EvoEmpirBench: Dynamic Spatial Reasoning with Agent-ExpVer

Pukun Zhao[1,*], Longxiang Wang[2,*], Miaowei Wang[3,*], Chen Chen[1], Fanqing Zhou[1], Haojian Huang[4,5†]

[1]Guangdong University of Finance and Economics,
[2]Chongqing University,
[3]The University of Edinburgh,
[4]Hong Kong University of Science and Technology (Guangzhou),
[5]The University of Hong Kong
{zhaopukun, Allen821}@student.gdufe.edu.cn, 20223610@stu.cqu.edu.cn,
M.Wang-123@sms.ed.ac.uk, zhoyfanqing@gmail.com, haojianhuang@connect.hku.hk

Abstract

Most existing spatial reasoning benchmarks focus on static or globally observable environments, failing to capture the challenges of long-horizon reasoning and memory utilization under partial observability and dynamic changes. We introduce two dynamic spatial benchmarks—locally observable maze navigation and match-2 elimination—that systematically evaluate models' abilities in spatial understanding and adaptive planning when local perception, environment feedback, and global objectives are tightly coupled. Each action triggers structural changes in the environment, requiring continuous update of cognition and strategy. We further propose a subjective experience-based memory mechanism for cross-task experience transfer and validation. Experiments show that our benchmarks reveal key limitations of mainstream models in dynamic spatial reasoning and long-term memory, providing a comprehensive platform for future methodological advances.

Introduction

Over the past few years, Large Language Models (LLMs)(Achiam et al. 2023; Touvron et al. 2023) have achieved impressive results across a wide range of static natural language benchmarks—including machine translation(Xu et al. 2024; Zhu et al. 2023; Enis and Hopkins 2024), question answering (Yao et al. 2023; Schick et al. 2023; Madaan et al. 2023), and code generation (Hong et al. 2023; Wu et al. 2023a)—by leveraging massive pretraining corpora and ever-larger context windows.

However, evaluating genuine reasoning and adaptability in LLMs, especially in dynamic, interactive scenarios, remains a significant challenge. Existing benchmarks for reasoning (Hao et al. 2024; Valmeekam et al. 2023; Saparov and He 2022; Hong et al. 2023; Wu et al. 2023a; Cobbe et al. 2021) predominantly rely on static datasets, rendering them vulnerable to data contamination, where models overfit to training patterns and are prone to rapid performance saturation due to their limited scope (Sainz et al. 2023; White et al. 2024; Kiela et al. 2021). While dynamic human-in-the-loop evaluations such as Chatbot Arena (Chiang et al. 2024) bring fresh perspectives, they tend to capture subjective user preferences rather than objective reasoning skills (Hu et al. 2024; Li, Angelopoulos, and Chiang 2024). Furthermore, Chatbot Arena struggles to evaluate specific reasoning capabilities, such as inductive and deductive reasoning, due to its reliance on unconstrained user inputs and the lack of precise control over the prompts' reasoning demands.

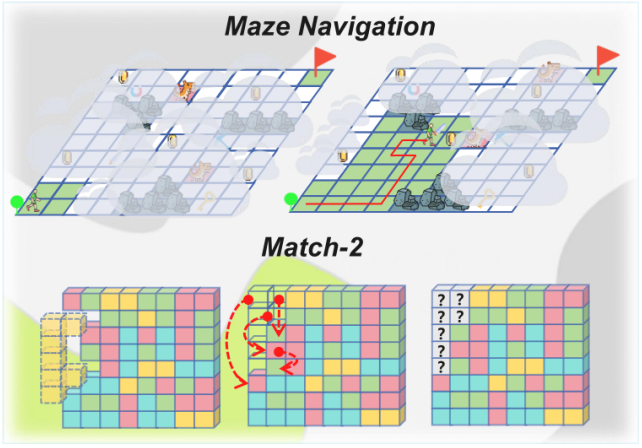

Figure 1: Overview of EvoEmpirBench: Locally observable maze navigation (up) and match-2 elimination (down), illustrating dynamic and interactive challenges for language agents.

Recent work (Chiang et al. 2024; Zheng et al. 2023; Zhao et al. 2024; Wu et al. 2023b; Leng et al. 2025a; Hu et al. 2024; Zhang et al. 2024b; Chen et al. 2024b) has begun exploring dynamic interactive tasks, but with limited generality. Many LLM-based agents excel only in narrowly defined domains and depend heavily on handcrafted prompts (Zhang et al. 2024b). For example, the Agent-Pro framework (Zhang et al. 2024b) enables agent learning in games such as Blackjack and Texas Hold'em, but these settings feature relatively shallow interaction structures and limited information complexity, hindering transfer to more realistic and complex

*Equal contribution
†Corresponding author.

tasks (Zhang et al. 2024b), such as navigation in unstructured environments (Kober, Bagnell, and Peters 2013) or real-time resource management (Mao et al. 2016). To address these deficits, we introduce a suite of dynamic and interactive benchmark environments that more faithfully capture the complexity and uncertainty of real-world reasoning problems. Our benchmarks include a maze game and a match-2 game that require multi-step spatial reasoning, tool use, strategic planning, and adaptive behavior under partial observability. Unlike prior work, our environments feature dynamic elements—such as discoverable tools or obstacles, score-based resource management, and non-stationary information structures—demanding agents to continually gather information, revise plans, and generalize learned strategies.

Given these dynamic settings, traditional static "collect-then-repeat" training paradigms (Mnih et al. 2015; Haarnoja et al. 2018) are ill-suited. In contrast, human learning involves continual cycles of abstraction and rule induction, where experiences are retrieved, refined, and integrated into transferable knowledge as new situations arise (Brown et al. 2020). Inspired by this, we propose a cognitively grounded online learning framework with three collaborating agents: the *GeoLink Agent* actively engages with the environment, the *InsightForce Agent* abstracts experiences and distills transferable rules ("truths"), and the *TruthWeaver Agent* manages these truths by merging, de-duplicating, and integrating new rules into structured memory. This architecture enables continual, adaptive learning without requiring offline data collection or post-deployment model updates. **Our contributions are as follows:**

- **Dynamic Reasoning Benchmark:** We present interactive, partially observable environments (maze and match-2 games) designed to rigorously test spatial, linguistic, and adaptive reasoning beyond static and narrow-domain evaluations.
- **Cognitively-inspired Online Learning:** We propose a novel human-inspired learning framework with memory abstraction and rule distillation, enabling continual, parameter-free adaptation and lifelong learning in dynamic environments.

