Dual-scale World Memory for LLM Agents Towards Hard-exploration Problems

Minsoo Kim, Seung-won Hwang
minsoo9574@gmail.com

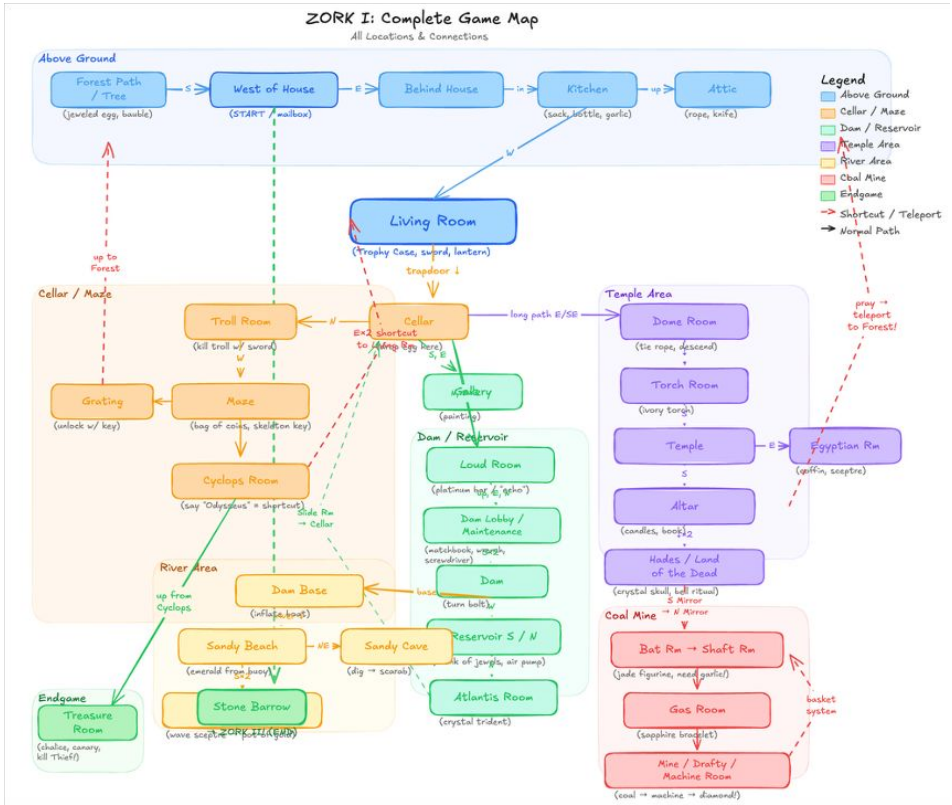
Motivation: Hard Exploration

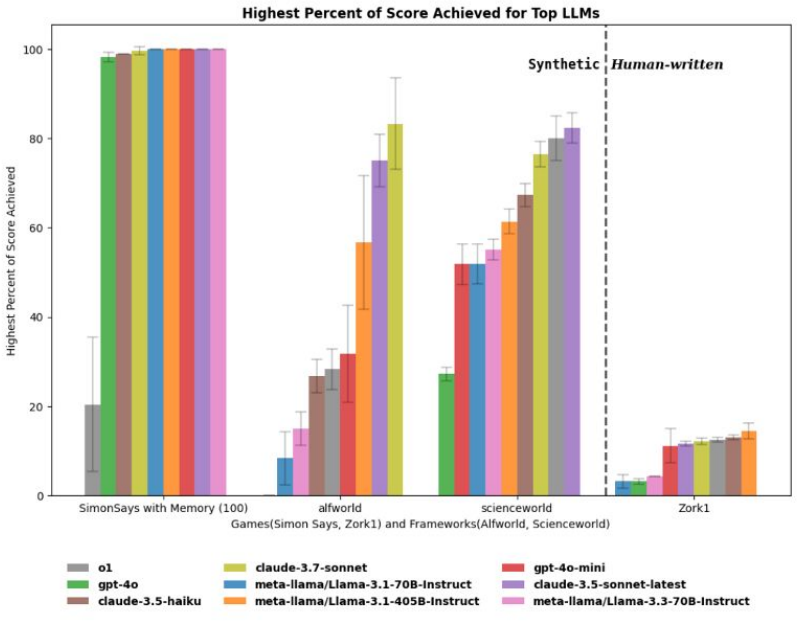
- Hard-exploration problems are characterized by:
  - Large & complex state–action spaces
  - Sparse rewards with delayed feedback
  - Deceptive local optima that trap naive exploration

