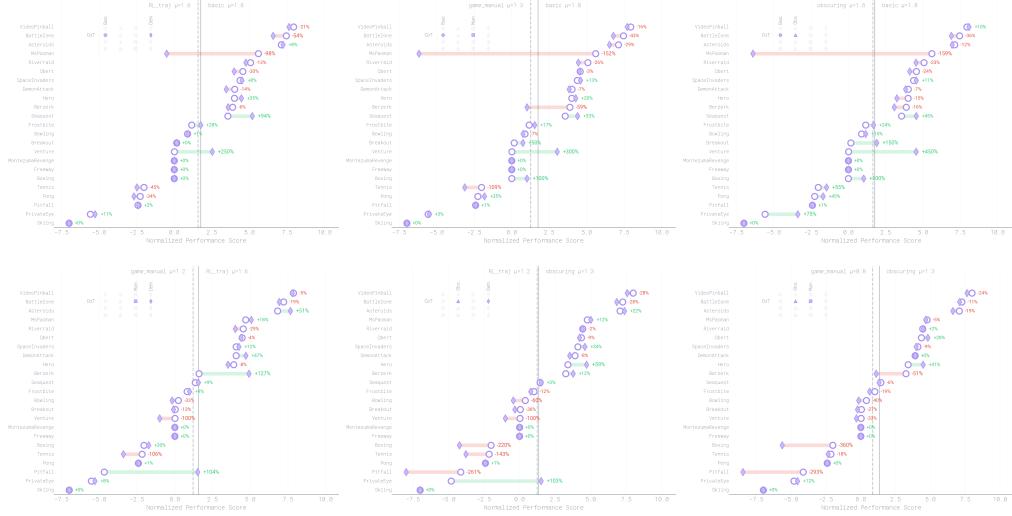


Figure 13: Performance comparison of Llama3.1-8B with Chain-of-Thought (CoT) agent across different scenarios. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Basic vs. RL Trajectory, (middle) Basic vs. Game Manual, and (right) Basic vs. Obscured scenarios. Bottom row: Comparison between (left) Game Manual vs. RL Trajectory, (middle) Obscured vs. RL Trajectory, and (right) Obscured vs. Game Manual scenarios.

All comparative analyses use one configuration as a baseline, allowing for direct assessment of relative improvements or degradations. This approach enables meaningful comparisons across different experimental conditions while controlling for game-specific factors that might otherwise confound interpretation.

As noted in the main text, the substantial computational demands of these experiments necessitated certain practical limitations, including a reduced evaluation horizon of 1,000 steps rather than the

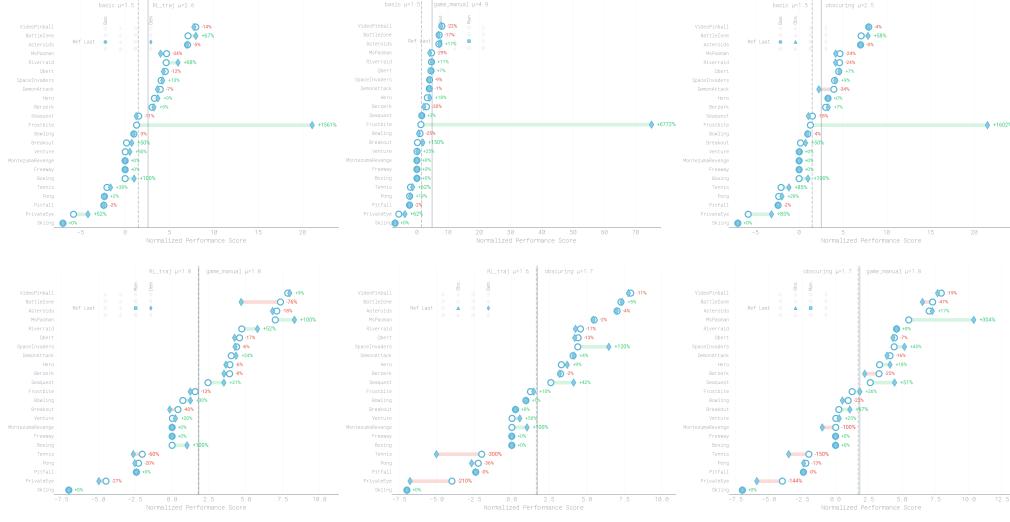


Figure 14: Performance comparison of Llama3.1-8B with Reflexion\_last agent (using the most recent reflection) across different scenarios. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Basic vs. RL Trajectory, (middle) Basic vs. Game Manual, and (right) Basic vs. Obscured scenarios. Bottom row: Comparison between (left) Game Manual vs. RL Trajectory, (middle) Obscured vs. RL Trajectory, and (right) Obscured vs. Game Manual scenarios.

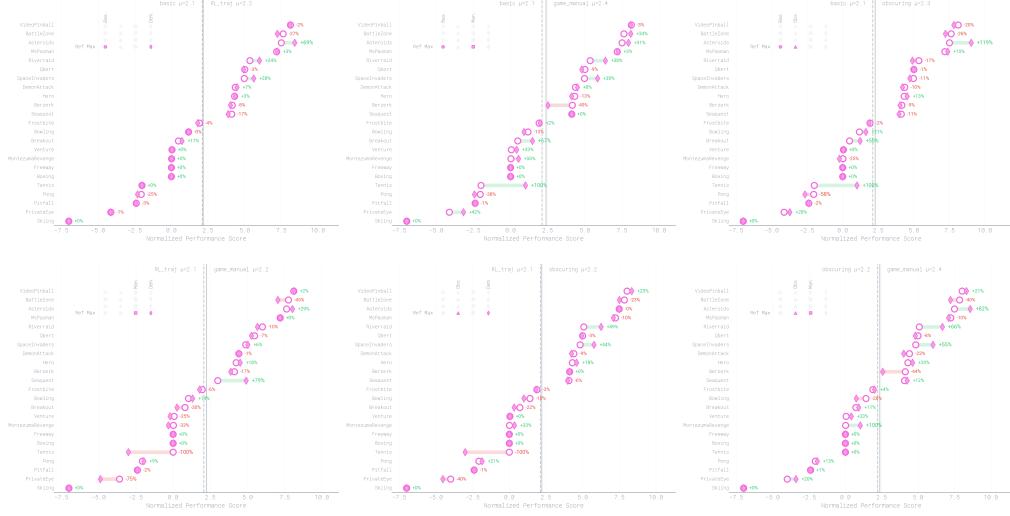


Figure 15: Performance comparison of Llama3.1-8B with Reflexion\_max agent (using the best-performing reflection) across different scenarios. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Basic vs. RL Trajectory, (middle) Basic vs. Game Manual, and (right) Basic vs. Obscured scenarios. Bottom row: Comparison between (left) Game Manual vs. RL Trajectory, (middle) Obscured vs. RL Trajectory, and (right) Obscured vs. Game Manual scenarios.

full 100,000 steps. While this reduction allows for broader experimental coverage, it may not fully capture the challenges of extremely long-horizon reasoning.

Performance on Atari games should be considered in the context of the specific challenges they present (spatial reasoning, planning, partial observability, and temporal reasoning) rather than as a general measure of language model capability. These games were selected specifically because they exercise different aspects of decision-making that are relevant to long-horizon planning.

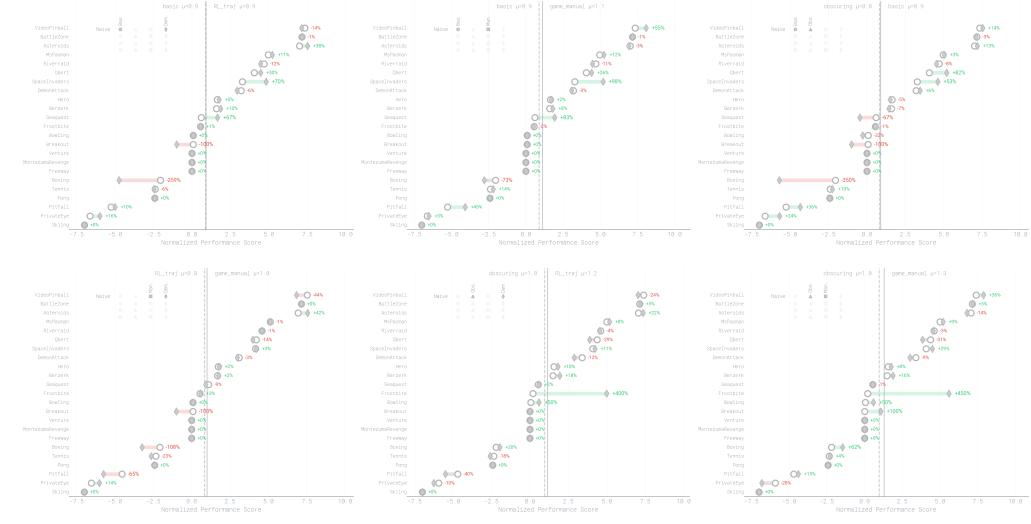


Figure 16: Performance comparison of Gemma-7B with Naive agent across different scenarios. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Basic vs. RL Trajectory, (middle) Basic vs. Game Manual, and (right) Basic vs. Obscured scenarios. Bottom row: Comparison between (left) Game Manual vs. RL Trajectory, (middle) Obscured vs. RL Trajectory, and (right) Obscured vs. Game Manual scenarios.

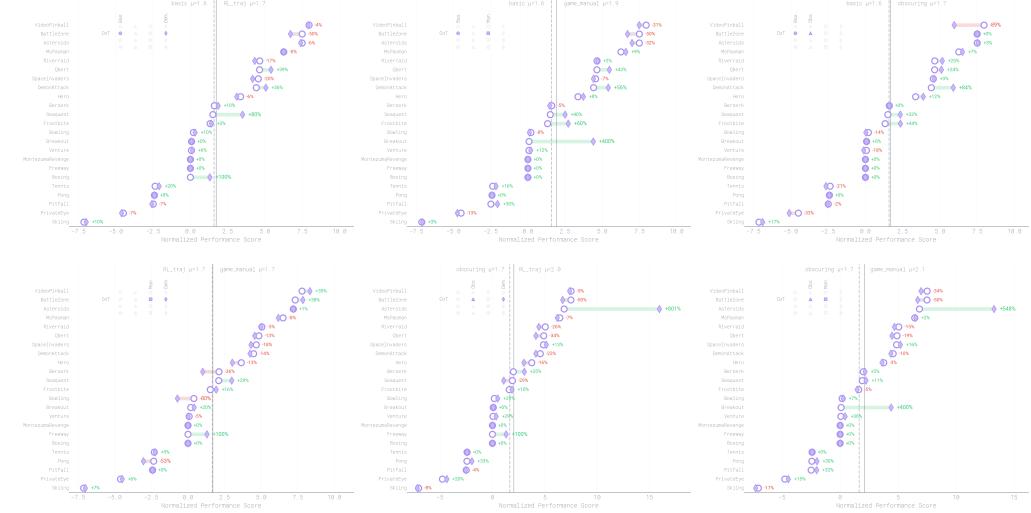


Figure 17: Performance comparison of Gemma-7B with Chain-of-Thought (CoT) agent across different scenarios. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Basic vs. RL Trajectory, (middle) Basic vs. Game Manual, and (right) Basic vs. Obscured scenarios. Bottom row: Comparison between (left) Game Manual vs. RL Trajectory, (middle) Obscured vs. RL Trajectory, and (right) Obscured vs. Game Manual scenarios.

The comprehensive nature of these visualizations allows for nuanced analysis of the interplay between model architectures, knowledge conditions, and prompting strategies in long-horizon sequential decision-making tasks. These visualizations provide valuable qualitative insights into the relative strengths and weaknesses of different approaches across the TextAtari benchmark. Together, they offer a foundation for understanding how different factors contribute to language agent performance on extended planning horizons.

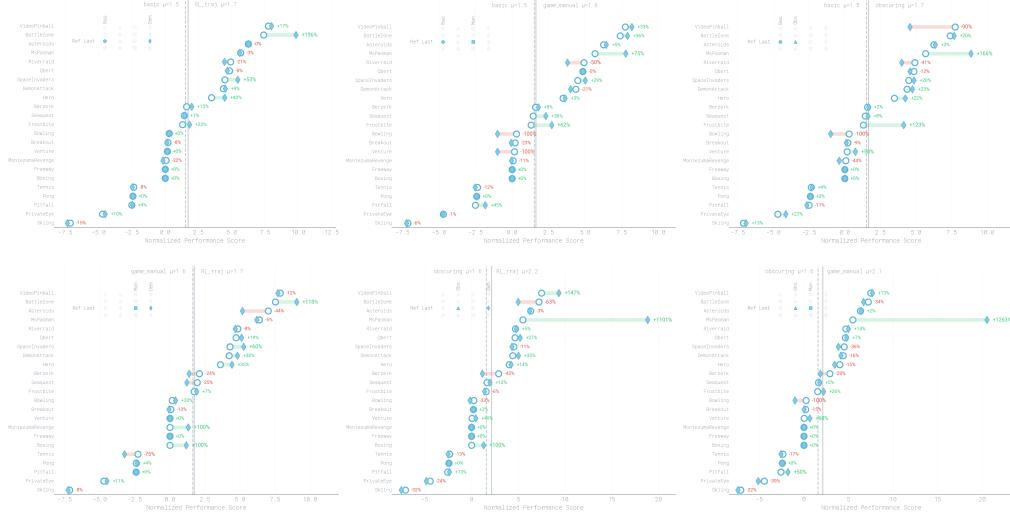


Figure 18: Performance comparison of Gemma-7B with Reflexion\_last agent (using the most recent reflection) across different scenarios. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Basic vs. RL Trajectory, (middle) Basic vs. Game Manual, and (right) Basic vs. Obscured scenarios. Bottom row: Comparison between (left) Game Manual vs. RL Trajectory, (middle) Obscured vs. RL Trajectory, and (right) Obscured vs. Game Manual scenarios.