Related Work

Benchmarks for LLM Reasoning. The assessment of LLM reasoning has evolved from static to dynamic benchmarks. Early static benchmarks—covering logical reasoning (Hao et al. 2024; Valmeekam et al. 2023; Saparov and He 2022), coding (Hong et al. 2023; Wu et al. 2023a), and mathematics (Cobbe et al. 2021)—probed LLMs using fixed datasets but struggled with data contamination (Sainz et al. 2023; White et al. 2024) and performance saturation (Kiela et al. 2021). To address these limitations, dynamic evaluation approaches have emerged. Methods involving human or LLM judges (Chiang et al. 2024; Zheng et al. 2023; Zhao et al. 2024) assess open-ended responses, but can introduce subjective biases (Chen et al. 2024a; Li, Angelopoulos, and Chiang 2024). Game-based benchmarks provide a more robust and interactive evaluation of real-world problem-solving (see Table 1). SmartPlay (Wu et al. 2023b) and CrossWordBench (Leng et al. 2025a) leverage static environments to study spatial and multimodal reasoning, but lack dynamic, real-time interaction. GameArena (Hu et al. 2024) incorporates human-LLM play in games like Akinator and Bluffing, yet does not provide fine-grained control over environmental dynamics or specific reasoning skills. Dynamic settings such as Agent-Pro (Zhang et al. 2024b), focused on game theory in Blackjack and Poker, and AutoManual (Chen et al. 2024b), which measures task completion, are limited by the scope and diversity of scenarios. In contrast, our benchmark introduces fully dynamic, interactive game environments that challenge LLMs with multi-dimensional reasoning tasks—including spatial analysis, language understanding, and adaptive decision-making—under uncertainty. Continuous exploration and active agent-environment interaction better reflect real-world complexity, requiring long-horizon reasoning across local and global objectives.

Continual Learning Mechanisms for LLM-based Agents. Continual learning (CL)(Wang et al. 2024) enables LLM-based intelligent agents to incrementally adapt to new tasks while minimizing catastrophic forgetting(Kirkpatrick et al. 2017). Recent work on CL for LLM-based agents (Ma et al. 2021; Zhang et al. 2024b; Chen et al. 2024b; Ghunaim et al. 2023; Kim, Seo, and Choi 2024) emphasizes memory management and online learning, drawing from cognitive principles. *Memory management* retains knowledge across tasks. Experience replay methods store past interactions for periodic retraining (Rolnick et al. 2019). CLIN (Majumder et al. 2023) updates memory with causal abstractions, while Generative Agents (Park et al. 2023) retrieve experiences based on recency and relevance. MemGPT (Packer et al. 2023) enables LLMs to manage working and long-term memory. However, these approaches target static domains, limiting scalability in interactive environments. *Online learning* supports real-time adaptation (Zhang, Lu, and Zhou 2018). Reflexion (Shinn et al. 2023), AdaPlanner (Sun et al. 2023), and ReAct (Yao et al. 2023) enable LLM agents to adjust planning based on feedback. RAP (Kagaya et al. 2024) uses retrieved experiences for task-switching decisions. Voyager (Wang et al. 2023) lets LLM agents store verified programs as reusable skills. These methods underperform in long-horizon tasks and lack autonomous policy evolution mechanisms. *Limitations remain*: Most frameworks for LLMs lack support for real-time, autonomous policy evolution in interactive environments (Feng et al. 2025). AutoManual (Chen et al. 2024b) introduces dynamic rules for adaptation, while Agent-Pro (Zhang et al. 2024b) refines strategies via policy-level reflection. Both are domain-specific and lack generalizability. Existing methods seldom allow parameter-free, real-time adaptation post-deployment (Zhang et al. 2024a). *Our approach* addresses these gaps by proposing a human-inspired online learning workflow (see section) that enables LLM agents to abstract and update knowledge through continuous interaction—supporting dynamic, real-time, and parameter-free continual adaptation.

Benchmark	Types / Trials	Reasoning Focus	Dynamic	Env Interaction	Real-World Complexity	Observability
SmartPlay (Wu et al. 2023b)	6 / 180	Spatial & Planning	No	Yes†	Low	Full
GameArena (Hu et al. 2024)	3 / 2000+	Deductive & Inductive	No	No	Medium	Full
CrossWordBench (Leng et al. 2025a)	1 / 350	Multimodal	No	No	Low	Full
Agent-Pro (Zhang et al. 2024b)	2 / 187	Game-Theoretic	Yes	Yes‡	Medium	Full
AutoManual (Chen et al. 2024b)	2 / 800+	Task Completion	Yes	Yes‡	Medium	Full
Ours	**2 / 180**	**Spatial & GLH**	**Yes**	**Yes**‡	**High**	**PO**

Table 1: Comparison of Game-Based Benchmarks for LLM Reasoning.†Agent perceives and is influenced by the environment, but actions do not modify it; ‡Agent perceives and modifies the environment, resulting in mutual influence; PO: Partially Observable; GLH: Global Long-Horizon.

Games in EvoEmpirBench

EvoEmpirBench (EEB) consists of two dynamic, interactive game environments designed to comprehensively evaluate LLM reasoning in complex, partially observable settings: a locally observable maze navigation task and a match-2 elimination game (Figure 1). Both games simulate real-world uncertainty and agent-environment dynamics, demanding skills in spatial analysis, language understanding, and adaptive planning. Each game provides 3 progressively difficult levels (Easy, Medium, Hard) for both training and testing, totaling 120 diverse task episodes. EEB supports both human and agent operation, enabling flexible evaluation under different paradigms. We now outline the core rules, difficulty scaling, and evaluation metrics.