Atari games with more immediate reward signals, simple action spaces

Hard Exploration: Montezuma's Revenge

Motivation: Exploration Problem in LLM Agents

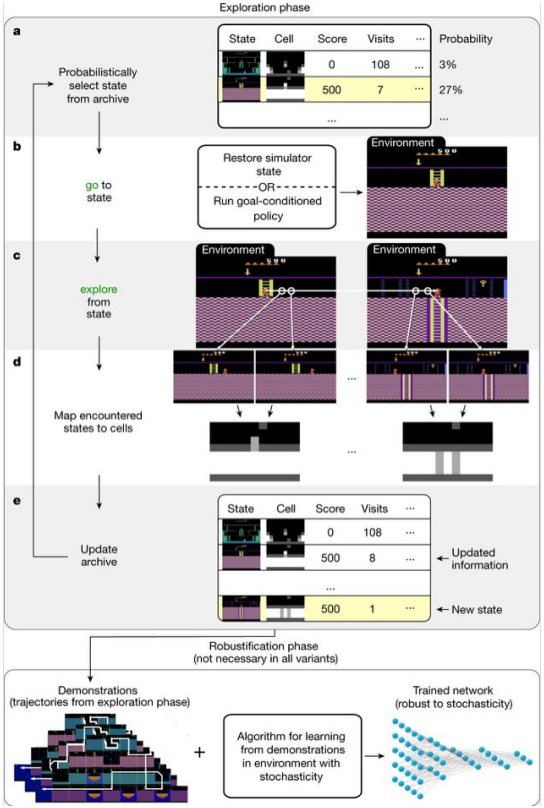


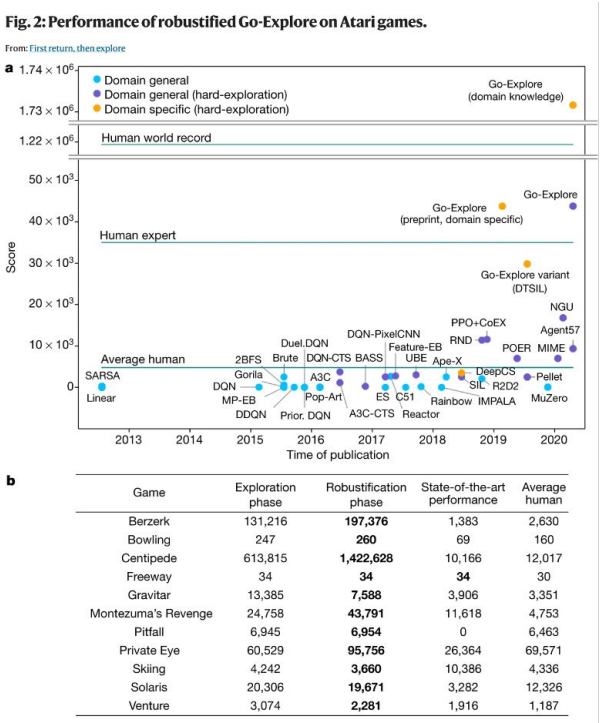


TALES: Text-Adventure Learning Environment Suite. Cui et al., arxiv 2025.

Background: Go-Explore

- Go-Explore decomposes hard exploration into **Go** and **Explore**
- Maintains a **state archive**
  - "**Go**" to a promising state
  - "**Explore**" from the state, expanding the archive





Game	Exploration phase	Robustification phase	State-of-the-art performance	Average human
Berzerk	131,216	**197,376**	1,383	2,630
Bowling	247	**260**	69	160
Centipede	613,815	**1,422,628**	10,166	12,017
Freeway	34	**34**	**34**	30
Gravitar	13,385	**7,588**	3,906	3,351
Montezuma's Revenge	24,758	**43,791**	11,618	4,753
Pitfall	6,945	**6,954**	0	6,463
Private Eye	60,529	**95,756**	26,364	69,571
Skiing	4,242	**3,660**	10,386	4,336
Solaris	20,306	**19,671**	3,282	12,326
Venture	3,074	**2,281**	1,916	1,187