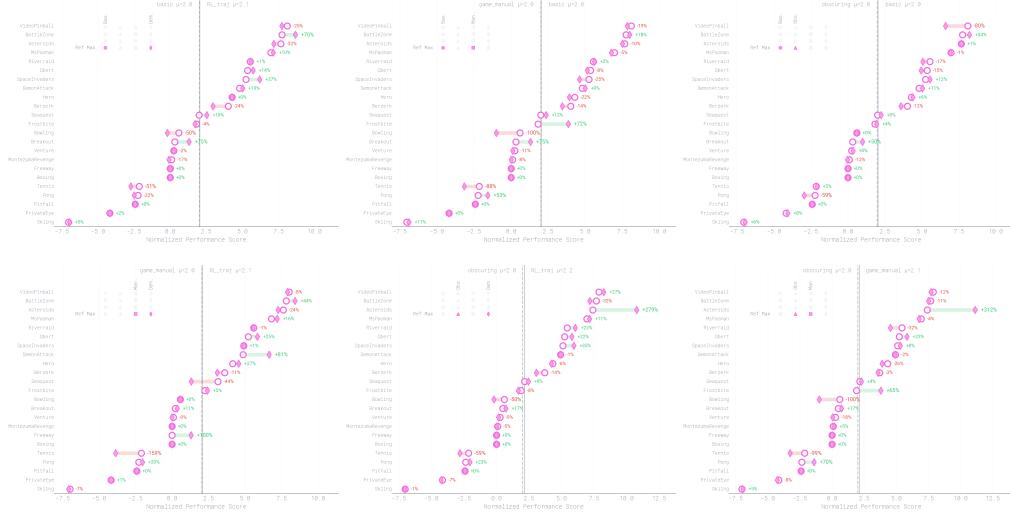


Figure 19: Performance comparison of Gemma-7B with Reflexion\_max agent (using the best-performing reflection) across different scenarios. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Basic vs. RL Trajectory, (middle) Basic vs. Game Manual, and (right) Basic vs. Obscured scenarios. Bottom row: Comparison between (left) Game Manual vs. RL Trajectory, (middle) Obscured vs. RL Trajectory, and (right) Obscured vs. Game Manual scenarios.

## F Border Impact

TextAtari introduces a benchmark for evaluating language agents on extremely long-horizon decision-making tasks, carrying various societal implications that warrant careful consideration. By establishing a rigorous evaluation standard for long-horizon reasoning, this benchmark may accelerate progress in temporal reasoning and strategic planning while highlighting the substantial gap between

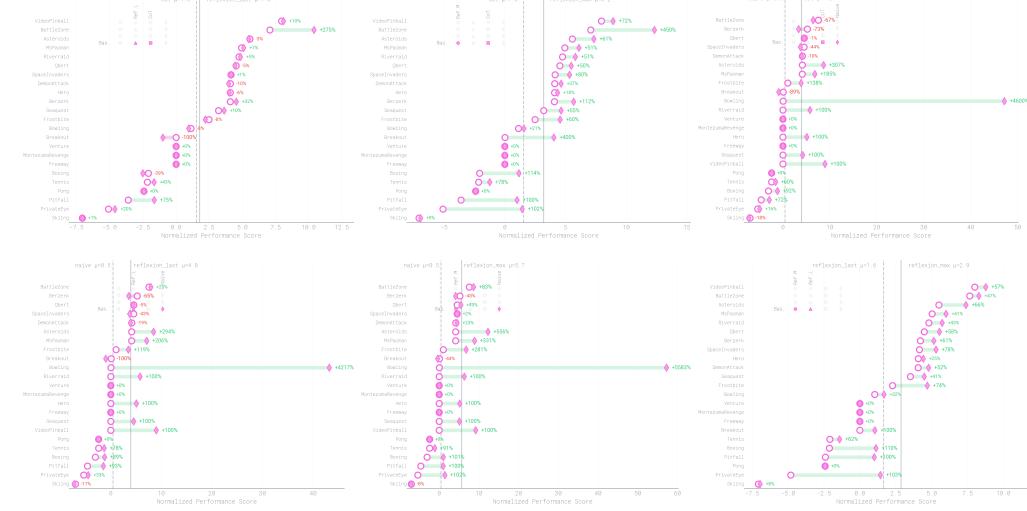


Figure 20: Performance comparison of Qwen2.5-7B in the Basic scenario across different agent types. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) CoT vs. Reflexion\_last, (middle) CoT vs. Reflexion\_max, and (right) Naive vs. CoT agents. Bottom row: Comparison between (left) Naive vs. Reflexion\_last, (middle) Naive vs. Reflexion\_max, and (right) Reflexion\_last vs. Reflexion\_max agents.

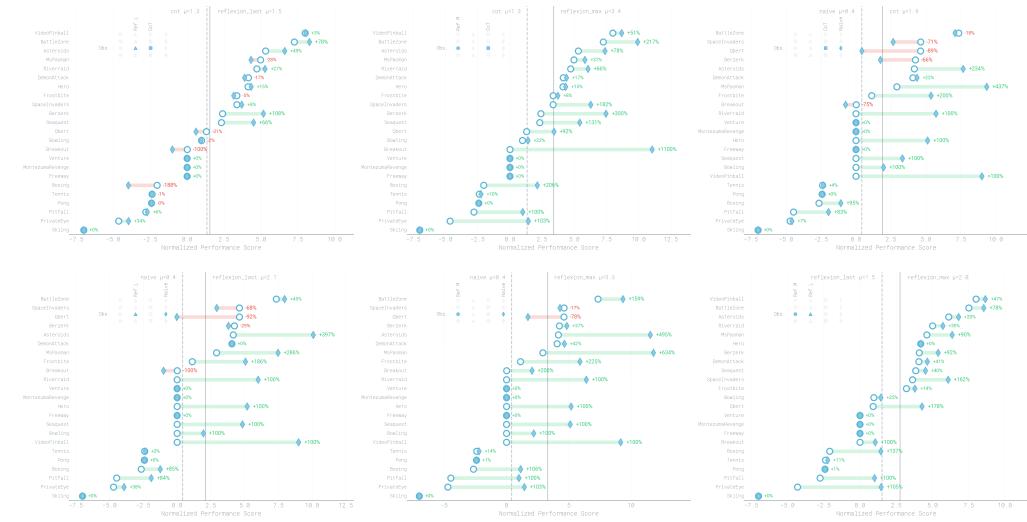


Figure 21: Performance comparison of Qwen2.5-7B in the Obscured scenario across different agent types. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) CoT vs. Reflexion\_last, (middle) CoT vs. Reflexion\_max, and (right) Naive vs. CoT agents. Bottom row: Comparison between (left) Naive vs. Reflexion\_last, (middle) Naive vs. Reflexion\_max, and (right) Reflexion\_last vs. Reflexion\_max agents.

current AI systems and human capabilities. However, the framework’s substantial computational demands raise important questions about research accessibility, environmental impact, and potential exacerbation of existing disparities in AI research.

Advances in long-horizon reasoning capabilities could transfer to beneficial applications across domains requiring extended planning, such as logistics optimization and healthcare coordination. Simultaneously, these same capabilities might enable more sophisticated autonomous systems for

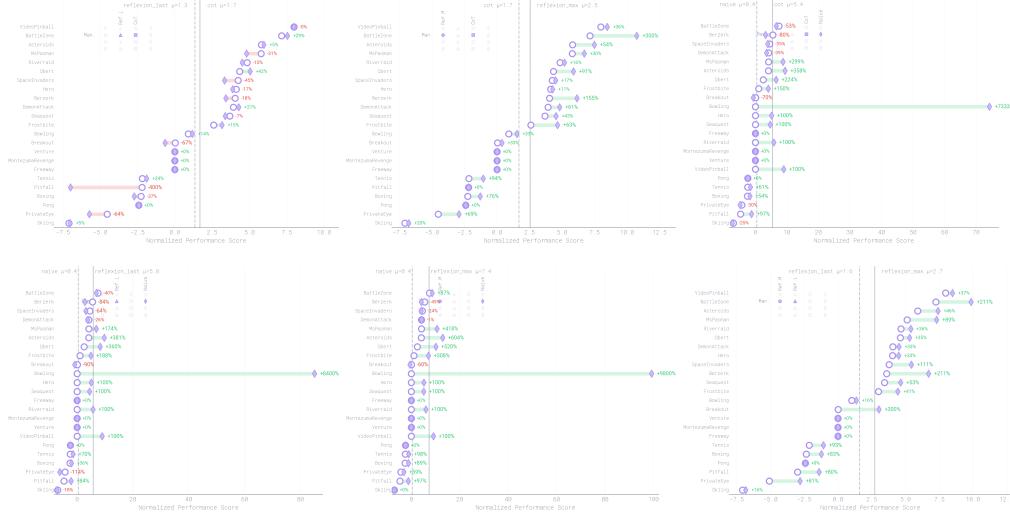


Figure 22: Performance comparison of Qwen2.5-7B in the Game Manual scenario across different agent types. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) CoT vs. Reflexion\_last, (middle) CoT vs. Reflexion\_max, and (right) Naive vs. CoT agents. Bottom row: Comparison between (left) Naive vs. Reflexion\_last, (middle) Naive vs. Reflexion\_max, and (right) Reflexion\_last vs. Reflexion\_max agents.

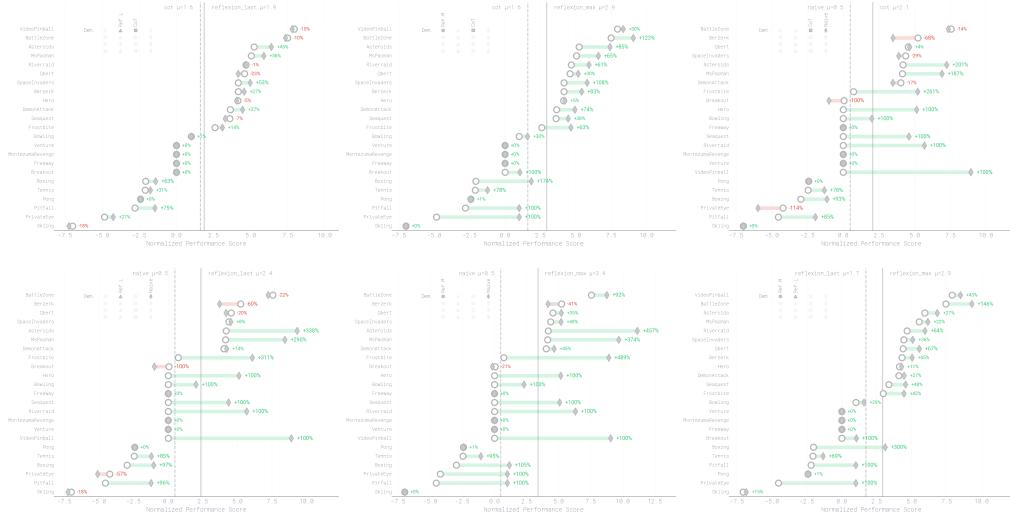


Figure 23: Performance comparison of Qwen2.5-7B in the RL Trajectory scenario across different agent types. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) CoT vs. Reflexion\_last, (middle) CoT vs. Reflexion\_max, and (right) Naive vs. CoT agents. Bottom row: Comparison between (left) Naive vs. Reflexion\_last, (middle) Naive vs. Reflexion\_max, and (right) Reflexion\_last vs. Reflexion\_max agents.

potentially harmful applications, including surveillance, automated disinformation campaigns, or autonomous weapons systems. There’s also risk that optimizing for performance on game-based benchmarks could prioritize capabilities that don’t transfer well to more nuanced real-world contexts involving ethical considerations, cultural sensitivity, or human welfare concerns.

TextAtari’s findings regarding the performance gap between language agents and humans in extended reasoning tasks could inform more effective human-AI collaboration frameworks, potentially leading

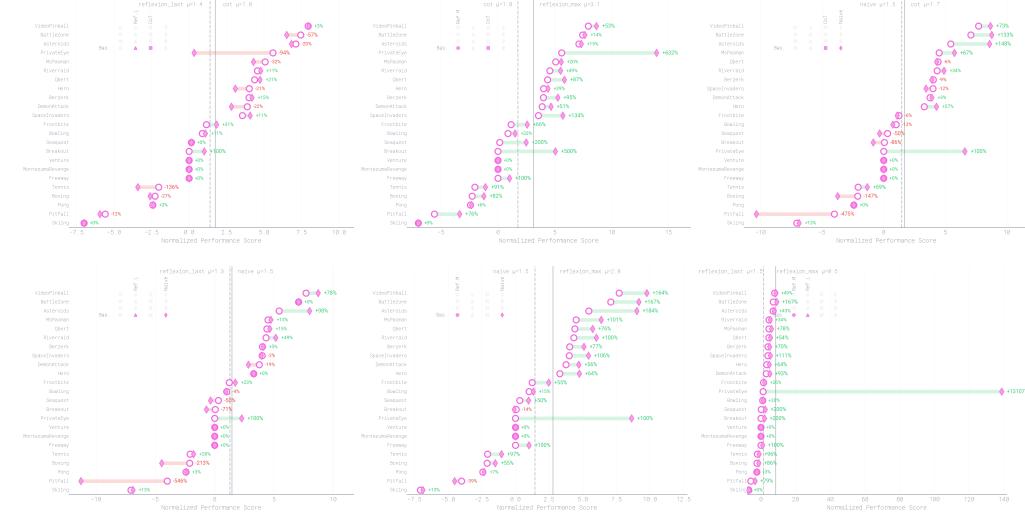


Figure 24: Performance comparison of Llama3.1-8B in the Basic scenario across different agent types. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) CoT vs. Reflexion\_last, (middle) CoT vs. Reflexion\_max, and (right) Naive vs. CoT agents. Bottom row: Comparison between (left) Naive vs. Reflexion\_last, (middle) Naive vs. Reflexion\_max, and (right) Reflexion\_last vs. Reflexion\_max agents.