Game Rules

Maze Navigation. In this 9×9 grid world, agents act under local observability—exploring to reveal new tiles while aiming to reach a goal with high score. Main objectives include collecting five gold coins and exploring unknown areas efficiently. As difficulty increases, agents encounter new elements: Easy levels feature a simple map; Medium introduces two moving monsters that threaten agent health; Hard adds four interactive items—Pickaxe (breaks obstacles), Iron Sword (unlimited monster combat), Magnet (area coin collection), and Key (needed for exit)—along with the existing challenges. Many agent actions, such as obtaining weapons or destroying obstacles, alter the environment dynamically, requiring adaptive, multi-step reasoning.

Match-2 Elimination. Agents operate on an 8×8 board, eliminating connected blocks of identical color (A–D) by matching two or more, with exponentially increasing rewards for larger matches. The primary goal is to meet per-color elimination targets within a limited number of moves, striving to maximize overall score and efficiency. Following each elimination, remaining blocks settle downward and new blocks spawn at the top. Agents may use step-limited props (row/col clear, bomb, hammer) at point costs to optimize gameplay. Difficulty levels vary step constraints and target requirements.

Difficulty Settings

In EvoEmpirBench, difficulty settings increase challenge by scaling environment complexity and resource constraints. For Maze Navigation, the difficulty progresses across three levels on a 9×9 grid. The Easy level includes no monsters and 5 coins as items. The Medium level introduces 2 moving monsters while maintaining 5 coins. The Hard level retains the 2 moving monsters but adds 4 additional items alongside the 5 coins, pushing planning and adaptation skills. For Match-2 Elimination, difficulty also spans three levels on an 8×8 grid. The Easy level allows 15–18 steps with a target of 8–12 eliminations per color. The Medium level tightens to 12–15 steps and increases the target to 12–16 eliminations per color. The Hard level further restricts steps to 10–13 and raises the target to 16–20 eliminations per color, requiring increasingly efficient strategies.

Evaluation Metrics

Performance on each game in EvoEmpirBench is assessed using complementary metrics that reflect both general task success and environment-specific skills. For both the Maze Navigation and Match-2 Elimination games, key metrics include Success Rate (*Suc.Rate*), which measures the episode completion ratio, and Average Score (*A.Score*), which represents the mean cumulative reward. In the Maze Navigation environment, additional metrics evaluate navigation efficiency and survival capabilities: Average Steps (*A.steps*) tracks the average steps per episode, Average Exploration (*A.Explor*) measures the explored map percentage, Average Gold Collected (*A.Gold*) quantifies the collected gold ratio, Remaining HP (*Rem.HP*) indicates surviving health points, Average Enemy Kills (*A.kills*) counts defeated enemies, and Average Barriers Destroyed (*A.Barr.*) measures destroyed barriers. For the Match-2 Elimination game, metrics focus on step efficiency and API optimization: Remaining/Max Steps Ratio (*R/M.S*) evaluates the remaining steps relative to the maximum, Score per Step (*Score/Step*) measures reward per move, Blocks Cleared per Step (*Clear/Step*) tracks cleared blocks per move, and API Efficiency (*API Eff.*) assesses the valid API call ratio. These metrics collectively provide a thorough evaluation of agent capabilities across the distinct task requirements of both games.

Methods

To address the limitations of static training paradigms in dynamic, partially observable environments, we present a human-inspired online learning framework: **Agent-ExpVer** (*Experience+Verification*), with its workflow illustrated in Figure 2. In contrast to classic reinforcement learn-