Background: Go-Explore

- Several works including ours builds on Go-Explore
- eXploiT-and-eXplore (XTX)
  - **Go:** Imitation learned-policy
  - **Explore:** DQN w/ curiosity rew
- Intelligent Go-Explore (IGE)
  - **Go:** LLM to decide state
  - **Explore:** ReAct
- Global-local World Memory (GLoW, Ours)
  - **Go:** LLM-based state selection with more principled value decomp.
  - **Explore:** Multiple explorations while inferring “advantage”

Algorithm 1 Go-Explore-based Algorithms

procedure GO-EXPLORE-FAMILY(s_0, n_{iter})
  $\mathcal{A} \leftarrow \{(s_0, 0)\}$ ▷ Archive of (state, score)
  $\mathcal{T} \leftarrow \emptyset$ ▷ Collected trajectories
  $\mathcal{F} \leftarrow \emptyset$ ▷ Trajectory Frontier
  for $i = 1$ to n_{iter} **do**
    — *Go Phase (State Selection)* —
    Go-Explore: $s_{next} \sim h(\mathcal{A})$ ▷ Hand-crafted heuristic (e.g., visit count, domain score)
    XTX: $s_{next} \leftarrow \text{ImitationLearning}(\mathcal{T})$ ▷ Imitation learning
    IGE: $s_{next} \leftarrow \text{LLM.SelectPromising}(\mathcal{A})$ ▷ Ill-defined promising-ness
    GLoW: $W_{\text{global}} \leftarrow g_{\text{LLM}}(\mathcal{F})$ ▷ Principled value decomposition (Sec. 3.1)
      $s_{next} \leftarrow \text{align}_{\text{LLM}}(\mathcal{A}, W_{\text{global}})$

    — *Explore Phase* —
    Go-Explore: $\tau \leftarrow \text{RandomActions}(s_{next})$ ▷ No learning
    XTX: $\tau \leftarrow \text{DQN}(s_{next})$ ▷ DQN with curiosity reward
    IGE: $\tau \leftarrow \text{ReAct}(s_{next})$ ▷ Standard LLM agent
    GLoW:
    for $j = 1$ to n **do** ▷ LLM agent with advantage-driven exploration (Sec. 3.2)
      $\tau_j \leftarrow \pi_{explore}(s_{next}, W_{\text{local}}, \{\tau_1, \ldots, \tau_{j-1}\}, \mathcal{F})$
      $W_{\text{local}} \leftarrow \text{MAR}(\{\tau_1, \ldots, \tau_j\}, \mathcal{F})$
    end for
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{\tau_1, \ldots, \tau_n\}$ ▷ Collect trajectories
    $\mathcal{F} \leftarrow \text{top-}k(\mathcal{F} \cup \{\tau_1, \ldots, \tau_n\}, v)$ ▷ Update trajectory frontier

    — *Archive Update* —
    for each state s' in τ **do**
      if IsNotRedundant($s', \mathcal{A}$) **then** ▷ Domain-specific novelty
        $\mathcal{A} \leftarrow \mathcal{A} \cup \{s'\}$
      end if
    end for
  end for
end procedure