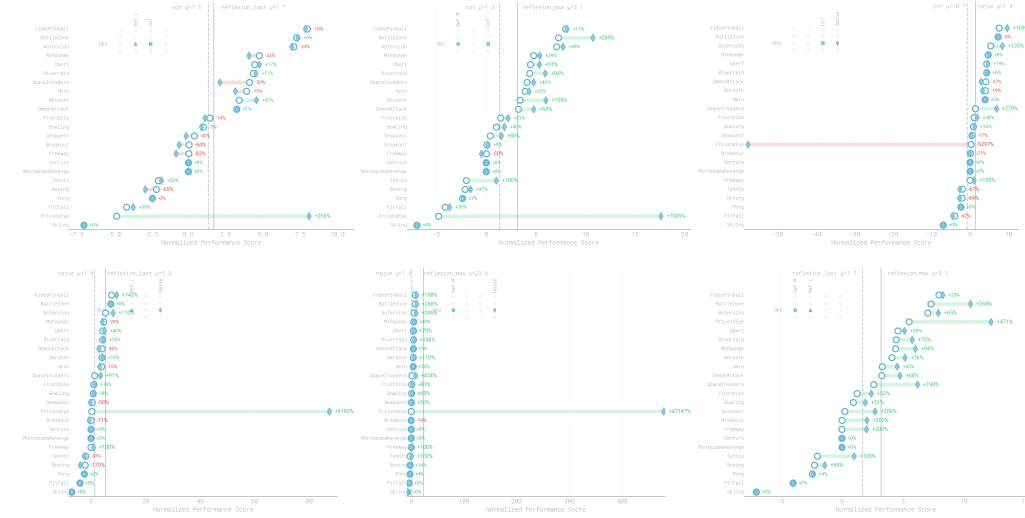


Figure 25: Performance comparison of Llama3.1-8B in the Obscured scenario across different agent types. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) CoT vs. Reflexion\_last, (middle) CoT vs. Reflexion\_max, and (right) Naive vs. CoT agents. Bottom row: Comparison between (left) Naive vs. Reflexion\_last, (middle) Naive vs. Reflexion\_max, and (right) Reflexion\_last vs. Reflexion\_max agents.

to more productive partnerships rather than full automation approaches in complex domains. The benchmark’s design, which requires explicit reasoning traces in some agent configurations, encourages more interpretable AI systems and enables analysis of how language models construct and maintain internal representations over time.

While TextAtari itself represents a controlled research environment with minimal direct risk, the capabilities it aims to measure and advance have significant implications for AI development trajectories.

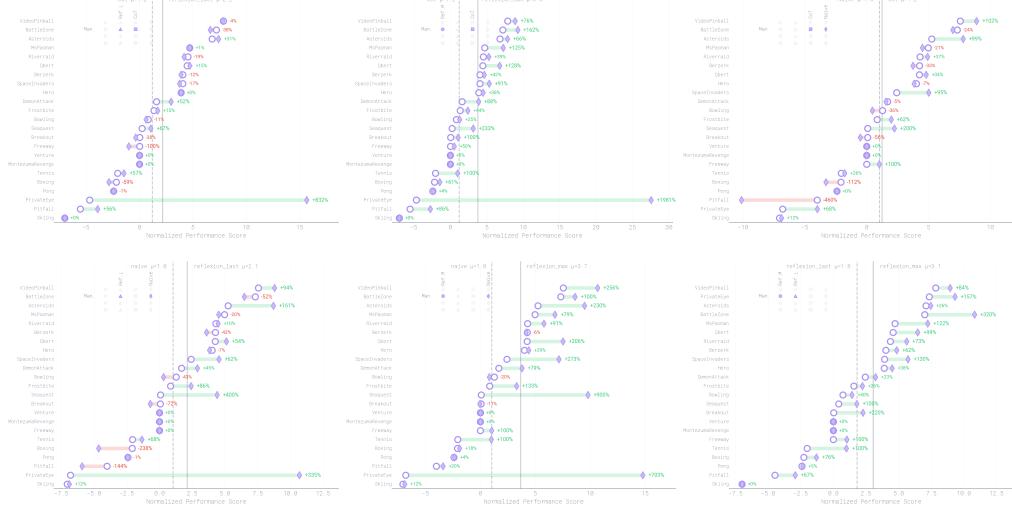


Figure 26: Performance comparison of Llama3.1-8B in the Game Manual scenario across different agent types. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) CoT vs. Reflexion\_last, (middle) CoT vs. Reflexion\_max, and (right) Naive vs. CoT agents. Bottom row: Comparison between (left) Naive vs. Reflexion\_last, (middle) Naive vs. Reflexion\_max, and (right) Reflexion\_last vs. Reflexion\_max agents.

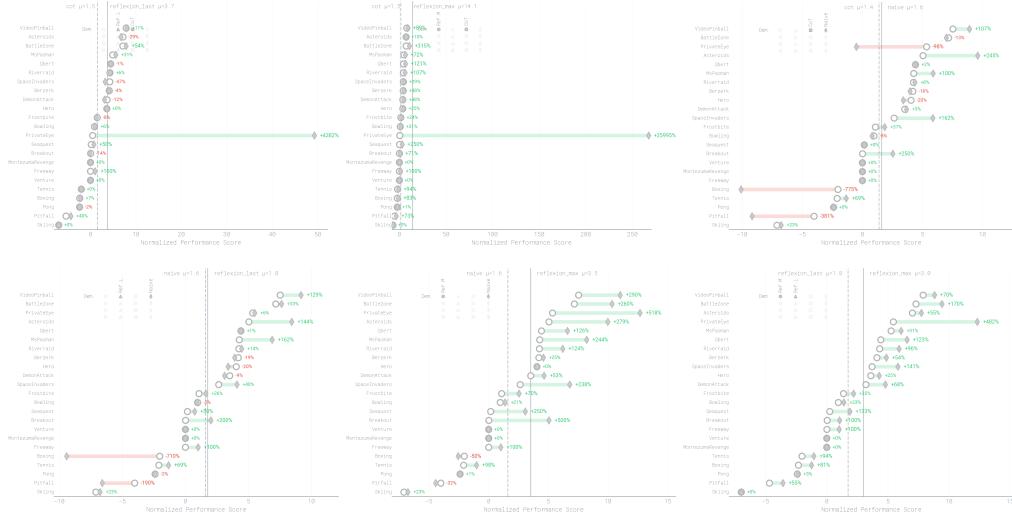


Figure 27: Performance comparison of Llama3.1-8B in the RL Trajectory scenario across different agent types. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) CoT vs. Reflexion\_last, (middle) CoT vs. Reflexion\_max, and (right) Naive vs. CoT agents. Bottom row: Comparison between (left) Naive vs. Reflexion\_last, (middle) Naive vs. Reflexion\_max, and (right) Reflexion\_last vs. Reflexion\_max agents.

The techniques developed to improve performance could be applied in both beneficial and harmful contexts—enhancing assistive technologies for individuals with cognitive impairments, but also potentially enabling more sophisticated autonomous systems for cyber attacks or manipulation. We encourage ongoing ethical reflection and governance discussions regarding long-horizon reasoning in autonomous systems as this research area progresses.

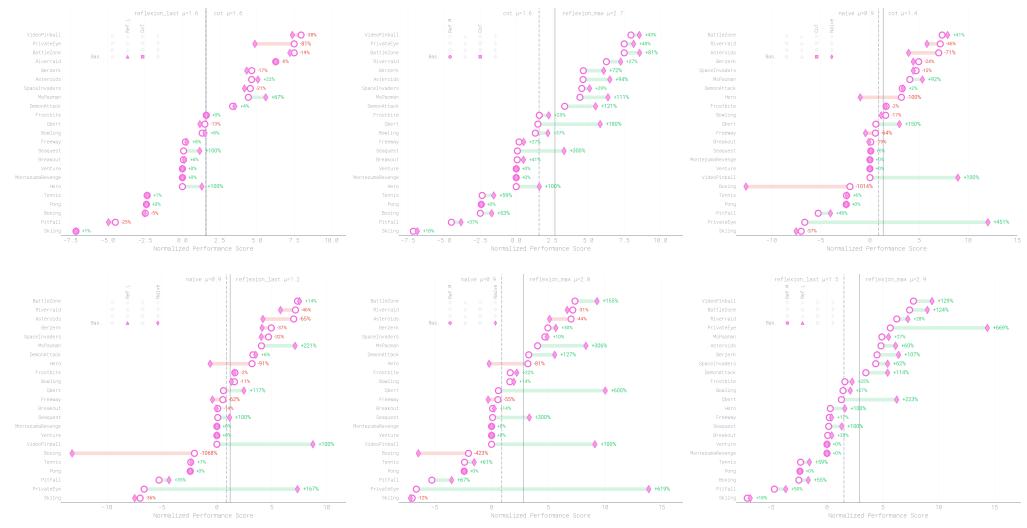


Figure 28: Performance comparison of Gemma-7B in the Basic scenario across different agent types. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) CoT vs. Reflexion\_last, (middle) CoT vs. Reflexion\_max, and (right) Naive vs. CoT agents. Bottom row: Comparison between (left) Naive vs. Reflexion\_last, (middle) Naive vs. Reflexion\_max, and (right) Reflexion\_last vs. Reflexion\_max agents.

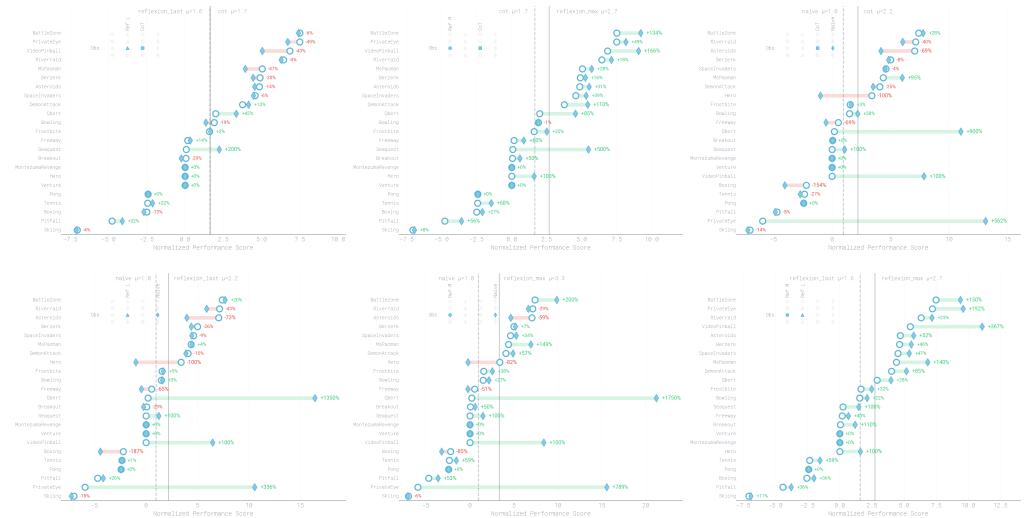


Figure 29: Performance comparison of Gemma-7B in the Obscured scenario across different agent types. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) CoT vs. Reflexion\_last, (middle) CoT vs. Reflexion\_max, and (right) Naive vs. CoT agents. Bottom row: Comparison between (left) Naive vs. Reflexion\_last, (middle) Naive vs. Reflexion\_max, and (right) Reflexion\_last vs. Reflexion\_max agents.

Table 5: Comprehensive language agent benchmark collection for sequential decision making (Part III). This table catalogs benchmarks for evaluating language agents across diverse domains. Each benchmark is characterized by its domain category (e.g., Video Game, Web, Software, Text Game, Card Game, Embodied), decision horizon (number of steps required for task completion), number of distinct tasks, and agent type (single-agent or multi-agent). This collection contextualizes TextAtari’s contribution to the benchmark landscape, particularly in addressing the challenge of very long-horizon decision-making tasks (up to 100,000 steps) for language agents.

ID	Name	Category	Horizon	Tasks	Agent Type
111	OmniAct (Kapoor et al., 2024)	Software, Web	5	10000	single-agent
112	InterCode (Yang et al., 2023)	Software	10	1000	single-agent
113	VisualWebArena (Koh et al., 2024)	Web	50	1000	single-agent
114	Mobile-Env (Zhang et al., 2023)	Software, Web	10	200	single-agent
115	AssistGUI (Gao et al., 2023)	Software	5	100	single-agent
116	ScienceWorld (Wang et al., 2022)	Video Game	100	50	single-agent
117	MoTIF (Klissarov et al., 2023)	Software	20	5000	single-agent
118	MLAgentBench (Huang et al., 2024a)	Software	20	10	single-agent
119	Spider 2.0 (Lei et al., 2025)	Software	50	500	single-agent
120	MetaTool (Wang et al., 2024d)	Software, Web	5	20000	single-agent
121	DDD (Wu et al., 2024a)	Card Game	10	5	multi-agent
122	AvalonBench (Light et al., 2023)	Card Game	20	1	multi-agent
123	SOTOPIA (Zhou et al., 2024b)	Text Game	20	500	multi-agent
124	ToolEmu (Ruan et al., 2024)	Software	10	100	single-agent
125	MiniGrid (Chevalier-Boisvert et al., 2023)	Video Game	100	20	single-agent
126	Alfred (Shridhar et al., 2020a)	Embodied	50	5	single-agent
127	NetHack LE (Küttler et al., 2020)	Text Game	100000	10	single-agent
128	Alchemy (Chen et al., 2019)	Video Game	200	1	single-agent
129	IVRE (Xu et al., 2023a)	Video Game	10	1	single-agent
130	UGIF (Venkatesh et al., 2022)	Software	5	500	single-agent
131	WebVoyager (He et al., 2024)	Web	500	20	single-agent
132	AMEX (Chai et al., 2024)	Software	10	3000	single-agent
133	AndroidArena (Xing et al., 2024)	Software	10	200	single-agent
134	AndroidLab (Xu et al., 2024d)	Software	30	100	single-agent
135	ARA (Alghamdi et al., 2024)	Software	5	10	single-agent
136	AsyncHow (Lin et al., 2024a)	Text Game	10	2000	single-agent
137	VirtualHome (Puig et al., 2018)	Embodied	100	3000	single-agent
138	WAH (Puig et al., 2021)	Embodied	250	1000	multi-agent
139	VideoWebArena (Jang et al., 2024)	Web	20	2000	single-agent
140	HandMeThat (Wan et al., 2022)	Embodied	50	30000	multi-agent
141	DialFRED (Gao et al., 2022)	Embodied	50	30000	multi-agent
142	TEACH (Padmakumar et al., 2022)	Embodied	100	5000	multi-agent
143	LIGHT (Urbanek et al., 2019)	Text Game	200	10000	multi-agent
144	Diplomacy (Bakhtin et al., 2021)	Card Game	100	1	multi-agent
145	AppWorld (Trivedi et al., 2024)	Software	20	1000	single-agent
146	ToolLLM (Qin et al., 2024)	Software, Web	5	10000	single-agent
148	ToolQA (Zhuang et al., 2023)	Software, Web	5	2000	single-agent
149	ToolLens (Qu et al., 2024)	Software, Web	5	20000	single-agent
150	Crafter (Hafner, 2022)	Video Game	1000	20	single-agent
151	Baba is AI (Cloos et al., 2024)	Video Game	100	5	single-agent
152	MiniHack (Samvelyan et al., 2021)	Text Game	100	100	single-agent
153	MLE-Bench (Chan et al., 2024)	Software	2000	100	single-agent
154	RE-Bench (Wijk et al., 2024)	Software	5000	10	single-agent
155	ScienceAgentBench (Chen et al., 2024d)	Software	10	100	single-agent
156	LlamaTouch (Zhang et al., 2024b)	Software	50	500	single-agent
157	AgentStudio (Zheng et al., 2024)	Software	10	200	single-agent
158	RoboGen (Wang et al., 2024e)	Embodied	10	100	single-agent
159	Clembench (Chalamalasetti et al., 2023)	Text Game	10	200	multi-agent
160	LMRL-Gym (Abdulhai et al., 2023)	Text Game	100	10	multi-agent
161	Game-theoretic LLM (Hua et al., 2024)	Text Game	20	10	multi-agent
162	LAMEN (Davidson et al., 2024)	Text Game	10	5	multi-agent
163	SPIN-Bench (Yao et al., 2025)	Text Game	50	5	multi-agent