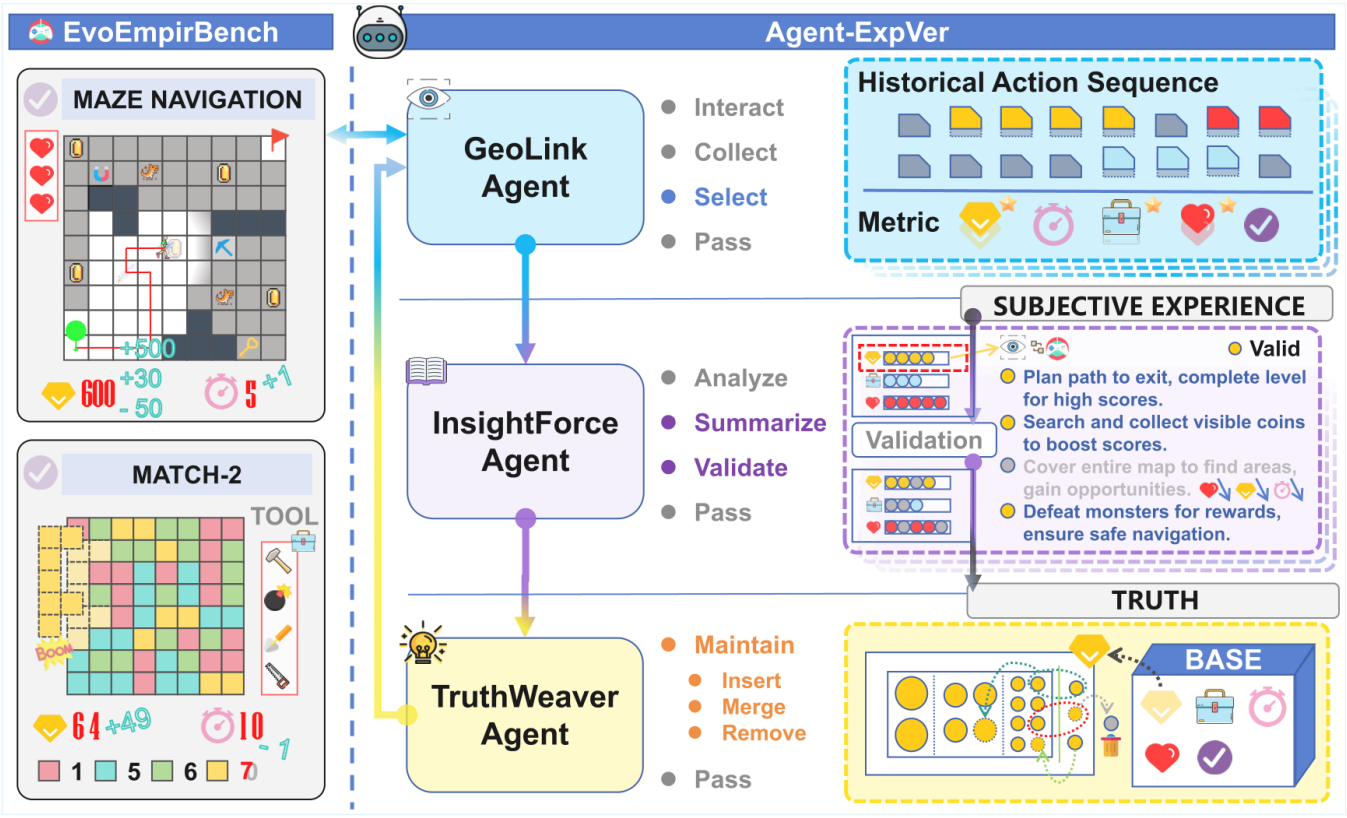

Figure 2: Workflow of the Agent-ExpVer System. The left side showcases EvoEmpirBench (EEB), our dynamic benchmark. The right side presents the Agent-ExpVer framework, comprising three agents: the GeoLink Agent collects and selects key historical actions based on game highlight metrics; the InsightForce Agent summarizes and validates subjective experiences; the TruthWeaver Agent maintains truths (insert, merge, remove) and passes them back to the GeoLink Agent. The figure highlights the processes of selection, summarization, validation, and maintenance.

ing (Zhang et al. 2021, 2022)—which relies on large-scale pre-collected data and can struggle to generalize—Agent-ExpVer leverages in-context learning (Dong et al. 2022) and continual experience abstraction (Zhang et al. 2024b). The agent's experiences are systematically organized, verified, and merged to build a reusable knowledge base that supports online adaptation.

Active Environment Interaction

Human learners actively reduce environmental uncertainty through exploration. Agent-ExpVer mimics this with an environment interaction agent called *GeoLink Agent* that iteratively interacts with the game environment.

At each timestep t, the agent observes the current state $\mathbf{s}_t$ (partially observable), chooses an action a_t based on its policy π_t, and returns the subsequent state $\mathbf{s}_{t+1}$ and reward r_t:

$$a_t \sim \pi_t(\mathbf{s}_t), \tag{1}$$

$$(\mathbf{s}_{t+1}, r_t) \sim \mathcal{E}(\mathbf{s}_t, a_t), \tag{2}$$

where $\mathcal{E}$ represents the environment dynamics. Across each episode, the agent accumulates an interaction history $\mathcal{H}_{0:T} = \{(\mathbf{s}_0, a_0, r_0), \ldots, (\mathbf{s}_T, a_T, r_T)\}$. This history is subsequently used for experience abstraction.

Subjective Experience Abstraction and Truth Distillation

After an episode, Agent-ExpVer employs a *InsightForce agent*— inspired by human episodic memory—to process the collected trajectory $\mathcal{H}_{0:T}$ and final metrics $\mathbf{m}$ (A.Score, Suc.Rate, etc.).

The InsightForce agent summarizes key decisions and outcomes as a subjective experience $\mathbf{e}$:

$$\mathbf{e} = f_{\text{sum}}(\mathcal{H}_{0:T}, \mathbf{m}) \tag{3}$$

where f_{sum} is executed by prompting the LLM. Experiences are added to a memory module $\mathcal{M}_{\text{exp}}$.

To verify and refine this knowledge, the agent then replays the episode, integrating $\mathbf{e}$ as a prompt extension to its policy. If the level is passed and resulting score improves, $\mathbf{e}$ is upgraded to a reusable *truth*, stored in truth repository $\mathcal{M}_{\text{truth}}$, managed by the *TruthWeaver agent*:

$$\mathcal{M}_{\text{truth}} \leftarrow \mathcal{M}_{\text{truth}} \cup \mathbf{e} \quad \text{if} \quad P \wedge (S' > S) \tag{4}$$

where P indicates whether the level is passed (*true* if passed, *false* otherwise), and $S' > S$ indicates an improved score.

The TruthWeaver Agent manages the incoming truths through the following operations: (1) merging incoming

Model	Sample	S.R. ↑	A.S. ↑	A.St. ↓	A.Ex. ↑	A.G. ↑	R.HP ↑	A.K. ↑	A.B. ↑
Human Baseline									
human	90	90.00	2914.67	20.6	77.54	68.22	2.43	0.83	1.43
Open-Source Models									
deepseek-V3	90	61.11	1649.78	50.61	83.15	75.56	1.71	0.77	0.80
llama-3.1-8b-instruct	90	23.33	-1213.67	54.42	51.00	24.89	1.28	0.03	0.27
llama-3.1-70b-instruct	90	31.11	489.89	39.19	58.30	51.11	0.87	0.30	0.23
qwen2.5-7b-instruct	90	3.33	-1071.67	62.74	65.45	51.56	0.53	0.27	0.60
qwen2.5-14b-instruct	90	36.67	24.67	53.69	65.31	55.56	1.30	0.27	0.23
qwen2.5-32b-instruct	90	42.22	1122.22	38.41	68.43	60.44	1.12	0.23	0.50
qwen2.5-32b-instruct-ours	90	54.44	1532.33	38.81	68.73	61.33	1.18	0.26	0.5
qwen2.5-72b-instruct	90	34.44	368.67	54.76	75.11	67.33	1.34	0.30	0.43
qwen3-30b-a3b	90	65.56	2359.22	33.10	81.93	73.33	1.73	0.43	0.57
Proprietary Models									
grok-3	90	67.78	2893.89	29.66	83.47	72.67	1.70	0.93	1.03
gemini-2.0-flash	90	28.89	-627.44	67.51	65.42	60.22	1.54	0.30	0.43
gemini-2.5-flash-preview	90	45.56	1426.11	48.30	83.37	74.44	1.14	0.77	0.80
gemini-2.5-flash-preview-ours	90	64.44	2158.67	36.88	80.81	71.33	1.68	0.50	0.70
claude-3-5-sonnet	90	50.00	1664.00	42.44	81.70	73.56	1.34	0.57	0.67
claude-3-7-sonnet	90	68.89	2793.44	27.40	82.59	73.33	1.70	0.47	0.77
claude-3-7-sonnet-ours	90	72.22	2858	31.33	83.35	76.44	1.78	0.73	0.83
gpt-4	90	43.33	1623.33	32.59	73.11	64.44	1.12	0.33	0.57
gpt-4.1	90	73.33	2562.33	34.03	83.63	74.89	1.98	0.47	0.67
gpt-4.1-ours	90	78.89	2805.67	32.77	83.69	72.67	1.99	0.73	0.77

