DeepMaze: A Maze Benchmark for Quantifying Reasoning Depth in Language Models

Anonymous ACL submission

Abstract

Existing benchmarks for evaluating reasoning in large language models primarily emphasize final-answer correctness, making it difficult to distinguish genuine multi-step reasoning from statistical shortcuts and prior knowledge exploitation. Accurately measuring reasoning depth requires environments where this confounding is eliminated; however, existing benchmarks retain data contamination and static test distribution bias that enable prior knowledge to masquerade as reasoning, preventing clean capability isolation. To overcome this, we introduce DEEPMAZE, a minimalist benchmark of procedurally generated environments with rigorously controlled topology. Its dual-task architecture—comprising *planning* under full observability and *exploration* under partial observability—inherently necessitates sustained, state-consistent reasoning by requiring models to dynamically track environmental states across sequential actions. Within this environment, we define a reasoning depth metric that quantifies the length of state-consistent action sequences, explicitly decoupling process quality from outcome success. This design isolates LLMs' core reasoning capabilities under controlled conditions, establishing a foundation for evaluating their true multi-step reasoning proficiency independent of domain-specific knowledge or outcome-driven shortcuts.

1 Introduction

Spatial reasoning ability refers to the capacity to construct, organize, manipulate, and transform spatial information and spatial relations (Newcombe and Shipley, 2014). In real-world and embodied tasks, this capability requires sustained, state-consistent reasoning over multi-step actions to interact effectively with dynamic environments. In recent years, a growing number of benchmarks have been proposed to evaluate the spatial reasoning capabilities of large language models and multimodal models, providing important tools for understanding models' spatial cognition. However, existing approaches still suffer from several critical limitations, which constrain their ability to faithfully characterize models' genuine reasoning capabilities.

First, existing benchmarks fail to disentangle and explicitly characterize reasoning depth. Reasoning depth requires quantifying the length of state-consistent action sequences—where each decision depends on dynamically updated internal states formed from prior actions—in ways that cannot be replaced by statistical shortcuts. Crucially, most evaluations equate reasoning with final-answer correctness, conflating genuine multi-step reasoning with outcome success. This prevents independent assessment of how much sustained, state-consistent reasoning a model performs during execution, as even memorized solutions yield correct answers without intermediate state updates.

Second, evaluation outcomes are heavily confounded by static test distribution bias. Fixed or limited-scale datasets contain inherent statistical regularities (e.g., frequent path patterns or layout symmetries), enabling models to exploit pattern matching and prior knowledge instead of explicit reasoning. Models succeed via statistical shortcuts that bypass multi-step state-dependent decisions, as environmental complexity alone cannot eliminate these biases when test distributions remain static. Consequently, performance reflects dataset memorization rather than adaptive reasoning capability.

Third, data contamination critically undermines evaluation validity. When test instances leak into pretraining or fine-tuning data—a common issue in benchmarks with limited procedural generation—models leverage prior knowledge to masquerade as reasoning. Performance then measures memorization of specific instances rather than state-consistent reasoning in novel situations, severely

weakening the benchmark's ability to isolate true reasoning competence. This contamination, combined with static distributions, creates an environment where spurious shortcuts dominate genuine capability assessment.

Guided by the principle of "less is more"—namely, isolating the core mechanisms of reasoning and memory by stripping away semantic content and prior-knowledge cues through a minimalist maze environment—we introduce DEEPMAZE, a benchmark built upon procedurally generated mazes. DEEPMAZE adopts a two-level progressive task hierarchy to systematically decompose and probe different dimensions of spatial reasoning ability: (1) **Planning**: given the full observation of maze layout, the model plans and executes moves step-by-step—with multiple attempts allowed—until reaching the goal, evaluating its long-horizon planning capability and enabling the assessment of reasoning depth; (2) **Exploration**: the model is restricted to local observations of its surrounding environment and must make decisions while exploring, thereby forcing it to construct and continuously update an internal spatial representation; Based on the above design philosophy and methodological considerations, this paper makes the following main contributions:

(1) We propose DEEPMAZE, a minimalist benchmark leveraging procedurally generated environments that isolates core reasoning capabilities from prior knowledge exploitation. The benchmark features a rigorously controlled dataset of over 50,000 distinct maze instances spanning 10 grid scales (from 5×5 to 29×29), where topological diversity eliminates statistical shortcuts while enhancing generalization.