Atari Games Classification and LLM Gaming Challenges (Part 1)		
Game	Category	Challenges for LLM
<b>Action Games</b>		
Asteroids	Space Shooter	Spatial reasoning for circular topology Reaction-based gameplay timing Continuous state space navigation Strategic target prioritization
BattleZone	First-Person Tank	3D spatial reasoning First-person partially observed environment Strategic target selection Situational awareness maintenance
Berzerk	Maze Shooter	Navigating complex maze layouts Dynamic obstacle avoidance Time pressure (Evil Otto pursuit) Multitasking (walls, enemies, bullets)
DemonAttack	Space Shooter	Pattern recognition in enemy waves Prediction of enemy movement patterns Progressive difficulty adaptation Timing of defensive movements
Hero	Exploration	Resource management (dynamite) Complex navigation in caverns Long-term planning for rescue Non-linear exploration paths
Pitfall	Platformer	Precise timing for obstacles Complex movement patterns Long horizon optimization (20-minute game-play) Risk/reward assessment for treasure collection
<b>Puzzle and Strategy Games</b>		
Breakout	Brick-breaker	Geometry understanding Ball trajectory prediction Strategic brick targeting Timing-sensitive paddle control
Frostbite	Building	Planning sequences under time constraints Risk assessment with moving platforms Multi-objective optimization (ice blocks vs. safety) Timing jumps between ice floes
MontezumaRevenge	Puzzle Platformer	Extremely sparse rewards Long causal chains Complex dependency hierarchies Precise timing for trap avoidance
PrivateEye	Detective	Sparse reward structure Long-term memory requirements Non-linear gameplay Large state space navigation
Qbert	Puzzle	Isometric spatial reasoning Planning efficient color-changing pathways Trap and enemy avoidance Progressive difficulty adaptation

Table 6: **Detailed Analysis of Atari Games and Their LLM Challenges (Part 1).** This table presents action and puzzle games with their specific challenges for LLMs. Color coding indicates challenge types: **spatial reasoning**, **planning & strategy**, **partial observability**, and **temporal reasoning**.

Atari Games Classification and LLM Gaming Challenges (Part 2)			
Game	Category	Challenges for LLM	
Sports Games			
Bowling	Sports	Physics understanding Precise parameter control Adapting to pin configurations Optimizing for strike probability	
Boxing	Sports	Adversarial reasoning Predicting opponent movements Tactical positioning Timing attack and defense moves	
Pong	Sports	Continuous control optimization Anticipating ball trajectory Timing paddle movements Optimizing paddle positioning	
Skiing	Sports	Precise timing for gates Path planning with fixed obstacles Speed/accuracy trade-offs Long-term performance optimization	
Tennis	Sports	Positioning for shots and court coverage Anticipating opponent strategy Shot selection and placement Balancing offensive and defensive play	
Arcade Classics			
Freeway	Obstacle Avoidance	Timing road crossings Pattern recognition in traffic flow Binary sparse reward navigation Risk assessment under time pressure	
MsPacman	Maze	Dynamic path planning Ghost behavior modeling (different behaviors) Strategic power pellet usage Adapting to maze layout variations	
Riveraid	Scrolling Shooter	Resource management (fuel) Prioritizing various hazard types Adapting to increasing difficulty Navigating narrow passages	
Seaquest	Underwater Shooter	Resource management (oxygen) Multi-objective balancing (rescue, combat, survival) Bidirectional threat assessment Risk/reward decisions for surfacing	
SpaceInvaders	Space Shooter	Strategy shifts based on remaining enemies Managing defensive shelters Adapting to increasing enemy speed Spatial awareness for shelter usage	
Venture	Dungeon Crawler	Room-to-room navigation strategy Partially observed state Risk/reward assessment for treasures Enemy pattern recognition	
VideoPinball	Simulation	Physics prediction Timing-based flipper control Understanding complex scoring mechanisms Long-term strategy for high scores	

Table 7: **Detailed Analysis of Atari Games and Their LLM Challenges (Part 2).** This table presents sports games and arcade classics with their specific challenges for LLMs. Color coding indicates challenge types: spatial reasoning, planning & strategy, partial observability, and temporal reasoning.

Detailed Description of Atari Games (Part 1)	
Game	Detailed Description
Asteroids	A space-themed shooter in vector graphics where the player controls a triangular ship in an asteroid field. The player must shoot and destroy asteroids while avoiding collisions. As asteroids are destroyed, they break into smaller pieces, creating more hazards. Occasionally, flying saucers appear and shoot at the player. The ship has momentum in a zero-gravity environment, requiring strategic thruster control and rotation. Players can hyperspace to a random location when in danger, though this carries risk.
BattleZone	One of the first 3D tank combat games using vector graphics to create a first-person perspective. Players control a tank on a flat plain with mountains in the background and various obstacles like blocks and pyramids. Enemy tanks, missiles, and flying saucers attack the player, requiring strategic positioning and aiming. A radar display helps locate enemies outside the visible field. The realistic control scheme requires separate controls for driving and turret rotation, demanding sophisticated spatial coordination.
Berzerk	A multi-directional shooter set in a maze of interconnected rooms. The player navigates through rooms filled with robots that shoot at the player. Touching robots, their bullets, or the electrified walls results in death. After a short time in any room, "Evil Otto" - an indestructible bouncing smiley face - appears and pursues the player, forcing quick movement to the next room. The robots' speech synthesis ("The humanoid must not escape!") was groundbreaking for its time. Each maze is procedurally generated, creating effectively endless gameplay.
Bowling	A simulation of ten-pin bowling where players control the position and curvature of the ball's path. The player can position their bowler horizontally, set the ball's curve, and time the release for optimal accuracy. The game accurately models pin physics, with realistic pin scatter and knockdown patterns. Players compete across 10 frames, aiming for strikes and spares to maximize their score. Different pin configurations after the first throw require adaptive targeting strategies for picking up spares.
Boxing	A top-down boxing simulation where two boxers face off in a ring. The player controls a boxer trying to land punches on the opponent while avoiding being hit. Players need to strategically position themselves, time their punches, and guard against counterattacks. The game has a three-minute time limit, and the winner is determined by either knockout or points scored through successful hits. Different punching angles and positions result in varying effectiveness, requiring strategic positioning and timing.
Breakout	A brick-breaking game where the player controls a paddle at the bottom of the screen to bounce a ball upward into layers of bricks. When hit, bricks disappear, and the ball bounces back. The goal is to eliminate all bricks without letting the ball pass the paddle. As more bricks are destroyed, the ball moves faster. Some versions feature multi-colored brick layers with higher-value bricks at the top. Breaking through to the top allows the ball to bounce around the top edge, potentially clearing many bricks rapidly.
DemonAttack	A fixed-shooter game where players control a cannon at the bottom of the screen defending against waves of demons descending from the top. Each wave features demons with different movement patterns, attack strategies, and point values. Some demons split into two when hit, while others drop bombs or dive-bomb the player. The player's cannon can move horizontally and fire upward. As levels progress, demons become faster and more aggressive, with more complex attack patterns. The game features distinctive sound effects and colorful graphics.

Table 8: **Detailed Descriptions of Selected Atari Games (Part 1).** This table provides comprehensive descriptions of seven classic Atari games, explaining their gameplay mechanics, objectives, and distinctive features.

Detailed Description of Atari Games (Part 2)	
Game	Detailed Description
Freeway	A simple but challenging game where the player controls a chicken trying to cross a ten-lane freeway filled with traffic. The goal is to reach the other side as many times as possible within the time limit. Each lane has vehicles moving at different speeds and directions. If the chicken collides with a vehicle, it is pushed backward. Players can only move up or down, requiring precise timing to navigate through gaps in traffic. The game supports two-player competitive mode where players race to get their chickens across more times than their opponent.
Frostbite	An arctic-themed game where players control Frostbite Bailey, who must build an igloo by jumping on floating ice floes in a frigid river. Each time the player lands on an ice floe, it changes color and adds a block to the igloo. Once the igloo is complete, the player must reach it to advance to the next level. Hazards include bears, geese, and crabs that patrol the ice floes. The temperature gradually drops, adding time pressure. Fish and clams appear as bonus items that can be collected for extra points.
Hero	A complex adventure game where players control Roderick Hero, a miner equipped with a helicopter backpack, dynamite, and a beam weapon. The objective is to navigate through multi-screen mine shafts to rescue trapped miners. Players must blast through walls with dynamite, defeat creatures with the beam weapon, and avoid hazards like magma and falling rocks. The helicopter backpack allows limited flight but consumes power. Lanterns throughout the mine provide light in dark areas and extra power when collected. The game features complex, non-linear level designs.
Montezuma's Revenge	A notoriously challenging platformer set in an Aztec temple. Players control an explorer named Panama Joe navigating through multiple screens of the temple to collect treasures. The game features locked doors requiring keys, deadly traps including fire pits and rolling skulls, and enemies like snakes and spiders. Players must use ladders, ropes, and disappearing floors to navigate between rooms. The game is known for its punishing difficulty, sparse rewards, and the need for precise timing and planning. Death results in returning to the starting room, making progress particularly demanding.
MsPacman	An enhanced version of Pac-Man featuring a female protagonist. Players navigate through four different maze designs consuming dots while avoiding ghosts. Power pellets allow temporary ghost consumption. Ms. Pac-Man improved upon the original with more varied mazes, ghosts with less predictable movement patterns, and bonus fruits that move around the maze rather than staying stationary. Between levels, brief cutscenes tell the love story between Ms. Pac-Man and Pac-Man. The game's less deterministic ghost behavior makes it more challenging and less susceptible to pattern-based strategies than the original.
Pitfall	A groundbreaking side-scrolling platformer where players control Pitfall Harry through a jungle collecting treasures within a 20-minute time limit. The jungle consists of 255 interconnected screens with various hazards including rolling logs, crocodiles, scorpions, quicksand, tar pits, and fires. Players navigate by running, jumping, and swinging on vines. Underground passages allow for faster travel between areas. The game was revolutionary for its era due to its large game world and fluid character animation. Points are scored by collecting treasures, with time penalties for falling into hazards.
Pong	One of the earliest and most iconic video games, simulating table tennis. Players control vertical paddles on opposite sides of the screen and must hit a ball back and forth. The ball bounces off the top and bottom edges, and points are scored when one player fails to return the ball. The ball's speed increases after several successful returns, increasing difficulty. The angle of the ball's rebound depends on which part of the paddle it hits, allowing for strategic aiming. Despite its simplicity, Pong established many foundational elements of video games and interactive entertainment.

Table 9: **Detailed Descriptions of Selected Atari Games (Part 2).** This table provides comprehensive descriptions of seven classic Atari games, explaining their gameplay mechanics, objectives, and distinctive features.

Detailed Description of Atari Games (Part 3)	
Game	Detailed Description
PrivateEye	A detective adventure game where players control Private Eye Pierre Touché solving cases by navigating a scrolling city environment. The main case involves recovering stolen items from the Goldfish Diamond case. Players drive a car through the city, entering buildings, and collecting evidence while avoiding gangsters and other hazards. The game features a complex scoring system based on catching criminals, recovering stolen items, and completing the case within the time limit. Incorrect accusations result in penalties. The game's non-linear design with multiple locations to explore was innovative for its time.
Qbert	A puzzle-platformer where players control Q*bert, an orange creature with a tubular nose, who hops around a pyramid of cubes. The objective is to change the color of each cube's top surface by hopping on it. Once all cubes are changed to the target color, the player advances to the next level. Enemies include Coily the snake, Wrong-Way and Ugg who travel along the sides of the pyramid, and red balls that fall from the top. Q*bert can use floating discs to escape to the top of the pyramid or to lure Coily off the edge. Later levels require changing each cube's color multiple times or in specific sequences.
Riveraid	A vertically scrolling shooter where players pilot a fighter jet over a river, destroying enemy helicopters, ships, fuel depots, and bridges while avoiding collisions with the shoreline. The plane consumes fuel continuously, requiring players to fly over fuel depots to refill. The river varies in width, creating narrow passages that demand precise navigation. Destroying bridges marks progression to new sections with increased difficulty. The game was notable for its use of a pseudo-random algorithm to generate the river layout, creating a different experience each play while using minimal memory.
Seaquest	An underwater shooter where players control a submarine that must rescue divers while defending against sharks, enemy submarines, and other sea creatures. The submarine can move in all directions and fire torpedoes horizontally. Players must manage their oxygen supply, surfacing periodically to replenish it and deliver rescued divers. Each surfacing with a full complement of divers awards bonus points. If oxygen depletes completely, the submarine is lost. As levels progress, enemies become more numerous and aggressive. Balancing rescue operations with defense and oxygen management creates a multi-objective challenge.
Skiing	A downhill skiing simulation where players navigate through a series of gates on a continuously scrolling course. The objective is to complete the course in the shortest time possible without missing gates, which incur time penalties. Players control their skier's horizontal position as they automatically move downward. Obstacles include trees and moguls that must be avoided. The game offers two modes: the slalom, with wider gate spacing, and the more challenging giant slalom with tighter gates. Precise control and planning the optimal line through gates are essential for achieving the best times.
SpaceInvaders	A fixed shooter game where players control a laser cannon moving horizontally at the bottom of the screen, defending against rows of descending aliens. The aliens move side to side, dropping bombs as they advance toward the bottom. Players must eliminate all aliens before they reach the ground. Protective bunkers provide temporary cover but degrade when hit by either player shots or alien bombs. As aliens are destroyed, the remaining invaders move faster. Strategic gameplay involves targeting specific aliens to manipulate their movement patterns and using the bunkers effectively for protection.
Tennis	A sports simulation of tennis from a side view of the court. Players control tennis players who can move around their side of the court and swing rackets to hit the ball over the net. The game implements basic tennis rules including serves, volleys, and scoring (15-30-40-game). Ball physics include appropriate bouncing and speed changes based on the type of hit. Strategic gameplay involves positioning for shots, timing racket swings correctly, and anticipating opponent movements. The game can be played against the computer or another human player, with varying difficulty levels for the AI opponent.