Table 2: All Levels Metrics Across Models — Maze Navigation Game Performance. Metric abbreviations: S.R. (Success Rate), A.S. (Avg. Score), A.St. (Avg. Steps), A.Ex. (Avg. Exploration), A.G. (Gold Collection Rate), R.HP (Remaining HP), A.K. (Avg. Kills), A.B. (Avg. Barrier Interactions). Top-1 values in each column are underlined.

truths that have high semantic similarity with existing truths, (2) removing incoming truths that are redundant with existing truths, and (3) inserting new truths.

Policy Evolution via Truth Integration

The current agent policy is constructed by combining the base prompt π_0 with all truths $\mathcal{M}_{\text{truth}}$:

$$\pi_t = \pi_0 \cup \bigcup_{\mathbf{e} \in \mathcal{M}_{\text{truth}}} \mathbf{e} \tag{5}$$

After each episode, new truths are added and validated as above, enabling continual improvement as the system accumulates and distills knowledge.

We define policy improvement as the average test score increment:

$$\Delta = \frac{1}{N} \sum_{i=1}^{N} (S_i^t - S_i^{t-1}), \tag{6}$$

where N is the number of test cases, S_i^t and S_i^{t-1} are scores using the current and previous policies, respectively. If policy performance degrades ($\Delta < 0$), the newest updates are reverted and abstraction is repeated.

This workflow supports continual, online learning, allowing Agent-ExpVer to bridge the gap between static benchmark settings and the adaptive problem-solving required in dynamic, real-world scenarios.

Experiment

Experiment Setting. We evaluated both proprietary and open-source models on our EvoEmpirBench (EEB). Proprietary models include GPT-4 (Achiam et al. 2023), GPT-4.1, Gemini-2.0-flash, Gemini-2.5-flash-preview (Leng et al. 2025b), Claude-3-5-sonnet, Claude-3-7-sonnet, and Grok-3. Open-source models comprise Deepseek-V3 (Liu et al. 2024), Llama-3.1-8B-instruct, Llama-3.1-70B-instruct (Dubey et al. 2024), Qwen2.5-7B-instruct, Qwen2.5-14B-instruct, Qwen2.5-32B-instruct, Qwen2.5-72B-instruct, and Qwen3-30B-a3B (Team 2025). EEB features two tasks—maze navigation and match-2 games—each at three difficulty levels with 30 instances per level (90 instances per task). We also applied our Agent-ExpVer workflow to GPT-4.1, Gemini-2.5-flash-preview, Claude-3-7-sonnet, and Qwen2.5-32B-instruct, observing marked improvements in their EEB performance. Finally, by benchmarking both models and human participants, we provide a clear view of current LLM capabilities and limitations in dynamic reasoning.

Maze Navigation Game. As shown in Table 2, applying **Agent-ExpVer** to both open-source and proprietary models boosts their Maze Navigation performance by an average of $+5.6\%$ in success rate and $+9.5\%$ in mean score. In particular, **GPT-4.1** achieves a $+5.56\%$ (absolute) / $+7.6\%$ (relative) increase in success and a $+243.3$–point / $+9.5\%$ rise in score, making it the closest to human-level performance. Among open-source models, **Qwen-32B-instruct** with Agent-ExpVer posts $+12.22\%$ / $+28.96\%$ success and $+410.1$–point / $+36.6\%$ score gains, surpassing nearly all peers and rivaling Deepseek-V3 and Qwen-30B-A3B. The comprehensive metrics in EvoEmpirBench effectively evaluate model reasoning capabilities in complex environments. Our analysis reveals that stronger models achieve higher completion rates with fewer steps, while maintaining greater residual health points and demonstrating superior performance in coin collection, monster elimination, and obstacle destruction. These metrics collectively assess spatial reasoning in dynamic environments, risk-reward optimization, long-term versus short-term reward trade-offs, and interactive proficiency with virtual ecosystems.

Model	Sample	S.R. ↑	A.S. ↑	R/M.S ↑	S./St. ↑	C./St. ↑	API E. ↑
Human Baseline							
human	90	86.67	350.2	22.99	34.40	6.20	–
Open-Source Models							
deepseek-V3	90	37.78	218.29	8.99	17.07	4.92	44.55
llama-3.1-8b-instruct	90	22.22	95.4	4.18	9.05	4.33	4.95
llama-3.1-70b-instruct	90	30.00	289.67	4.94	21.59	4.65	32.82
qwen2.5-7b-instruct	90	17.78	85.2	3.61	8.09	3.42	2.19
qwen2.5-14b-instruct	90	36.67	170.91	9.38	13.42	4.86	37.86
qwen2.5-32b-instruct	90	33.33	203.07	8.80	15.60	5.11	34.29
qwen2.5-32b-instruct-ours	90	41.11	197.42	8.93	15.52	5.30	57.76
qwen2.5-72b-instruct	90	35.56	194.36	7.06	14.86	4.88	38.87
qwen3-30b-a3b	90	38.89	266.30	6.43	20.31	4.91	54.43
Proprietary Models							
grok-3	90	42.22	201.11	11.04	15.87	5.08	35.43
gemini-2.0-flash	90	34.44	202.24	8.52	15.76	5.12	49.19
gemini-2.5-flash-preview	90	37.78	410.76	5.57	30.58	4.85	69.73
gemini-2.5-flash-preview-ours	90	37.78	415.49	7.29	32.33	5.04	88.85
claude-3-5-sonnet	90	26.67	342.02	4.88	25.26	4.57	53.27
claude-3-7-sonnet	90	41.11	298.96	7.76	23.04	4.92	50.49
claude-3-7-sonnet-ours	90	47.19	291.01	17.49	26.70	5.64	88.06
gpt-4	90	31.11	142.27	5.26	27.80	4.43	36.67
gpt-4.1	90	40.00	245.04	7.52	18.67	5.03	46.12
gpt-4.1-ours	90	53.33	234.6	17.31	19.90	5.86	75.72