(2) We introduce a *Reasoning Depth Metric* that quantifies sustained state-consistent reasoning through consecutive valid actions, decoupling process quality from outcome success. This metric provides the first standardized method to measure reasoning chain length independent of final-answer correctness.

(3) We establish a dual-task evaluation framework—comprising *planning* (full observability) and *exploration* (partial observability)—that systematically probes state-tracking capabilities under increasing environmental complexity. Comprehensive analyses across state-of-the-art models reveal a fundamental *reasoning gap*: high success rates often mask failures in maintaining state-consistent logic as complexity grows.

2 Related Works

2.1 General Reasoning Benchmarks

Evaluating multi-step reasoning in LLMs has led to diverse benchmarks spanning mathematical, logical, and physical domains. Mathematical benchmarks like MATH (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021) assess solution correctness but **fail to disentangle reasoning depth from formulaic shortcuts**. For example, a model may solve a geometry problem by directly recalling "area = πr^2" without reconstructing the derivation—compressing multi-step spatial reasoning into a single memorized step. Similarly, theorem-proving benchmarks (e.g., Lean (Yang et al., 2023)) rely on formal syntax that allows proof-skipping via lemma reuse. Logical reasoning datasets (e.g., ReClor (Yu et al., 2020), LOGIQA (Liu et al., 2020)) emphasize deductive validity but lack metrics for intermediate state consistency, conflating correct conclusions with sound inference processes. Crucially, these benchmarks share a core flaw: static test distributions enable exploitation of secondary conclusions, preventing isolation of pure reasoning depth. Our work bridges this gap by designing a domain-agnostic evaluation framework where reasoning depth is inherently uncompressible due to procedural topology control.

2.2 Spatial Reasoning Benchmarks

Existing spatial benchmarks (e.g., SpatialBench (Xu et al., 2025), ALFWorld (Shridhar et al., 2021)) inherit the above limitations. Real-world scene datasets embed semantic priors (e.g., "kitchens contain fridges"), allowing models to bypass geometric reasoning via object co-occurrence statistics. Even procedurally generated environments (e.g., VirtualHome (Puig et al., 2018)) suffer from layout regularities that enable memorization. Critically, **none decouple reasoning depth from outcome success**—a gap our state-consistent metric directly addresses.

2.3 Maze-based Benchmarks and Studies

Maze tasks have long served as a classical paradigm for evaluating multi-step reasoning due to their simple structure, abstract semantics, and inherent requirement for sequential decision-making. In reinforcement learning literature, benchmarks such as MiniGrid (Chevalier-Boisvert et al., 2023) and BabyAI (Chevalier-Boisvert et al., 2019)

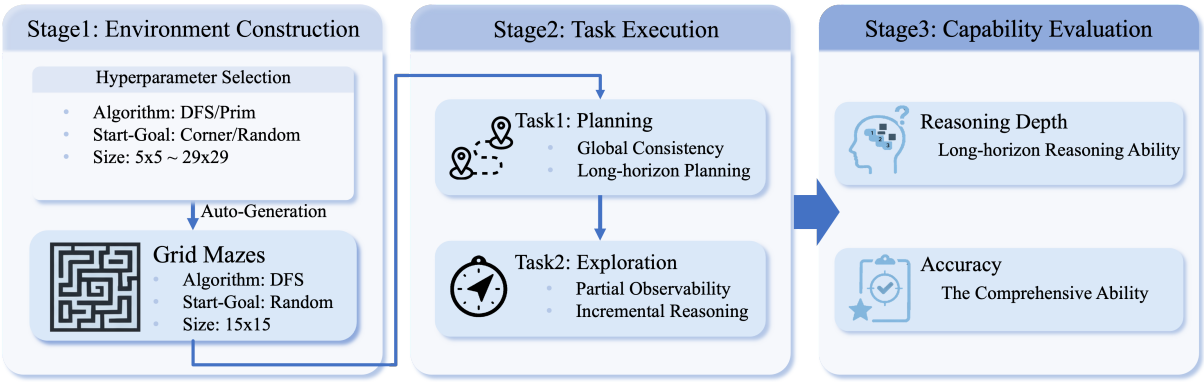

Figure 1: Overview of DeepMaze

have been widely adopted to assess navigation capabilities. However, these frameworks primarily target RL agents and lack standardized metrics for evaluating reasoning depth in language models.