Table 10: **Detailed Descriptions of Selected Atari Games (Part 3).** This table provides comprehensive descriptions of seven classic Atari games, explaining their gameplay mechanics, objectives, and distinctive features.

Detailed Description of Atari Games (Part 4)	
Game	Detailed Description
Venture	An exploration game where players control Winky, an adventurer navigating through a multi-room dungeon to collect treasures. Each room contains different monsters guarding treasure, requiring specific strategies to overcome. Players view the dungeon layout from an overhead perspective but transition to a zoomed-in view when entering a room. If players take too long in a room, the invincible "Hallmonster" appears, forcing swift action. The game features four different dungeons with increasing difficulty and unique monsters in each room, from snakes and trolls to giant spiders and the Grim Reaper.
VideoPinball	A digital recreation of pinball that simulates the physics and features of a traditional pinball machine. Players control left and right flippers to keep the ball in play, aiming to hit various targets to score points. The table includes bumpers, spinners, rollover targets, and bonus areas. Players can tilt the table (with limits) to influence ball direction. The game features realistic ball physics including momentum, ricochet angles, and speed changes. Special features include multiball play and bonus rounds that can be activated through specific target combinations. Scoring emphasizes both quick reflexes and strategic target selection.

**Table 11: Detailed Descriptions of Selected Atari Games (Part 4).** This table provides comprehensive descriptions of the remaining classic Atari games, explaining their gameplay mechanics, objectives, and distinctive features.

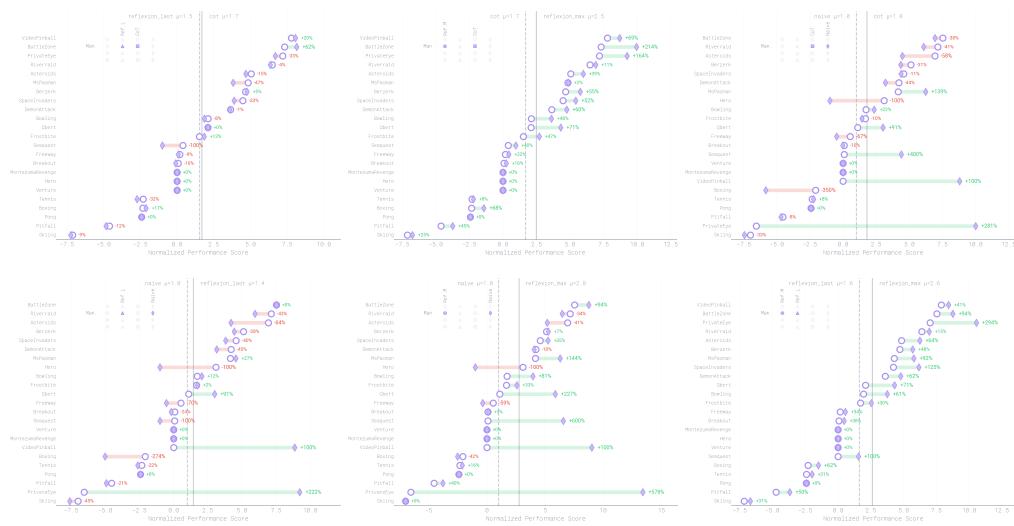


Figure 30: Performance comparison of Gemma-7B in the Game Manual scenario across different agent types. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) CoT vs. Reflexion\_last, (middle) CoT vs. Reflexion\_max, and (right) Naive vs. CoT agents. Bottom row: Comparison between (left) Naive vs. Reflexion\_last, (middle) Naive vs. Reflexion\_max, and (right) Reflexion\_last vs. Reflexion\_max agents.

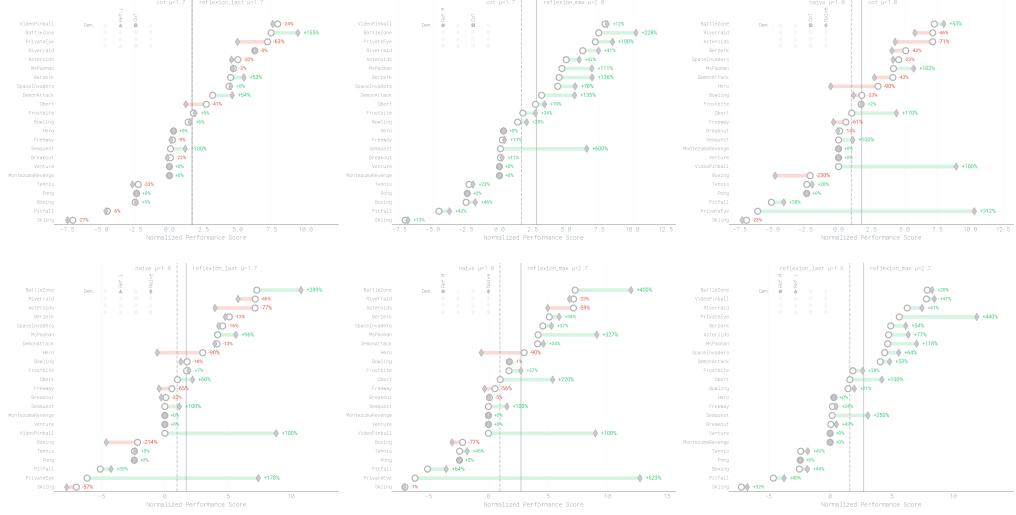


Figure 31: Performance comparison of Gemma-7B in the RL Trajectory scenario across different agent types. Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) CoT vs. Reflexion\_last, (middle) CoT vs. Reflexion\_max, and (right) Naive vs. CoT agents. Bottom row: Comparison between (left) Naive vs. Reflexion\_last, (middle) Naive vs. Reflexion\_max, and (right) Reflexion\_last vs. Reflexion\_max agents.

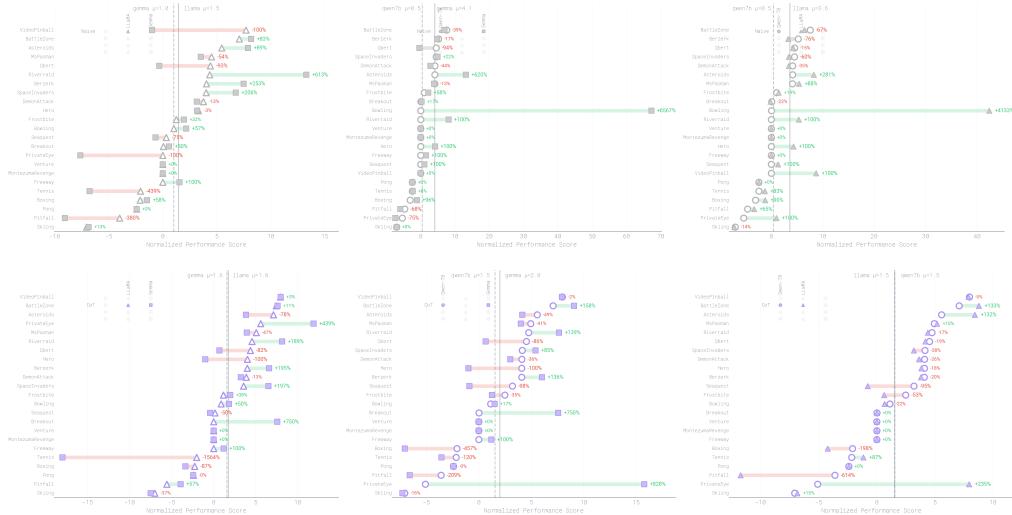


Figure 32: Cross-model performance comparison in the Basic scenario using Naive agent (top row) and Chain-of-Thought (CoT) agent (bottom row). Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Llama3.1-8B vs. Gemma-7B, (middle) Qwen2.5-7B vs. Gemma-7B, and (right) Qwen2.5-7B vs. Llama3.1-8B using the Naive agent. Bottom row: Same model comparisons using the CoT agent.

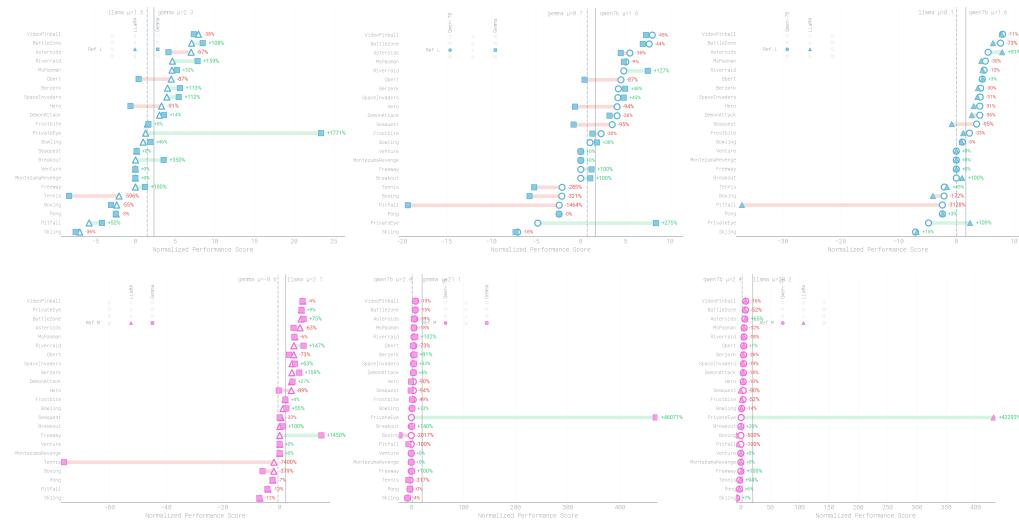


Figure 33: Cross-model performance comparison in the Basic scenario using Reflexion\\_last agent (top row) and Reflexion\\_max agent (bottom row). Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Llama3.1-8B vs. Gemma-7B, (middle) Qwen2.5-7B vs. Gemma-7B, and (right) Qwen2.5-7B vs. Llama3.1-8B using the Reflexion\\_last agent. Bottom row: Same model comparisons using the Reflexion\\_max agent.

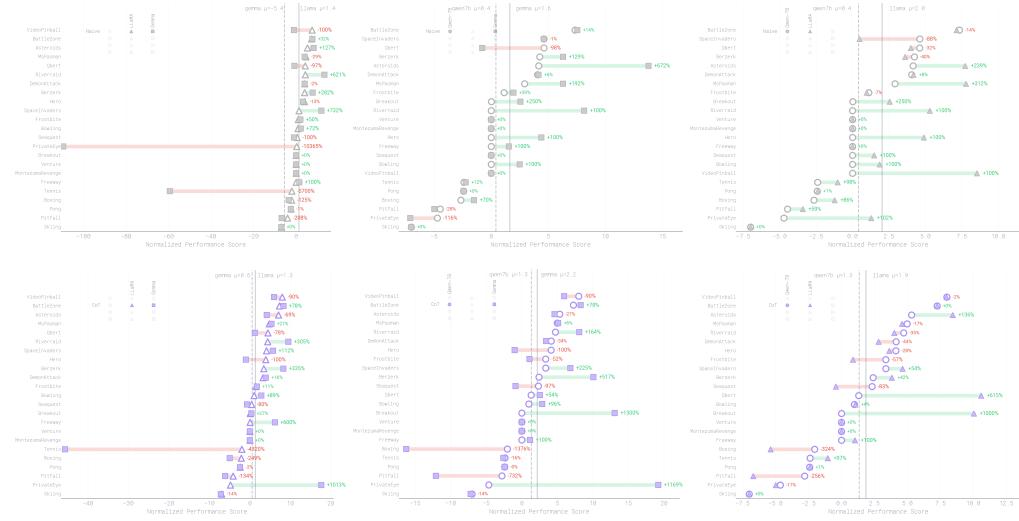


Figure 34: Cross-model performance comparison in the Obscured scenario using Naive agent (top row) and Chain-of-Thought (CoT) agent (bottom row). Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Llama3.1-8B vs. Gemma-7B, (middle) Qwen2.5-7B vs. Gemma-7B, and (right) Qwen2.5-7B vs. Llama3.1-8B using the Naive agent. Bottom row: Same model comparisons using the CoT agent.

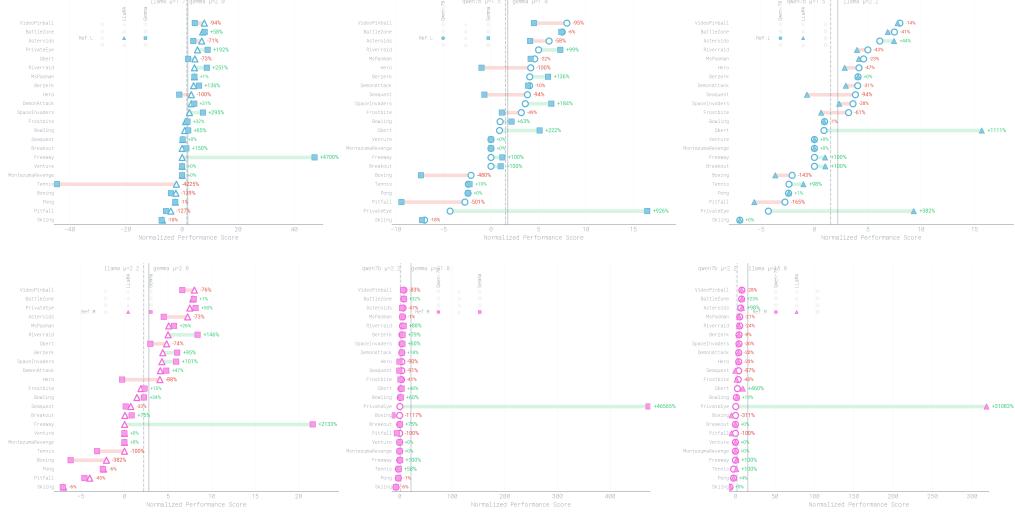


Figure 35: Cross-model performance comparison in the Obscured scenario using Reflexion\\_last agent (top row) and Reflexion\\_max agent (bottom row). Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Llama3.1-8B vs. Gemma-7B, (middle) Qwen2.5-7B vs. Gemma-7B, and (right) Qwen2.5-7B vs. Llama3.1-8B using the Reflexion\\_last agent. Bottom row: Same model comparisons using the Reflexion\\_max agent.