Table 3: All Levels Metrics Across Models — Match-2 Elimination Game Performance. Metric abbreviations: S.R. (Success Rate), A.S. (Avg. Score), R/M.S (Redundant/Missing Steps), S./St. (Score per Step), C./St. (Clearance per Step), API E. (API Efficiency). Top-1 values (excluding human baseline) in each column are underlined.

Match-2 Elimination Game. Table 3 highlights the greater challenge of the Match-2 task, where baseline LLMs average only 33.7% success. Open-source contenders Deepseek-V3 (37.78%) and Qwen-30B-a3b (38.89%), and proprietary models Grok-3 (42.22%), Claude-3-7-sonnet (41.11%), and GPT-4.1 (40.00%), all underperform humans by a wide margin. Integrating Agent-ExpVer yields consistent gains of $+13.3\%$ in success rate, $+9.8\%$ more steps remaining, $+1.2$ elimination score per step, and $+0.8$ eliminations per step. Notably, GPT-4.1 achieves $+13.33\%$ success—becoming the top non-human agent—though its average score dips by 4.3% (–10.44 points). Conversely, Gemini-2.5-flash-preview maintains its success rate while gaining $+1.2\%$ in score (+4.73 points), even surpassing human benchmarks. This underscores EEB's emphasis on compound objectives: agents must judiciously trade immediate elimination efficiency (via in-game items like bombs) against long-term score maximization.

Ablation Studies

Is organizing and managing a truth repository important? To assess the importance of managing the truth repository, we conduct experiments on the EEB benchmark, comparing the full Agent-ExpVer framework against a variant without the TruthWeaver Agent, using Qwen2.5-32B and GPT-4.1 as bases. As reported in Table 4, removing TruthWeaver causes a marked drop in both tasks: for Qwen2.5-32B, average scores fall by 16.9% and success rates decline by 6.1%. Moreover, reasoning efficiency suffers—maze completion requires 3.5% more steps, and remaining steps in Match-2 decrease by 4.0%. For GPT-4.1, average scores fall by 5.8% and success rates decline by 8.3%; maze completion requires 0.9% more steps, while remaining steps in Match-2 decrease by 26.8%. These results confirm that TruthWeaver is essential for (1) mitigating catastrophic forgetting via dynamic truth consolidation and (2) stabilizing belief distributions to avoid redundant hypotheses.

How much do EEB's constraints affect model performance? To evaluate, we conduct ablation studies on both games. As reported in Table 5, in the Maze Game, removing exploration via global visibility (Full-Vision) leads to a notable performance boost: GPT-4.1's success rate rises from 73.33% (EEB) to 93.33%, and its average score increases from 2562.33 to 3412.22. Qwen2.5 shows a similar trend, with success rate improving from 42.22% to 57.78%. Both models also require fewer steps under Full-Vision, indicating reduced reasoning complexity.In the Match-2 Game, prohibiting tool use (NoProps) slightly improves scores for both models, but the gains are less pronounced than in Maze. For example, GPT-4.1's success rate increases from 40.00% (EEB) to 43.33%, and Qwen2.5's score rises from 203.07 to 411.62.These results demonstrate that EEB's design—through partial observability and tool-based reasoning—significantly increases task difficulty and reasoning demands, as evidenced by consistent performance drops compared to control variants.

Further Analysis

What truth did the model learn with Agent-ExpVer? With Agent-ExpVer, models like *gpt-4.1-ours* and *claude-3-7-sonnet-ours* in the Maze Game showed greater risk awareness—achieving higher remaining health points (1.99 and 1.78) but slightly reduced exploration and gold collection (See Appendix). This suggests they internalized that survival is key for long-term success, sometimes favoring cau-

		Maze Game			Match-2 Game		
Model	**Sample**	**S.R. ↑**	**A.S. ↑**	**A.St. ↓**	**S.R. ↑**	**A.S. ↑**	**R/M.S ↓**
GPT-4.1	90	73.33	2562.33	34.03	40.00	245.04	7.52
GPT-4.1 w/o TW	90	77.78	2765.33	33.06	48.89	219.53	12.66
GPT-4.1-ours	90	78.89	2805.67	32.77	53.33	234.60	17.31
Qwen2.5-32B	90	42.22	1122.22	38.41	33.33	203.07	8.80
Qwen2.5-32B w/o TW	90	51.11	1272.11	37.07	35.56	187.89	8.57
Qwen2.5-32B-ours	90	54.44	1532.33	35.81	41.57	197.42	8.93

Table 4: Cross-Game Ablation Study of the TruthWeaver Framework. Metrics for the Maze Game (Cols 3–5) and Match-2 Game (Cols 6–8).