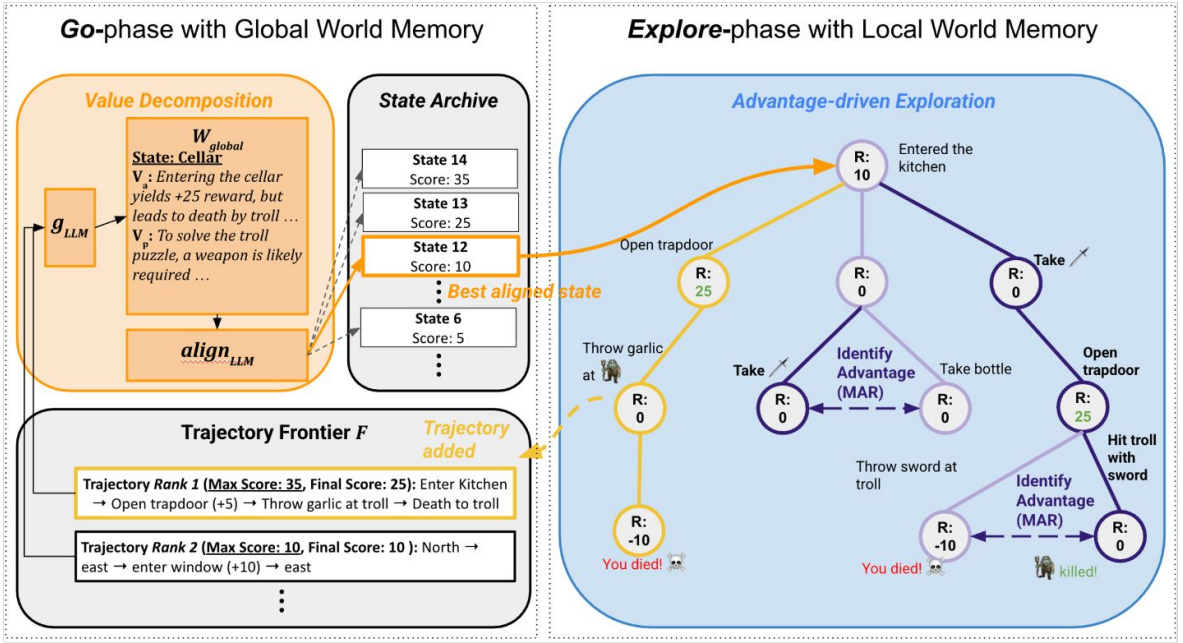
GLoW: Global-Local “World Memory”

Global World Memory (Go)

- Trajectory Frontier F
  - Top-k highest-value trajectories
- A form of value decomposition for selection (via LLM):
  - v for representing achieved values from state
  - v' for capturing estimated future potential

Local World Memory (Explore)

- Multi-path Advantage Reflection (MAR)
  - Compare n trajectories starting from same state
  - LLM infers semantic advantages at key states in an (s,a)-like form



“Go” with GWM

Value-Ranked Trajectory Frontier

- $F = \{\tau_1, \tau_2, \ldots, \tau_k\}$ are top-k trajectories ranked by max cumulative reward

LLM Value Decomposition

- $W_{global} = g_{LLM}(F) = \{(s_1, v_1, v'_1), (s_2, v_2, v'_2), \ldots\}$
- v = achieved value from state (empirical)
- v' = LLM estimate of future potential (*semantic form* of optimism under uncertainty)

Inspired by UCB: the potential value v' serves the same role as the exploration bonus, but through LLM reasoning about bottlenecks rather than visit counts.

W_{global} → State selection using LLM

Global World Memory Example (Zork1)

```
Strategic Analysis of Game Trajectories

==================================================

1.  FRONTIER & EXPLORATION STATUS

Successfully Reached Areas:
- Starting point:  Mailbox and House exterior (north, south, east,
west of house)
- Forest Path and Clearing with grating
- Up a Tree (bird's nest with egg and leaflet)
- Behind House (window entry)
- Kitchen and Living Room inside the house
- Attic (Trajectory 1)
```

...

```
- +25 for entering cellar (significant milestone)
- +5 for moving north from troll room to passage
- -10 on death and respawn (penalty)

Highest Reward Actions:
- Descending into cellar (+25)
- Collecting key items early (+5 to +10)
- Progressing past major checkpoints

Scoring Patterns:
- Early exploration yields small incremental rewards
- Major area transitions yield large rewards
- Death penalizes score, encouraging cautious play

==================================================

5.  NEXT INVESTIGATION GOALS

Specific Objectives:
- Develop reliable strategy to defeat or bypass troll without dying
- Investigate use of rope and knife for puzzles or combat
- Explore crawlway south and forbidding hole west in troll room
- Find safe method to cross or bypass chasm
- Attempt to open or circumvent nailed gothic door

Most Promising Unexplored Areas:
- Crawlway south and forbidding hole west in cellar/troll room
- Upstairs dark staircase (with lantern or other light source)
- Beyond barred trap door if it can be reopened
- Areas beyond chasm once safe crossing method found
```