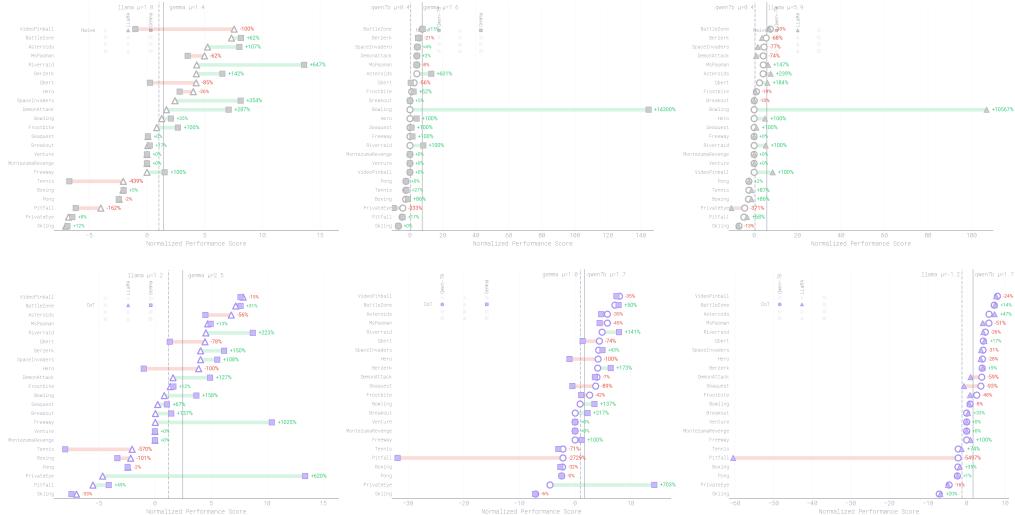


Figure 36: Cross-model performance comparison in the Game Manual scenario using Naive agent (top row) and Chain-of-Thought (CoT) agent (bottom row). Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Llama3.1-8B vs. Gemma-7B, (middle) Qwen2.5-7B vs. Gemma-7B, and (right) Qwen2.5-7B vs. Llama3.1-8B using the Naive agent. Bottom row: Same model comparisons using the CoT agent.

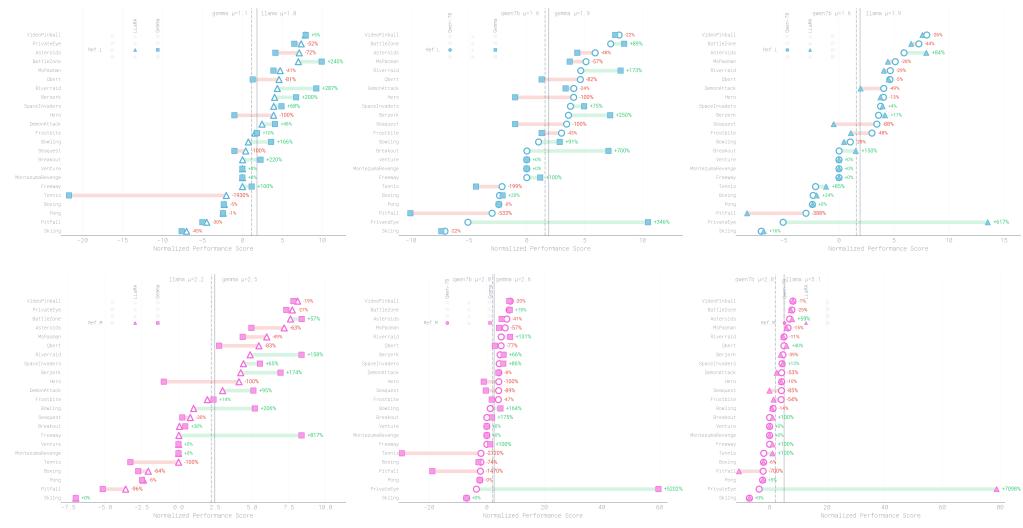


Figure 37: Cross-model performance comparison in the Game Manual scenario using `Reflexion_last` agent (top row) and `Reflexion_max` agent (bottom row). Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Llama3.1-8B vs. Gemma-7B, (middle) Qwen2.5-7B vs. Gemma-7B, and (right) Qwen2.5-7B vs. Llama3.1-8B using the `Reflexion_last` agent. Bottom row: Same model comparisons using the `Reflexion_max` agent.

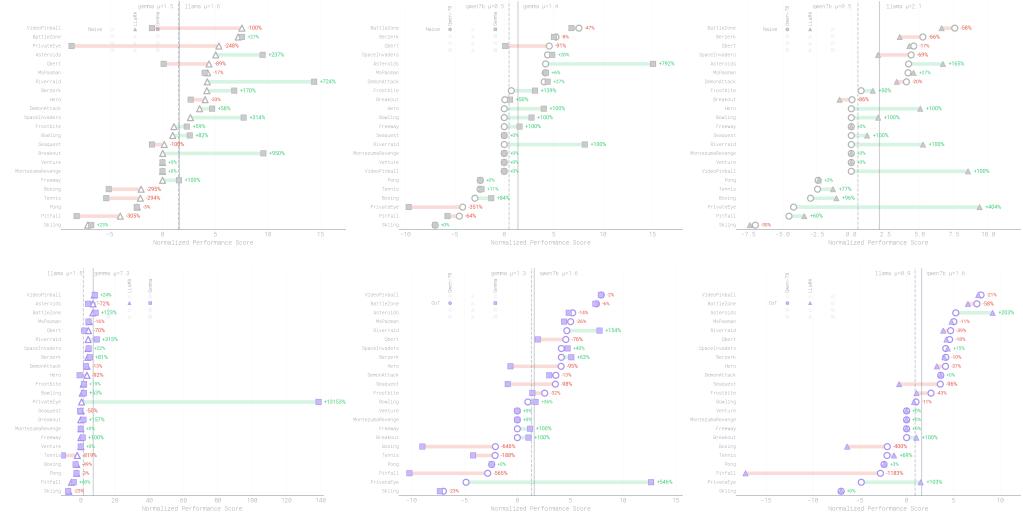


Figure 38: Cross-model performance comparison in the RL Trajectory scenario using `Naive` agent (top row) and `Chain-of-Thought` (CoT) agent (bottom row). Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Llama3.1-8B vs. Gemma-7B, (middle) Qwen2.5-7B vs. Gemma-7B, and (right) Qwen2.5-7B vs. Llama3.1-8B using the `Naive` agent. Bottom row: Same model comparisons using the `CoT` agent.

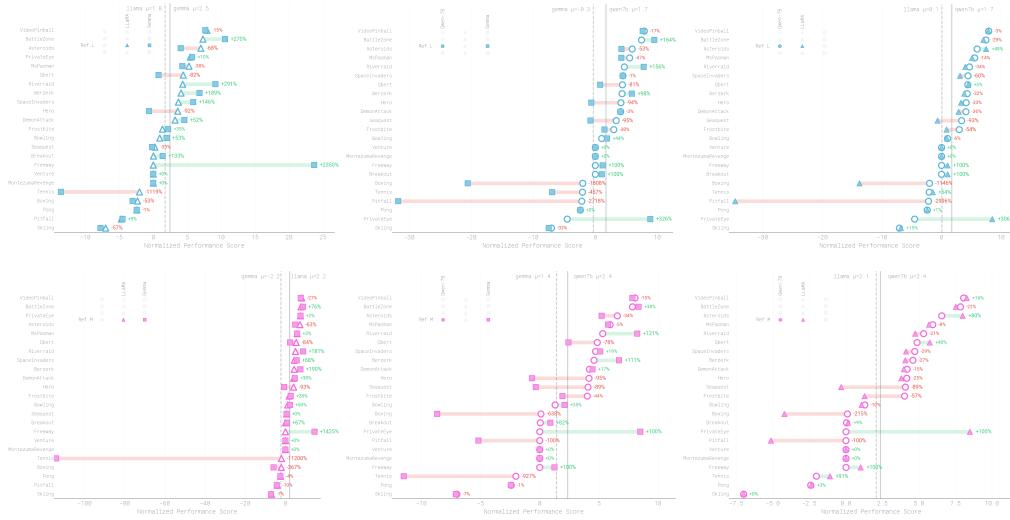


Figure 39: Cross-model performance comparison in the RL Trajectory scenario using Reflexion\_last agent (top row) and Reflexion\_max agent (bottom row). Each plot shows relative performance differences (bars) and normalized average scores (vertical lines) across all 23 Atari environments. Top row: Comparison between (left) Llama3.1-8B vs. Gemma7B, (middle) Qwen2.5-7B vs. Gemma7B, and (right) Qwen2.5-7B vs. Llama3.1-8B using the Reflexion\_last agent. Bottom row: Same model comparisons using the Reflexion\_max agent.

## Appendix References

- Abdulhai, M., White, I., Snell, C., Sun, C., Hong, J., Zhai, Y., Xu, K., and Levine, S. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models. *arXiv preprint arXiv:2311.18232*, 2023.
- Agashe, S., Fan, Y., Reyna, A., and Wang, X. E. Llm-coordination: evaluating and analyzing multi-agent coordination abilities in large language models. *arXiv preprint arXiv:2310.03903*, 2023.
- Alghamdi, E. A., Masoud, R. I., Alnuhait, D., Alomairi, A. Y., Ashraf, A., and Zaytoon, M. Aratrust: An evaluation of trustworthiness for llms in arabic. *CoRR*, 2024.
- Bailis, S., Friedhoff, J., and Chen, F. Werewolf arena: A case study in llm evaluation via social deduction. *arXiv preprint arXiv:2407.13943*, 2024.
- Bakhtin, A., Wu, D., Lerer, A., and Brown, N. No-press diplomacy from scratch. *Advances in Neural Information Processing Systems*, 34:18063–18074, 2021.
- Boisvert, L., Thakkar, M., Gasse, M., Caccia, M., de Chezelles, T., Cappart, Q., Chapados, N., Lacoste, A., and Drouin, A. Workarena++: Towards compositional planning and reasoning-based common knowledge work tasks. *Advances in Neural Information Processing Systems*, 37: 5996–6051, 2024.
- Bonatti, R., Zhao, D., Dupont, D., Abdali, S., Li, Y., Lu, Y., Wagle, J., Koishida, K., Bucker, A., Jang, L. K., et al. Windows agent arena: Evaluating multi-modal os agents at scale. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024.
- Cao, R., Lei, F., Wu, H., Chen, J., Fu, Y., Gao, H., Xiong, X., Zhang, H., Hu, W., Mao, Y., et al. Spider2-v: How far are multimodal agents from automating data science and engineering workflows? *Advances in Neural Information Processing Systems*, 37:107703–107744, 2024.
- Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.
- Carta, T., Romac, C., Wolf, T., Lamprier, S., Sigaud, O., and Oudeyer, P.-Y. Grounding large language models in interactive environments with online reinforcement learning. In *International Conference on Machine Learning*, pp. 3676–3713. PMLR, 2023.
- Chai, Y., Huang, S., Niu, Y., Xiao, H., Liu, L., Zhang, D., Gao, P., Ren, S., and Li, H. Amex: Android multi-annotation expo dataset for mobile gui agents. *CoRR*, 2024.
- Chai, Y., Li, H., Zhang, J., Liu, L., Liu, G., Wang, G., Ren, S., Huang, S., and Li, H. A3: Android agent arena for mobile gui agents. *arXiv preprint arXiv:2501.01149*, 2025.
- Chalamalasetti, K., Götze, J., Hakimov, S., Madureira, B., Sadler, P., and Schlangen, D. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11174–11219, 2023.
- Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., Starace, G., Liu, K., Maksin, L., Patwardhan, T., et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024.
- Chang, M., Chhablani, G., Clegg, A., Cote, M. D., Desai, R., Hlavac, M., Karashchuk, V., Krantz, J., Mottaghi, R., Parashar, P., et al. Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks. *arXiv preprint arXiv:2411.00081*, 2024.
- Chen, G., Chen, P., Hsieh, C.-Y., Lee, C.-K., Liao, B., Liao, R., Liu, W., Qiu, J., Sun, Q., Tang, J., et al. Alchemy: A quantum chemistry dataset for benchmarking ai models. *arXiv preprint arXiv:1906.09427*, 2019.

- Chen, J., Hu, X., Liu, S., Huang, S., Tu, W.-W., He, Z., and Wen, L. Llmarena: Assessing capabilities of large language models in dynamic multi-agent environments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13055–13077, 2024a.
- Chen, J., Yuen, D., Xie, B., Yang, Y., Chen, G., Wu, Z., Yixing, L., Zhou, X., Liu, W., Wang, S., et al. Spa-bench: A comprehensive benchmark for smartphone agent evaluation. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024b.
- Chen, Y., Yuan, Y., Zhang, Z., Zheng, Y., Liu, J., Ni, F., and Hao, J. Sheetagent: A generalist agent for spreadsheet reasoning and manipulation via large language models. In *ICML 2024 Workshop on LLMs and Cognition*, 2024c.
- Chen, Z., Chen, S., Ning, Y., Zhang, Q., Wang, B., Yu, B., Li, Y., Liao, Z., Wei, C., Lu, Z., et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024d.
- Chen, Z., Du, W., Zhang, W., Liu, K., Liu, J., Zheng, M., Zhuo, J., Zhang, S., Lin, D., Chen, K., et al. T-eval: Evaluating the tool utilization capability of large language models step by step. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9510–9529, 2024e.
- Chevalier-Boisvert, M., Dai, B., Towers, M., Perez-Vicente, R., Willem, L., Lahou, S., Pal, S., and Castro, P. S. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in Neural Information Processing Systems*, 36:73383–73394, 2023.
- Cloos, N., Jens, M., Naim, M., Kuo, Y.-L., Cases, I., Barbu, A., and Cueva, C. J. Baba is ai: Break the rules to beat the benchmark. In *ICML 2024 Workshop on LLMs and Cognition*, 2024.
- Costarelli, A., Allen, M., Hauksson, R., Sodunke, G., Hariharan, S., Cheng, C., Li, W., Clymer, J. M., and Yadav, A. Gamebench: Evaluating strategic reasoning abilities of llm agents. In *Language Gamification-NeurIPS 2024 Workshop*, 2024.
- Côté, M.-A., Kádár, A., Yuan, X., Kybartas, B., Barnes, T., Fine, E., Moore, J., Hausknecht, M., El Asri, L., Adada, M., et al. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*, pp. 41–75. Springer, 2019.
- Davidson, T. R., Veselovsky, V., Josifoski, M., Peyrard, M., Bosselut, A., Kosinski, M., and West, R. Evaluating language model agency through negotiations. In *ICLR 2024*, 2024.
- Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., and Su, Y. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.
- Dong, Y., Zhu, X., Pan, Z., Zhu, L., and Yang, Y. Villageragent: A graph-based multi-agent framework for coordinating complex task dependencies in minecraft. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 16290–16314, 2024.
- Drouin, A., Gasse, M., Caccia, M., Laradji, I. H., Del Verme, M., Marty, T., Vazquez, D., Chapados, N., and Lacoste, A. Workarena: How capable are web agents at solving common knowledge work tasks? In *International Conference on Machine Learning*, pp. 11642–11662. PMLR, 2024.
- Duan, J., Zhang, R., Diffenderfer, J., Kailkhura, B., Sun, L., Stengel-Eskin, E., Bansal, M., Chen, T., and Xu, K. Gtbench: Uncovering the strategic reasoning capabilities of llms via game-theoretic evaluations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- Gao, D., Ji, L., Bai, Z., Ouyang, M., Li, P., Mao, D., Wu, Q., Zhang, W., Wang, P., Guo, X., et al. Assistgui: Task-oriented desktop graphical user interface automation. *arXiv preprint arXiv:2312.13108*, 2023.