Configuration	**Success (%)**	**Score**	**Steps**
Maze Game			
GPT-4.1 Maze EEB	73.33	2562.33	34.03
GPT-4.1 Maze Full-Vis	93.33	3412.22	19.53
Qwen2.5 Maze EEB	42.22	1122.22	38.41
Qwen2.5 Maze Full-Vis	57.78	1737.78	18.47
Match-2 Game			
GPT-4.1 Match-2 EEB	40.00	245.04	7.52
GPT-4.1 Match-2 NoProps	43.33	412.67	10.04
Qwen2.5 Match-2 EEB	33.33	203.07	8.80
Qwen2.5 Match-2 NoProps	32.22	411.62	6.24

Table 5: Model Performance Comparison on Maze and Match-2 Games. Metrics based on 90 samples per model. Abbreviations: Full-Vis = Full Visual Information, NoProps = No Propositional Information.

tion over immediate rewards. At the same time, increases in average kills and barrier interactions indicate more proactive engagement. In Match-2, models such as *gpt-4.1-ours* and *gemini-2.5-flash-preview-ours* reached higher success rates (53.33% and 41.11%) post-training, though sometimes with lower average scores. This trade-off reflects a shift toward strategies that prioritize completion, supported by gains in clearance per step and API efficiency, and demonstrates more efficient, deliberate actions. The evolution of these "truths" is further illustrated in Figure 3. In the early rounds of truth induction, both Maze and Match-2 models experienced a temporary drop in success rate and score (e.g., Maze success rate fell from 66.67% to 56.67%, and Match-2 from 6.67% to 0%), likely due to the immaturity and overgeneralization of initial hypotheses. For instance, a model might infer that "bold exploration yields high rewards" after a lucky episode, only to encounter negative outcomes when this risky strategy is applied indiscriminately. Notably, the improvement observed in the Maze task can be attributed to the model's ability to learn survival-oriented principles through the ExpVer module—details of which are provided in Appendix. However, after 3–4 rounds of iterative learning, both success rate and score steadily improved (Maze: success rate rose to 76.66%, score to 3370; Match-2: success rate to 16.67%, score to 258), reflecting the model's ability to refine its truths—shifting from reckless exploration to more cautious, stepwise progress. Besides, Figure 4 reveals that models trained with Agent-ExpVer not only improved in aggregate metrics but also became more efficient in their execution. The distribution of steps required for successful completion narrowed and shifted toward lower values for

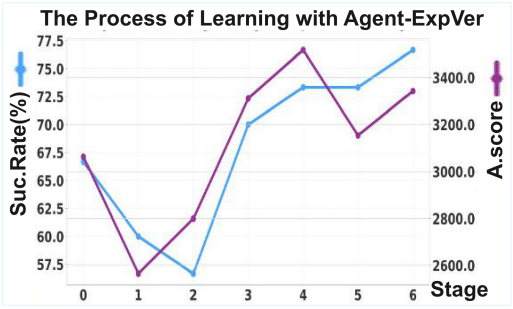

Figure 3: A.Score and Suc.Rate Trends of the Model Across Learning Episodes of Maze Navigation(See Match-2 in appendix)

both *gpt-4.1-ours* and *gemini-2.5-flash-preview-ours*, indicating enhanced planning and decision-making efficiency. In Maze, this translates to faster, safer navigation with reduced exposure to hazards, while in Match-2, it results in higher clearance rates and scores per step.

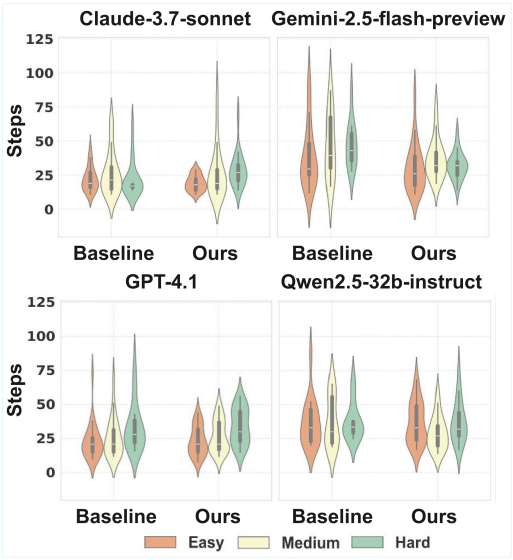

Figure 4: Distribution of steps required to complete the maze navigation game before and after learning

Conclusion

We introduce EvoEmpirBench, a benchmark for spatial and high-level reasoning in dynamic, interactive environments, featuring Maze Navigation and Match-2 tasks. We also present Agent-ExpVer, a three-agent framework for environment interaction, experience synthesis, and adaptive truth management; experiments show it drives effective online learning and markedly improves agent reasoning and interactivity. However, performance remains tied to model capacity, with even top systems falling short of human baselines. Future work will primarily enhance Agent-ExpVer's adaptability and expand EvoEmpirBench with more complex reasoning tasks, alongside refining experience management.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, G. H.; Chen, S.; Liu, Z.; Jiang, F.; and Wang, B. 2024a. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.

Chen, M.; Li, Y.; Yang, Y.; Yu, S.; Lin, B.; and He, X. 2024b. Automanual: Constructing instruction manuals by llm agents via interactive environmental learning. *Advances in Neural Information Processing Systems*, 37: 589–631.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhu, B.; Zhang, H.; Jordan, M.; Gonzalez, J. E.; et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Liu, T.; et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Enis, M.; and Hopkins, M. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *arXiv preprint arXiv:2404.13813*.

Feng, T.; Wang, X.; Zhou, Z.; Wang, R.; Zhan, Y.; Li, G.; Li, Q.; and Zhu, W. 2025. EvoAgent: Agent Autonomous Evolution with Continual World Model for Long-Horizon Tasks. *arXiv preprint arXiv:2502.05907*.

Ghunaim, Y.; Bibi, A.; Alhamoud, K.; Alfarra, M.; Al Kader Hammoud, H. A.; Prabhu, A.; Torr, P. H.; and Ghanem, B. 2023. Real-time evaluation in online continual learning: A new hope. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11888–11897.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. Pmlr.

Hao, S.; Gu, Y.; Luo, H.; Liu, T.; Shao, X.; Wang, X.; Xie, S.; Ma, H.; Samavedhi, A.; Gao, Q.; et al. 2024. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*.

Hong, S.; Zheng, X.; Chen, J.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; Zhou, L.; et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 3(4): 6.

Hu, L.; Li, Q.; Xie, A.; Jiang, N.; Stoica, I.; Jin, H.; and Zhang, H. 2024. GameArena: Evaluating LLM Reasoning through Live Computer Games. *arXiv preprint arXiv:2412.06394*.