State Selection ($align_{LLM}$) Prompt

```
=== STRATEGIC GAME ANALYSIS ===
{Analysis of frontier trajectories W_global}
==================================================

Based on the above analysis, select the state from the archive
that:
- Best aligns with the identified investigation goals
- Can help overcome identified bottlenecks
- Explores promising frontiers
- Has potential for high rewards based on patterns

Current state archive:

0:  [Score:  X, Steps:  Y, Visits:  Z]
  Observation:  {state observation}
  Inventory:  {state inventory}

1:  [Score:  X, Steps:  Y, Visits:  Z]
  Observation:  {state observation}
  Inventory:  {state inventory}

...

Choose state index (0-N).
Respond in JSON format:
{
  "thought":  "Your reasoning about which state best aligns with the
strategic goals",
  "index":  <number>
}
```

“Explore” with LWM

Motivation: Advantages over Q-values

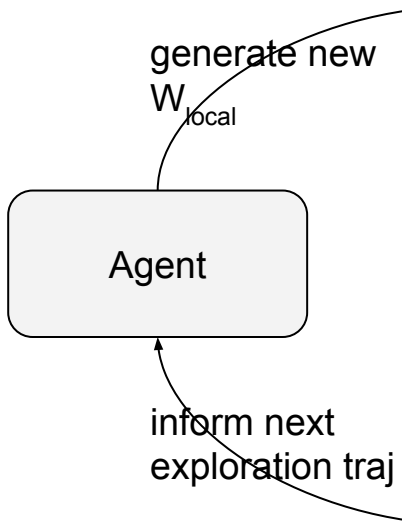
$A(s,a) = Q(s,a) - V(s)$

- Q-value estimation from sparse rewards
  → prone to high variance
- Single-trajectory reflection
  → prone to incorrect causal attribution

Multipath Advantage Reflection (MAR):

$W_{local} = MAR(T_s, F) = \{(s^*_1, As^*_1), ..., (s^*□, As^*□)\}$

1. Sample n trajectories from same starting state
2. LLM compares outcomes to identify semantic advantages
3. Frontier F serves as stable value baseline
4. Each trajectory τ_i benefits from insights of τ_{i-1}



Local World Memory Example (Zork1)

Based on detailed exploration attempts, here are KEY STATES with discovered advantages:

==========

STATE: The Troll Room
Small room with passages east/south, hole west; troll blocks all exits; sword glows

ADVANTAGES:
- "hit troll with sword" → kills troll, clears all exits for exploration (score: 40)
- "take axe" after troll death → obtains valuable weapon for later puzzles
- avoid engaging without glowing sword → likely failure
- optimal: kill troll first, then take axe (mandatory to proceed)

==========

STATE: Maintenance Room (Flood Control Dam #3)
Room with colored buttons, tool chests, wrench, screwdriver, toothpaste, leaking pipe

ADVANTAGES:
- "take wrench, screwdriver, tube" → essential tools for environment interaction
- "push blue button" → triggers leak, raises water level to access new areas
- "push red button" → toggles lights, affects water level
- avoid throwing lantern → breaks critical light source
- optimal: collect tools → manage buttons → control water without drowning

==========

STATE: Temple / Torch Room / Dome Room / Altar
Large temple with inscriptions; dome with railing; rope for descent; ivory torch; brass bell; gold coffin

ADVANTAGES:
- "take ivory torch" → stable light for deeper cave exploration
- "take bell" → key item for spirit/wraith interaction
- "ring bell at Entrance to Hades" → paralyzes wraiths, enables passage
- "blow out candles" → enables safe descent or passage
- optimal: acquire torch → bell → sceptre → manipulate altar → control spirits

==========

STATE: East-West Passage / Chasm Area
Narrow passage with stairs; chasm with paths; multiple routes (north/east/west/up/down)

ADVANTAGES:
- "east" then "north" → leads to Reservoir South and further areas
- "tie rope to railing" → enables safe descent into lower levels
- avoid getting stuck in loops → wastes moves
- optimal: explore chasm edges → use rope for vertical → access Dome/Torch

==========

Cross-Cutting Insights:
- Inventory Management: Strategic dropping/picking essential for critical artifacts
- Light Preservation: Maintaining lantern/torch crucial for dark exploration
- Combat Readiness: Glowing sword indicates combat opportunity (essential for progress)

Experiments: Baselines

RL Baselines:

- **DRRN:** Value-based RL (DQN)
- **KG-A2C:** Advantage Actor Critic (A2C) augmented by a dynamic knowledge graph
- **eXploit-then-eXplore (XTX):** Current state-of-the-art in Jericho, implementing Go-Explore with imitation learning on promising trajectories for state selection, and DQN with intrinsic curiosity reward for exploration

MCTS Baselines:

- **MC-LAVE**: Combines MCTS with language driven exploration, concentrating search effort on promising actions identified based on value estimates from semantically similar past actions.
- **MC-DML**: Enhances MCTS by incorporating LLMs as action priors in the PUCT algorithm, which balances exploration and exploitation during tree search.