Gao, X., Gao, Q., Gong, R., Lin, K., Thattai, G., and Sukhatme, G. S. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056, 2022.

Gong, R., Huang, J., Zhao, Y., Geng, H., Gao, X., Wu, Q., Ai, W., Zhou, Z., Terzopoulos, D., Zhu, S.-C., et al. Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20483–20495, 2023.

Gong, R., Huang, Q., Ma, X., Noda, Y., Durante, Z., Zheng, Z., Terzopoulos, D., Fei-Fei, L., Gao, J., and Vo, H. Mindagent: Emergent gaming interaction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3154–3183, 2024.

Gonzalez-Pumariega, G., Yean, L. S., Sunkara, N., and Choudhury, S. Robotouille: An asynchronous planning benchmark for llm agents. *arXiv preprint arXiv:2502.05227*, 2025.

Guertler, L., Cheng, B., Yu, S., Liu, B., Choshen, L., and Tan, C. Textarena. *arXiv preprint arXiv:2504.11442*, 2025.

Guo, Z., Cheng, S., Niu, Y., Wang, H., Zhou, S., Huang, W., and Liu, Y. Stabletoolbench-mirrorapi: Modeling tool environments as mirrors of 7,000+ real-world apis. *arXiv preprint arXiv:2503.20527*, 2025.

Hafner, D. Benchmarking the spectrum of agent capabilities. In *International Conference on Learning Representations*, 2022.

He, H., Yao, W., Ma, K., Yu, W., Dai, Y., Zhang, H., Lan, Z., and Yu, D. Webvoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6864–6890, 2024.

Hill, W., Liu, I., Koch, A. D. M., Harvey, D., Kumar, N., Konidaris, G., and James, S. Mineplanner: A benchmark for long-horizon planning in large minecraft worlds. *arXiv preprint arXiv:2312.12891*, 2023.

Hu, L., Li, Q., Xie, A., Jiang, N., Stoica, I., Jin, H., and Zhang, H. Gamearena: Evaluating llm reasoning through live computer games. *arXiv preprint arXiv:2412.06394*, 2024.

Hua, W., Liu, O., Li, L., Amayuelas, A., Chen, J., Jiang, L., Jin, M., Fan, L., Sun, F., Wang, W., et al. Game-theoretic llm: Agent workflow for negotiation games. *arXiv preprint arXiv:2411.05990*, 2024.

Huang, Q., Vora, J., Liang, P., and Leskovec, J. Mlagentbench: Evaluating language agents on machine learning experimentation. In *International Conference on Machine Learning*, pp. 20271–20309. PMLR, 2024a.

Huang, S., Zhong, W., Lu, J., Zhu, Q., Gao, J., Liu, W., Hou, Y., Zeng, X., Wang, Y., Shang, L., et al. Planning, creation, usage: Benchmarking llms for comprehensive tool utilization in real-world complex scenarios. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 4363–4400, 2024b.

Huang, Y., Wang, X., Liu, H., Kong, F., Qin, A., Tang, M., Wang, X., Zhu, S.-C., Bi, M., Qi, S., et al. Adasociety: An adaptive environment with social structures for multi-agent decision-making. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024c.

James, S., Ma, Z., Arrojo, D. R., and Davison, A. J. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Jang, L. K., Li, Y., Ding, C., Lin, J., Liang, P. P., Zhao, D., Bonatti, R., and Koishida, K. Videowebarena: Evaluating long context multimodal agents with video understanding web tasks. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024.

- Jansen, P., Côté, M.-A., Khot, T., Bransom, E., Dalvi Mishra, B., Majumder, B. P., Tafjord, O., and Clark, P. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. *Advances in Neural Information Processing Systems*, 37:10088–10116, 2024.
- Jin, T., Zhu, Y., and Kang, D. Elt-bench: An end-to-end benchmark for evaluating ai agents on elt pipelines. *arXiv preprint arXiv:2504.04808*, 2025.
- Jin, X., Wang, Z., Du, Y., Fang, M., Zhang, H., and Wang, J. Learning to discuss strategically: A case study on one night ultimate werewolf. *Advances in Neural Information Processing Systems*, 37:77060–77097, 2024.
- Kapoor, R., Butala, Y. P., Russak, M., Koh, J. Y., Kamble, K., AlShikh, W., and Salakhutdinov, R. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In *European Conference on Computer Vision*, pp. 161–178. Springer, 2024.
- Karten, S., Nguyen, A. L., and Jin, C. Pok\`echamp: an expert-level minimax language agent. *arXiv preprint arXiv:2503.04094*, 2025.
- Klissarov, M., D’Oro, P., Sodhani, S., Raileanu, R., Bacon, P.-L., Vincent, P., Zhang, A., and Henaff, M. Motif: Intrinsic motivation from artificial intelligence feedback. In *Second Agent Learning in Open-Endedness Workshop*, 2023.
- Koh, J. Y., Lo, R., Jang, L., Duvvur, V., Lim, M., Huang, P.-Y., Neubig, G., Zhou, S., Salakhutdinov, R., and Fried, D. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 881–905, 2024.
- Küttler, H., Nardelli, N., Miller, A., Raileanu, R., Selvatici, M., Grefenstette, E., and Rocktäschel, T. The nethack learning environment. *Advances in Neural Information Processing Systems*, 33: 7671–7684, 2020.
- Lee, J., Min, T., An, M., Hahm, D., Lee, H., Kim, C., and Lee, K. Benchmarking mobile device control agents across diverse configurations. *arXiv preprint arXiv:2404.16660*, 2024.
- Lei, F., Chen, J., Ye, Y., Cao, R., Shin, D., Hongjin, S., SUO, Z., Gao, H., Hu, W., Yin, P., et al. Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Li, H., Su, J., Chen, Y., Li, Q., and ZHANG, Z.-X. Sheetcopilot: Bringing software productivity to the next level through large language models. *Advances in Neural Information Processing Systems*, 36:4952–4984, 2023a.
- Li, H., Cao, Y., Yu, Y., Javaji, S. R., Deng, Z., He, Y., Jiang, Y., Zhu, Z., Subbalakshmi, K., Xiong, G., et al. Investorbench: A benchmark for financial decision-making tasks with llm-based agent. *arXiv preprint arXiv:2412.18174*, 2024a.
- Li, M., Zhao, Y., Yu, B., Song, F., Li, H., Yu, H., Li, Z., Huang, F., and Li, Y. Api-bank: A comprehensive benchmark for tool-augmented llms. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b.
- Li, M., Zhao, S., Wang, Q., Wang, K., Zhou, Y., Srivastava, S., Gokmen, C., Lee, T., Li, E. L., Zhang, R., et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534, 2024b.
- Li, W., Bishop, W. E., Li, A., Rawles, C., Campbell-Ajala, F., Tyamagundlu, D., and Riva, O. On the effects of data scale on ui control agents. *Advances in Neural Information Processing Systems*, 37: 92130–92154, 2024c.
- Li, Y., He, J., Zhou, X., Zhang, Y., and Baldridge, J. Mapping natural language instructions to mobile ui action sequences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8198–8210, 2020.

- Light, J., Cai, M., Shen, S., and Hu, Z. Avalonbench: Evaluating llms playing the game of avalon. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.
- Lim, S., Kim, S., Yu, J., Lee, S., Chung, J., and Yu, Y. Visescape: A benchmark for evaluating exploration-driven decision-making in virtual escape rooms. *arXiv preprint arXiv:2503.14427*, 2025.
- Lin, F., Malfa, E. L., Hofmann, V., Yang, E. M., Cohn, A. G., and Pierrehumbert, J. B. Graph-enhanced large language models in asynchronous plan reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 30108–30134, 2024a.
- Lin, K. Q., Li, L., Gao, D., Wu, Q., Yan, M., Yang, Z., Wang, L., and Shou, M. Z. Videogui: A benchmark for gui automation from instructional videos. *CoRR*, 2024b.
- Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., and Stone, P. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791, 2023.
- Liu, E. Z., Guu, K., Pasupat, P., Shi, T., and Liang, P. Reinforcement learning on web interfaces using workflow-guided exploration. In *International Conference on Learning Representations*, 2018.
- Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al. Agentbench: Evaluating llms as agents. In *ICLR*, 2024a.
- Liu, X., Zhang, T., Gu, Y., Iong, I. L., Xu, Y., Song, X., Zhang, S., Lai, H., Liu, X., Zhao, H., et al. Visualagentbench: Towards large multimodal models as visual foundation agents. *CoRR*, 2024b.
- Lu, X. H., Kasner, Z., and Reddy, S. Weblinx: Real-world website navigation with multi-turn dialogue. In *International Conference on Machine Learning*, pp. 33007–33056. PMLR, 2024.
- Ma, C., Zhang, J., Zhu, Z., Yang, C., Yang, Y., Jin, Y., Lan, Z., Kong, L., and He, J. Agentboard: An analytical evaluation board of multi-turn llm agents. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a.
- Ma, Z., Huang, W., Zhang, J., Gupta, T., and Krishna, R. m & m's: A benchmark to evaluate tool-use for m ulti-step m ulti-modal tasks. In *European Conference on Computer Vision*, pp. 18–34. Springer, 2024b.
- Ma, Z., Zhang, B., Zhang, J., Yu, J., Zhang, X., Zhang, X., Luo, S., Wang, X., and Tang, J. Spreadsheetbench: Towards challenging real world spreadsheet manipulation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024c.
- Mandi, Z., Jain, S., and Song, S. Roco: Dialectic multi-robot collaboration with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 286–299. IEEE, 2024.
- Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., and Scialom, T. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2024.
- Mu, T., Ling, Z., Xiang, F., Yang, D., Li, X., Tao, S., Huang, Z., Jia, Z., and Su, H. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. In *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- Mukobi, G., Erlebach, H., Lauffer, N., Hammond, L., Chan, A., and Clifton, J. Welfare diplomacy: Benchmarking language model cooperation. In *Socially Responsible Language Modelling Research*, 2022.
- Nasir, M. U., James, S., and Togelius, J. Gametraversalbenchmark: Evaluating planning abilities of large language models through traversing 2d game maps. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- Nasiriany, S., Maddukuri, A., Zhang, L., Parikh, A., Lo, A., Joshi, A., Mandlekar, A., and Zhu, Y. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *RSS 2024 Workshop: Data Generation for Robotics*, 2024.

- Nathani, D., Madaan, L., Roberts, N., Bashlykov, N., Menon, A., Moens, V., Budhiraja, A., Magka, D., Vorotilov, V., Chaurasia, G., et al. Mlgym: A new framework and benchmark for advancing ai research agents. *arXiv preprint arXiv:2502.14499*, 2025.
- Padmakumar, A., Thomason, J., Shrivastava, A., Lange, P., Narayan-Chen, A., Gella, S., Piramuthu, R., Tur, G., and Hakkani-Tur, D. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.
- Pagliari, D., Cupiał, B., Coward, S., Piterbarg, U., Wolczyk, M., Khan, A., Pignatelli, E., Kuciński, Ł., Pinto, L., Fergus, R., et al. Balrog: Benchmarking agentic llm and vlm reasoning on games. *arXiv preprint arXiv:2411.13543*, 2024.
- Pan, Y., Kong, D., Zhou, S., Cui, C., Leng, Y., Jiang, B., Liu, H., Shang, Y., Zhou, S., Wu, T., et al. Webcanvus: Benchmarking web agents in online environments. In *Agentic Markets Workshop at ICML 2024*, 2024.
- Payan, J., Mishra, S., Singh, M., Negreanu, C. S., Poelitz, C., Baral, C., Roy, S., Chakravarthy, R., Van Durme, B., and Nouri, E. Instructexcel: A benchmark for natural language instruction in excel. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- Peng, Y., Li, S., Gu, W., Li, Y., Wang, W., Gao, C., and Lyu, M. R. Revisiting, benchmarking and exploring api recommendation: How far are we? *IEEE Transactions on Software Engineering*, 49(4):1876–1897, 2022.
- Pérez-Rodríguez, M., Hidalgo, M. J., Mendoza, A., González, L. T., Rodríguez, F. L., Goicoechea, H. C., and Pellerano, R. G. Measuring trace element fingerprinting for cereal bar authentication based on type and principal ingredient. *Food Chemistry*: X, 18:100744, 2023.
- Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., and Torralba, A. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8494–8502, 2018.
- Puig, X., Shu, T., Li, S., Wang, Z., Liao, Y.-H., Tenenbaum, J. B., Fidler, S., and Torralba, A. Watch-and-help: A challenge for social perception and human-ai collaboration. In *ICLR*, 2021.
- Qi, S., Chen, S., Li, Y., Kong, X., Wang, J., Yang, B., Wong, P., Zhong, Y., Zhang, X., Zhang, Z., et al. Civrealm: A learning and reasoning odyssey in civilization for decision-making agents. In *The Twelfth International Conference on Learning Representations*, 2024.
- Qian, C., Han, P., Luo, Q., He, B., Chen, X., Zhang, Y., Du, H., Yao, J., Yang, X., Zhang, D., et al. Escapebench: Pushing language models to think outside the box. *arXiv preprint arXiv:2412.13549*, 2024.
- Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *The Twelfth International Conference on Learning Representations*, 2024.
- Qu, C., Dai, S., Wei, X., Cai, H., Wang, S., Yin, D., Xu, J., and Wen, J.-R. Towards completeness-oriented tool retrieval for large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1930–1940, 2024.
- Rawles, C., Li, A., Rodriguez, D., Riva, O., and Lillicrap, T. Android in the wild: A large-scale dataset for android device control, 2023. URL <https://arxiv.org/abs/2307.10088>, 2023.
- Rawles, C., Clinckemaillie, S., Chang, Y., Waltz, J., Lau, G., Fair, M., Li, A., Bishop, W., Li, W., Campbell-Ajala, F., et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*, 2024.
- Rein, D., Becker, J., Deng, A., Nix, S., Canal, C., O’Connel, D., Arnott, P., Bloom, R., Broadley, T., Garcia, K., et al. Hcast: Human-calibrated autonomy software tasks. *arXiv preprint arXiv:2503.17354*, 2025.
- Ren, Y., Tertikas, K., Maiti, S., Han, J., Zhang, T., Süsstrunk, S., and Kokkinos, F. Vgrp-bench: Visual grid reasoning puzzle benchmark for large vision-language models. *arXiv preprint arXiv:2503.23064*, 2025.