Einarsson (Einarsson, 2025) introduced MazeEval to evaluate LLM decision-making in coordinate-based mazes using sparse feedback signals, with findings showing significant performance degradation in larger mazes and non-English contexts. Dao et al. (Dao and Vu, 2025) developed AlphaMaze with tokenized maze representations trained via SFT and GRPO, but their approach suffers from quadratic sequence growth with maze size and lacks partial observability support, limiting scalability for complex reasoning evaluation.

Critically, existing maze-based studies predominantly treat mazes as *end tasks* rather than *reasoning probes*. They focus on navigation success rates within the maze domain without establishing how maze performance reflects fundamental reasoning capabilities such as state consistency maintenance or resistance to statistical shortcuts. These works fail to isolate and quantify the core reasoning mechanisms that transcend specific maze configurations.

In contrast, our benchmark treats mazes as a *minimalist reasoning microscope*. By stripping away semantic content and prior-knowledge cues through controlled procedural generation, DeepMaze directly measures state-consistent reasoning depth—quantified by consecutive valid actions—across two complementary conditions: planning under full observability and exploration under partial observability. This design enables precise characterization of how reasoning quality degrades with environmental complexity, providing transferable insights about model reasoning capabilities beyond maze navigation.

3 Benchmark Description

DeepMaze defines three progressively challenging tasks to isolate reasoning capabilities. Built on procedural maze generation, it enables rigorous control over topology and observation constraints.

3.1 Environment Formalization

A maze is a grid graph $\mathcal{M} = (V, E)$ where $V \subseteq [H] \times [W]$ contains traversable cells. State $s_t = (p_t, g)$ tracks agent position p_t and goal g. Actions $\mathcal{A} = \{\texttt{U}, \texttt{D}, \texttt{L}, \texttt{R}\}$ trigger deterministic transitions:

$$s_{t+1} = \mathcal{T}(s_t, a_t) = \begin{cases} (p_t + \Delta a_t, g) & \text{if valid} \\ s_t & \text{otherwise} \end{cases}$$

Observations depend on radius r: $o_t = \mathcal{O}(p_t, g; \mathcal{M}, r)$.

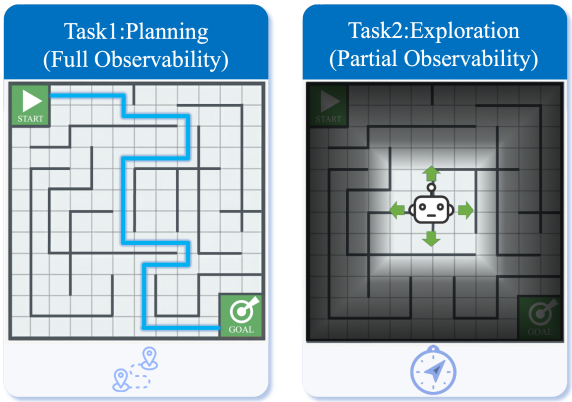

Figure 2: DeepMaze's three-task framework: (Left) Global planning with full observability; (Right) Exploration under partial observability.

3.2 Task 1: Global Planning

The agent receives full maze observation o_t^{global} and outputs a single action sequence $\mathbf{a}_{t:t+K}$.

Table 1: Comparison of Maze-Based Benchmarks for LLM Reasoning Evaluation

Evaluation Dimension	MazeEval (Einarsson, 2025)	MazeBench (Dao and Vu, 2025)	DeepMaze (Ours)
Reasoning depth quantification	✗	✗	✓
State-consistency measurement	✗	✗	✓
Observation modes supported	Partial only	Global only	**Global & Partial**
Maze size range	5×5–15×15	5×5	**5×5–29×29**
Evaluation dimensions	Outcome-only	Outcome-only	**Process & Outcome**
Token efficiency (global)	N/A	$\approx 32n^2$	$\mathbf{n^2}$
Primary purpose	Navigation task	Navigation task	**Reasoning probe**
Task paradigm	Exploration	Planning	**Planning & Exploration**

Execution follows strict all-or-nothing rule:

$$\mathcal{E}(s_t, \mathbf{a}) = \begin{cases} s_t & \text{if any invalid action} \\ s_{t+K} & \text{otherwise} \end{cases}$$

Unlike prior work, we evaluate plan coherence beyond endpoint success.