LLM Baselines:

- **ReAct**: Standard LLM agent interleaving reasoning and acting
- **Reflexion**: Building on ReAct, incorporating self-reflection on each episode to guide future episodes
- **In-context Reinforcement Learning (ICRL)**: Concatenate window of previous episodes, rewards and instruction to behave based on observed episodes and rewards
- **Intelligent Go-Explore (IGE)**: Implement Go-explore using LLM-based selection, and ReAct-based exploration

Experiments: Results

- GLoW outperforms other LLM-based approaches, setting highest perf on 7/10 games
- On par with RL/MCTS-based methods, with 100-800x fewer interactions, showing effective exploration

Games	*RL-based*				*MCTS-based*		*LLM-based*				
	DRRN	**KG-A2C**	**RC-DQN**	**XTX**	**MC-LAVE**	**MC-DML**	**ReAct**	**Reflexion**	**ICRL**	**IGE**	**GLoW (Ours)**
Steps	1,000,000	1,600,000	100,000	800,000	~400,000	~400,000	1000	1000	1000	1000	1000
Enchanter	20	12.1	20	52.0	–	20±0.0	46.7±9.4	48.3±9.4	43.3±8.5	50.0±7.1	**61.7**±20.1
Zork1	32.6	40.2±0.4	38.8	**103.4**±10.9	45.2	48.66±1.89	48.3±4.7	48.0±5.0	51.7±4.7	44.3±0.5	73.0±4.5
Zork3	0.5	0.0	2.83	4.2±0.1	–	3±0.0	3.0±0.0	2.7±0.5	3.0±0.0	3.7±0.9	**4.3**±0.9
Deephome	1	20±2.1	1	**77.7**±2.1	35	67±1.41	11.0±4.2	22.0±1.6	24.0±5.7	71.3±4.9	75.0±8.7
Ludicorp	13.8	19.8±1.0	17	**78.8**	22.8	19.67±1.7	19.7±0.9	21.7±1.2	32.0±7.1	28.3±11.3	73.7±11.0
Balances	10	10	10	**24**	10	10±0.0	10±0.0	10±0.0	11.7±2.4	10.0±0.0	16.7±2.4
Pentari	27.2	44±0.9	43.8	49.6	68	**70**±0.0	30.0±0.0	30.0±0.0	26.7±4.7	30.0±0.0	30.0±0.0
Detective	197.8	338±3.4	291.3	312.2	330	**346.67**±9.43	113.3±4.7	166.7±20.5	233.3±47.8	316.7±4.7	310.0±8.2
Temple	7.4	8	8	–	8±0.0	8±0.0	8.7±0.9	8.7±0.9	8±0.0	**13.7**±0.9	13.0±0.0
Ztuu	21.6	5±0.0	–	–	7	23.67±1.9	18.7±2.4	18.3±2.6	16.7±4.1	15.0±9.1	**29.3**±4.0

Experiments: Ablations

Ablation Variants	Zork1	Zork3	Enchanter	Deephome	Ludicorp	Balances
GLoW (Full)	**73.0**$\pm$4.5	**4.3**$\pm$0.9	**61.7**$\pm$20.1	**75.0**$\pm$8.7	**73.7**$\pm$11.0	**16.7**$\pm$2.4
✗ [Local WM] Multi-path Advantage Reflection (MAR)	70.0$\pm$13.6	4.3$\pm$0.5	51.7$\pm$9.4	56.7$\pm$21.7	54.7$\pm$22.4	11.7$\pm$2.4
✗ [Global WM] State selection with W_{global}	62.0$\pm$15.6	4.3$\pm$0.9	60.0$\pm$10.8	61.3$\pm$26.0	63.3$\pm$14.7	13.3$\pm$2.4
✗ [Global WM] Trajectory frontier $\mathcal{F}$	61.7$\pm$1.9	4.0$\pm$0.8	53.3$\pm$10.3	57.7$\pm$23.3	63.3$\pm$12.3	11.7$\pm$2.4
✗ All above	51.3$\pm$5.2	4.3$\pm$0.9	51.7$\pm$9.4	56.0$\pm$21.2	22.0$\pm$0.8	10.0$\pm$0.0
Standard IGE	44.3$\pm$0.5	3.7$\pm$0.9	50.0$\pm$7.1	71.3$\pm$4.9	28.3$\pm$11.3	10.0$\pm$0.0