- Ruan, K., Huang, M., Wen, J.-R., and Sun, H. Benchmarking llms' swarm intelligence. *arXiv preprint arXiv:2505.04364*, 2025.
- Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C. J., and Hashimoto, T. Identifying the risks of lm agents with an lm-emulated sandbox. In *The Twelfth International Conference on Learning Representations*, 2024.
- Samvelyan, M., Kirk, R., Kurin, V., Parker-Holder, J., Jiang, M., Hambro, E., Petroni, F., Kuttler, H., Grefenstette, E., and Rocktäschel, T. Minihack the planet: A sandbox for open-ended reinforcement learning research. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.
- Shen, H., Li, Y., Meng, D., Cai, D., Qi, S., Zhang, L., Xu, M., and Ma, Y. Shortcutsbench: A large-scale real-world benchmark for api-based agents. *CoRR*, 2024a.
- Shen, Y., Song, K., Tan, X., Zhang, W., Ren, K., Yuan, S., Lu, W., Li, D., and Zhuang, Y. Taskbench: Benchmarking large language models for task automation. *Advances in Neural Information Processing Systems*, 37:4540–4574, 2024b.
- Shi, T., Karpathy, A., Fan, L., Hernandez, J., and Liang, P. World of bits: An open-domain platform for web-based agents. In *International Conference on Machine Learning*, pp. 3135–3144. PMLR, 2017.
- Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., and Fox, D. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10740–10749, 2020a.
- Shridhar, M., Yuan, X., Cote, M.-A., Bisk, Y., Trischler, A., and Hausknecht, M. Alfworld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2020b.
- Song, Y., Thai, K., Pham, C. M., Chang, Y., Nadaf, M., and Iyyer, M. Bearcubs: A benchmark for computer-using web agents. *arXiv preprint arXiv:2503.07919*, 2025.
- Sun, H., Zhang, S., Ren, L., Xu, H., Fu, H., Yuan, C., and Wang, X. Collab-overcooked: Benchmarking and evaluating large language models as collaborative agents. *arXiv preprint arXiv:2502.20073*, 2025.
- Sun, L., Chen, X., Chen, L., Dai, T., Zhu, Z., and Yu, K. Meta-gui: Towards multi-modal conversational agents on mobile gui. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6699–6712, 2022.
- Tan, W., Ding, Z., Zhang, W., Li, B., Zhou, B., Yue, J., Xia, H., Jiang, J., Zheng, L., Xu, X., et al. Towards general computer control: A multimodal agent for red dead redemption ii as a case study. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.
- Tang, Q., Deng, Z., Lin, H., Han, X., Liang, Q., Cao, B., and Sun, L. Toolalpacca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023.
- Tang, W., Zhou, Y., Xu, E., Cheng, K., Li, M., and Xiao, L. Dsgbench: A diverse strategic game benchmark for evaluating llm-based agents in complex decision-making environments. *arXiv preprint arXiv:2503.06047*, 2025.
- Tang, X., Li, J., Liang, Y., Zhu, S.-C., Zhang, M., and Zheng, Z. Mars: Situated inductive reasoning in an open-world environment. *Advances in Neural Information Processing Systems*, 37:17830–17869, 2024.
- Thomas, G., Chan, A. J., Kang, J., Wu, W., Christianos, F., Greenlee, F., Toulis, A., and Purtorab, M. Webgames: Challenging general-purpose web-browsing ai agents. *arXiv preprint arXiv:2502.18356*, 2025.
- Trivedi, H., Khot, T., Hartmann, M., Manku, R., Dong, V., Li, E., Gupta, S., Sabharwal, A., and Balasubramanian, N. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, 2024.

- Urbanek, J., Fan, A., Karamcheti, S., Jain, S., Humeau, S., Dinan, E., Rocktäschel, T., Kiela, D., Szlam, A., and Weston, J. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 673–683, 2019.
- Valmeekam, K., Marquez, M., Olmo, A., Sreedharan, S., and Kambhampati, S. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36:38975–38987, 2023.
- Venkatesh, S. G., Talukdar, P., and Narayanan, S. Ugif: Ui grounded instruction following. *arXiv preprint arXiv:2211.07615*, 2022.
- Wan, Y., Mao, J., and Tenenbaum, J. Handmethat: Human-robot communication in physical and social environments. *Advances in Neural Information Processing Systems*, 35:12014–12026, 2022.
- Wang, J., Zerun, M., Li, Y., Zhang, S., Chen, C., Chen, K., and Le, X. Gta: a benchmark for general tool agents. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a.
- Wang, L., Deng, Y., Zha, Y., Mao, G., Wang, Q., Min, T., Chen, W., and Chen, S. Mobileagentbench: An efficient and user-friendly benchmark for mobile llm agents. *arXiv preprint arXiv:2406.08184*, 2024b.
- Wang, R., Jansen, P., Côté, M.-A., and Ammanabrolu, P. Scienceworld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022.
- Wang, W., Zhang, D., Feng, T., Wang, B., and Tang, J. Battleagentbench: A benchmark for evaluating cooperation and competition capabilities of language models in multi-agent systems. *arXiv preprint arXiv:2408.15971*, 2024c.
- Wang, X., Li, D., Zhao, Y., Wang, H., et al. Metatool: Facilitating large language models to master tools with meta-task augmentation. *CoRR*, 2024d.
- Wang, X., Zhuang, B., and Wu, Q. Are large vision language models good game players? *arXiv preprint arXiv:2503.02358*, 2025a.
- Wang, Y., Xian, Z., Chen, F., Wang, T.-H., Wang, Y., Fragkiadaki, K., Erickson, Z., Held, D., and Gan, C. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. In *International Conference on Machine Learning*, pp. 51936–51983. PMLR, 2024e.
- Wang, Y., Guo, Y., Zheng, Y., Yin, Z., Chen, S., Yang, J., Chen, J., Huang, X., and Qiu, X. Familytool: A multi-hop personalized tool use benchmark. *arXiv preprint arXiv:2504.06766*, 2025b.
- Wang, Z., Cui, Y., Zhong, L., Zhang, Z., Yin, D., Lin, B. Y., and Shang, J. Officebench: Benchmarking language agents across multiple applications for office automation. *CoRR*, 2024f.
- White, I., Nottingham, K., Maniar, A., Robinson, M., Lillemark, H., Maheshwari, M., Qin, L., and Ammanabrolu, P. Collaborating action by action: A multi-agent llm framework for embodied reasoning. *arXiv preprint arXiv:2504.17950*, 2025.
- Wijk, H., Lin, T., Becker, J., Jawhar, S., Parikh, N., Broadley, T., Chan, L., Chen, M., Clymer, J., Dhyani, J., et al. Re-bench: Evaluating frontier ai r&d capabilities of language model agents against human experts. *arXiv preprint arXiv:2411.15114*, 2024.
- Wornow, M., Narayan, A., Viggiano, B., Khare, I., Verma, T., Thompson, T., Hernandez, M., Sundar, S., Trujillo, C., Chawla, K., et al. Wonderbread: A benchmark for evaluating multimodal foundation models on business process management tasks. *Advances in Neural Information Processing Systems*, 37:115963–116021, 2024.
- Wu, D., Shi, H., Sun, Z., and Liu, B. Deciphering digital detectives: Understanding llm behaviors and capabilities in multi-agent mystery games. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 8225–8291, 2024a.

- Wu, Y., Tang, X., Mitchell, T., and Li, Y. Smartplay: A benchmark for llms as intelligent agents. In *The Twelfth International Conference on Learning Representations*, 2024b.
- Xi, Z., Ding, Y., Chen, W., Hong, B., Guo, H., Wang, J., Yang, D., Liao, C., Guo, X., He, W., et al. Agentgym: Evolving large language model-based agents across diverse environments. *CoRR*, 2024.
- Xie, J., Zhang, R., Chen, Z., Wan, X., and Li, G. Whodunitbench: Evaluating large multimodal agents via murder mystery games. *Advances in Neural Information Processing Systems*, 37:86655–86687, 2024a.
- Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T. J., Cheng, Z., Shin, D., Lei, F., et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024b.
- Xing, M., Zhang, R., Xue, H., Chen, Q., Yang, F., and Xiao, Z. Understanding the weakness of large language model agents within a complex android environment. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6061–6072, 2024.
- Xu, F. F., Song, Y., Li, B., Tang, Y., Jain, K., Bao, M., Wang, Z. Z., Zhou, X., Guo, Z., Cao, M., et al. Theagentcompany: benchmarking llm agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*, 2024a.
- Xu, K., Kordi, Y., Nayak, T., Asija, A., Wang, Y., Sanders, K., Byerly, A., Zhang, J., Van Durme, B., and Khashabi, D. Tur [k] ingbench: A challenge benchmark for web agents. *arXiv preprint arXiv:2403.11905*, 2024b.
- Xu, L., Hu, Z., Zhou, D., Ren, H., Dong, Z., Keutzer, K., Ng, S. K., and Feng, J. Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7315–7332, 2024c.
- Xu, M., Jiang, G., Liang, W., Zhang, C., and Zhu, Y. Interactive visual reasoning under uncertainty. *Advances in Neural Information Processing Systems*, 36:42409–42432, 2023a.
- Xu, Q., Hong, F., Li, B., Hu, C., Chen, Z., and Zhang, J. On the tool manipulation capability of open-sourced large language models. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023b.
- Xu, Y., Liu, X., Sun, X., Cheng, S., Yu, H., Lai, H., Zhang, S., Zhang, D., Tang, J., and Dong, Y. Androidlab: Training and systematic benchmarking of android autonomous agents. *arXiv preprint arXiv:2410.24024*, 2024d.
- Yang, J., Prabhakar, A., Narasimhan, K., and Yao, S. Intercode: Standardizing and benchmarking interactive coding with execution feedback. *Advances in Neural Information Processing Systems*, 36:23826–23854, 2023.
- Yang, S., Zhao, B., and Xie, C. Aqa-bench: An interactive benchmark for evaluating llms' sequential reasoning ability. *CoRR*, 2024.
- Yao, J., Wang, K., Hsieh, R., Zhou, H., Zou, T., Cheng, Z., Wang, Z., and Viswanath, P. Spin-bench: How well do llms plan strategically and reason socially? *arXiv preprint arXiv:2503.12349*, 2025.
- Yao, S., Chen, H., Yang, J., and Narasimhan, K. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.
- Yao, S., Shinn, N., Razavi, P., and Narasimhan, K.  $\tau$ -bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.
- Yu, P., Yang, Y., Li, J., Zhang, Z., Wang, H., Feng, X., and Zhang, F. Multi-mission tool bench: Assessing the robustness of llm based agents through related and dynamic missions. *arXiv preprint arXiv:2504.02623*, 2025.

- Yu, X., Fu, J., Deng, R., and Han, W. Mineland: Simulating large-scale multi-agent interactions with limited multimodal senses and physical needs. *CoRR*, 2024.
- Yuan, X., Moss, M. M., Feghali, C. E., Singh, C., Moldavskaya, D., MacPhee, D., Caccia, L., Pereira, M., Kim, M., Sordoni, A., et al. debug-gym: A text-based environment for interactive debugging. *arXiv preprint arXiv:2503.21557*, 2025.
- Zhang, D., Shen, Z., Xie, R., Zhang, S., Xie, T., Zhao, Z., Chen, S., Chen, L., Xu, H., Cao, R., et al. Mobile-env: Building qualified evaluation benchmarks for llm-gui interaction. *arXiv preprint arXiv:2305.08144*, 2023.
- Zhang, H., Guo, H., Guo, S., Cao, M., Huang, W., Liu, J., and Zhang, G. Ing-vp: Mllms cannot play easy vision-based games yet. *arXiv preprint arXiv:2410.06555*, 2024a.
- Zhang, L., Wang, S., Jia, X., Zheng, Z., Yan, Y., Gao, L., Li, Y., and Xu, M. Llamatouch: A faithful and scalable testbed for mobile ui task automation. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–13, 2024b.
- Zhang, S., Xu, Z., Liu, P., Yu, X., Li, Y., Gao, Q., Fei, Z., Yin, Z., Wu, Z., Jiang, Y.-G., et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. *arXiv preprint arXiv:2412.18194*, 2024c.
- Zheng, L., Huang, Z., Xue, Z., Wang, X., An, B., and Shuicheng, Y. Agentstudio: A toolkit for building general virtual agents. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024.
- Zheng, X., Li, L., Yang, Z., Yu, P., Wang, A. J., Yan, R., Yao, Y., and Wang, L. V-mage: A game evaluation framework for assessing visual-centric capabilities in multimodal large language models. *arXiv preprint arXiv:2504.06148*, 2025.
- Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., et al. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024a.
- Zhou, X., Zhu, H., Mathur, L., Zhang, R., Yu, H., Qi, Z., Morency, L.-P., Bisk, Y., Fried, D., Neubig, G., et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024b.
- Zhuang, Y., Yu, Y., Wang, K., Sun, H., and Zhang, C. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36:50117–50143, 2023.