Kagaya, T.; Yuan, T. J.; Lou, Y.; Karlekar, J.; Pranata, S.; Kinose, A.; Oguri, K.; Wick, F.; and You, Y. 2024. Rap: Retrieval-augmented planning with contextual memory for multimodal llm agents. *arXiv preprint arXiv:2402.03610*.

Kiela, D.; Bartolo, M.; Nie, Y.; Kaushik, D.; Geiger, A.; Wu, Z.; Vidgen, B.; Prasad, G.; Singh, A.; Ringshia, P.; et al. 2021. Dynabench: Rethinking benchmarking in NLP. *arXiv preprint arXiv:2104.14337*.

Kim, B.; Seo, M.; and Choi, J. 2024. Online continual learning for interactive instruction following agents. *arXiv preprint arXiv:2403.07548*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Kober, J.; Bagnell, J. A.; and Peters, J. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11): 1238–1274.

Leng, J.; Huang, C.; Huang, L.; Lin, B. Y.; Cohen, W. W.; Wang, H.; and Huang, J. 2025a. CrossWordBench: Evaluating the Reasoning Capabilities of LLMs and LVLMs with Controllable Puzzle Generation. *arXiv preprint arXiv:2504.00043*.

Leng, J.; Huang, C.; Huang, L.; Lin, B. Y.; Cohen, W. W.; Wang, H.; and Huang, J. 2025b. CrossWordBench: Evaluating the Reasoning Capabilities of LLMs and LVLMs with Controllable Puzzle Generation. arXiv:2504.00043.

Li, T.; Angelopoulos, A.; and Chiang, W.-L. 2024. Does style matter? disentangling style and substance in chatbot arena, August 2024a. *URL https://blog. lmarena. ai/blog/2024/style-control*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ma, H.; Sun, Y.; Li, J.; Tomizuka, M.; and Choi, C. 2021. Continual multi-agent interaction behavior prediction with conditional generative memory. *IEEE Robotics and Automation Letters*, 6(4): 8410–8417.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Welleck, S.; Majumder, B. P.; Gupta, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-Refine: Iterative Refinement with Self-Feedback. *ArXiv*, abs/2303.17651.

Majumder, B. P.; Mishra, B. D.; Jansen, P.; Tafjord, O.; Tandon, N.; Zhang, L.; Callison-Burch, C.; and Clark, P. 2023. Clin: A continually learning language agent for rapid task adaptation and generalization. *arXiv preprint arXiv:2310.10134*.

Mao, H.; Alizadeh, M.; Menache, I.; and Kandula, S. 2016. Resource management with deep reinforcement learning. In *Proceedings of the 15th ACM workshop on hot topics in networks*, 50–56.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533.

Packer, C.; Fang, V.; Patil, S.; Lin, K.; Wooders, S.; and Gonzalez, J. 2023. MemGPT: Towards LLMs as Operating Systems.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Rolnick, D.; Ahuja, A.; Schwarz, J.; Lillicrap, T.; and Wayne, G. 2019. Experience replay for continual learning. *Advances in neural information processing systems*, 32.

Sainz, O.; Campos, J. A.; García-Ferrero, I.; Etxaniz, J.; de Lacalle, O. L.; and Agirre, E. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. *arXiv preprint arXiv:2310.18018*.

Saparov, A.; and He, H. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *ArXiv*, abs/2302.04761.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 8634–8652.

Sun, H.; Zhuang, Y.; Kong, L.; Dai, B.; and Zhang, C. 2023. Adaplanner: Adaptive planning from feedback with language models. *Advances in neural information processing systems*, 36: 58202–58245.

Team, Q. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Valmeekam, K.; Marquez, M.; Olmo, A.; Sreedharan, S.; and Kambhampati, S. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36: 38975–38987.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Wang, L.; Zhang, X.; Su, H.; and Zhu, J. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

White, C.; Dooley, S.; Roberts, M.; Pal, A.; Feuer, B.; Jain, S.; Shwartz-Ziv, R.; Jain, N.; Saifullah, K.; Naidu, S.; et al. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. 2023a. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Wu, Y.; Tang, X.; Mitchell, T. M.; and Li, Y. 2023b. Smartplay: A benchmark for llms as intelligent agents. *arXiv preprint arXiv:2310.01557*.

Xu, H.; Sharaf, A.; Chen, Y.; Tan, W.; Shen, L.; Van Durme, B.; Murray, K.; and Kim, Y. J. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Zhang, H.; Lei, Y.; Gui, L.; Yang, M.; He, Y.; Wang, H.; and Xu, R. 2024a. Cppo: Continual learning for reinforcement learning with human feedback. In *The Twelfth International Conference on Learning Representations*.

Zhang, L.; Lu, S.; and Zhou, Z.-H. 2018. Adaptive online learning in dynamic environments. *Advances in neural information processing systems*, 31.

Zhang, W.; Tang, K.; Wu, H.; Wang, M.; Shen, Y.; Hou, G.; Tan, Z.; Li, P.; Zhuang, Y.; and Lu, W. 2024b. Agent-pro: Learning to evolve via policy-level reflection and optimization. *arXiv preprint arXiv:2402.17574*.

Zhang, W.; Zhao, K.; Li, P.; Zhu, X.; Shen, Y.; Ma, Y.; Chen, Y.; and Lu, W. 2022. A closed-loop perception, decision-making and reasoning mechanism for human-like navigation. *arXiv preprint arXiv:2207.11901*.

Zhang, W.; Zhao, K.; Li, P.; Zhu, X.; Ye, F.; Jiang, W.; Fu, H.; and Wang, T. 2021. Learning to navigate in a vuca environment: Hierarchical multi-expert approach. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9254–9261. IEEE.

Zhao, R.; Zhang, W.; Chia, Y. K.; Xu, W.; Zhao, D.; and Bing, L. 2024. Auto-Arena: Automating LLM Evaluations with Agent Peer Battles and Committee Discussions. *arXiv preprint arXiv:2405.20267*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zhu, W.; Liu, H.; Dong, Q.; Xu, J.; Huang, S.; Kong, L.; Chen, J.; and Li, L. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.