- MAR performs better than Reflection for guiding exploration
- W_{global} is effective for state selection
- The global and local components are complementary

Experiments: Scaling to stronger LLMs

- GLoW is effective using GPT-4.1: Performance scales across games
- Surpasses XTX on 5/6 games
- State of the art on 4/6 games

Games	*RL*	*LLM-based*									
		GPT-4.1 mini					*GPT-4.1*				
	XTX	**ReAct**	**Rfl**	**ICRL**	**IGE**	**GLoW**	**ReAct**	**Rfl**	**ICRL**	**IGE**	**GLoW**
Steps	800K	1K	1K	1K	1K	1K	1K	1K	1K	1K	1K
Enchanter	52.0	46.7±9.4	48.3±9.4	43.3±8.5	50.0±7.1	61.7±20.1	38.3±2.4	58.3±2.4	45±7.1	68.3±18.4	**98.3**±4.7
Zork1	**103.4**±10.9	48.3±4.7	48.0±5.0	51.7±4.7	44.3±0.5	73.0±4.5	45.0±0.0	54.3±4.5	48.0±2.8	86.7±24.1	103.0±6.8
Zork3	4.2±0.1	3.0±0.0	2.7±0.5	3.0±0.0	3.7±0.9	4.3±0.9	3.3±0.5	2.7±0.5	3.0±0.8	3.0±0.0	**5.0**±0.0
Deephome	77.7±2.1	11.0±4.2	22.0±1.6	24.0±5.7	71.3±4.9	75.0±8.7	32.3±19.6	22.3±1.7	34.7±18.7	82.0±8.6	**114.7**±27.8
Ludicorp	78.8	19.7±0.9	21.7±1.2	32.0±7.1	28.3±11.3	73.7±11.0	31.0±2.8	29.0±0.8	31.7±0.5	**89.0**±7.8	79.0±16.8
Balances	24	10±0.0	10±0.0	11.7±2.4	10.0±0.0	16.7±2.4	18.3±2.4	18.3±2.4	16.7±2.4	16.7±2.4	**26.7**±2.4

Qualitative Analysis of Failure Modes

Case 1: Zork1

Game Context:
- Reached deeper frontier states through Troll Room to Maze (score 50) and Chasm (score 45).
- State archive includes:
  - State 72: Score 50, Location: Maze, Inventory: lantern, sword, ...
  - State 58: Score 45, Location: Chasm, Inventory: lantern, sword
  - State 37: Score 15, Location: Kitchen, Inventory: empty

Observed Behavior:
Repetitive state selection:
- Selection Iter. 2: **State 37** (score 15)
- Selection Iter. 3: **State 37** (score 15)
- Selection Iter. 4: **State 37** (score 15)

Reasoning for state selection: *"State 37 is in the living room where the brass lantern and the elvish sword are present. Acquiring and activating these items is crucial for safe exploration... This state aligns well with investigation goals..."*

Case 2: Deephome

Game Context:
- Agent has activated the City Generator (+30 points), which powers the Railway Station controls.
- Further progress requires using the Railway to access intermediate areas (Blacksmith, Waterfall).

Observed Behavior:
- Step 827: Activates generator (+30 points, score 65)

Failed subsequent multi-stage progression:
- Steps 828–1000:
  - Railway Station visits: **0**
  - Blacksmith visits: **0**
  - Waterfall visits: **0**
  - Water Works: 2 attempts at wheel, both fail

Mode 1: Conservative state selection
- (Zork1 example) Overprioritizing safety and early-game potential over deeper state with uncertainty

Mode 2: Multi-step dependency reasoning failure
- (Deephome example) Solves one puzzle which is a dependency for others, but fails to capitalize, reverting to exploring familiar areas leading to no new progress

Commonly, LLM's understanding of value tends to be myopic, handling few-step lookahead well, but struggling with longer causal chains

Thank you