3.3 Task 2: Local Exploration

The agent observes only a local window of radius $r \ll \min(H, W)$:

$$o_t = \mathcal{O}(p_t, g; \mathcal{M}, r).$$

It outputs short action sequences incrementally based on history $o_{\leq t}$, evaluating online reasoning under uncertainty.

3.4 Evaluation Metrics

For N episodes $\mathcal{E}$ with max steps T_{max}:

Success Rate

$$\mathbb{I}_{\text{succ}}(e) = \begin{cases} 1 & \text{if reached } g \text{ within } T_{\text{max}} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Success} = \frac{1}{N} \sum_{e \in \mathcal{E}} \mathbb{I}_{\text{succ}}(e).$$

Reasoning Depth For episode e with m_e planning attempts:

$$k_{e,i} = \max\{k \mid \forall j \leq k\text{: } a_j^{(i)} \text{ valid}\}$$

$$\text{RD}(e) = \frac{1}{m_e} \sum_{i=1}^{m_e} k_{e,i}$$

$$\text{Reasoning Depth} = \frac{1}{N} \sum_{e \in \mathcal{E}} \text{RD}(e)$$

4 Experiments

4.1 Maze Configuration

We generate mazes using the *Recursive Division Algorithm*, producing connected acyclic structures in ASCII format (walls '#', empty cells '0', agent 'A', goal 'G'). Maze sizes are scaled continuously:

- Task 1: 5×5 to 29×29 grids (critical thresholds at 11×11, 13×13)
- Task 2: 5×5 to 9×9 grids with partial observability

For Task 2, we implement configurable visibility radii (1-7 cells) using ray-casting for occlusion simulation, with 3-cell radius as default. All experiments use 1,000 trials per condition (600 for Qwen-3-max-preview due to computational constraints).

4.2 Models and Context Mechanism

We evaluate the following models with DeepSeek-V3.2 as our primary baseline:

- **DeepSeek-V3.2** (baseline): Standard variant with temperature=0 and 1.0 settings
- **DeepSeek-V3.2-Thinking**: Enhanced reasoning variant with iterative problem-solving capability
- **Proprietary models**: GPT-4o, GPT-5.1, GPT-o3
- **Other open-source**: Qwen-3-max-preview

All models operate with a context window containing the **most recent k interaction cycles**, where each cycle consists of an *action sequence* proposed by the model followed by *execution feedback* from the environment (valid moves, collisions, and partial observations). This forms a closed-loop interaction history that enables state tracking and adaptive decision-making. For Task 2, we systematically evaluate different context window sizes ($k \in \{3, 5, 7\}$) to quantify memory requirements for effective exploration.

Task 2 constraints: Step limits ($k \times n^2$, $k \in \{1.0, 2.0\}$), exploration patience ($m \in \{3, 5, 7, 9\}$ consecutive non-exploring steps), memory window sizes (3-7 interaction cycles), and visibility radii (1-7 cells). Section 3.3 presents our ablation study on memory window size, revealing critical thresholds for maintaining state consistency during exploration.

4.3 Results and Analysis

4.3.1 Task 1

Accuracy trends. The experimental results reveal critical insights into model performance under increasing spatial complexity. As shown in Figure 3, all non-reasoning models exhibit a clear negative correlation between maze size and success rate, with degradation becoming pronounced beyond 11×11 grids. However, the rate and severity of decline vary dramatically across architectures and configurations.

The *DeepSeek-V3.2-Thinking* variant demonstrates exceptional robustness, maintaining perfect success rates (1.000) across all grid sizes from 5×5 to 29×29. This suggests that its "thinking" mechanism—enabling multi-step reasoning and iterative planning—effectively mitigates combinatorial complexity in larger mazes. In contrast, the non-reasoning *DeepSeek-V3.2 (temperature=0)* configuration suffers a sharp performance cliff: while achieving 0.80 success at 5×5, it drops to 0.00 for all grids $\geq$ 13×13. This abrupt failure indicates that greedy decoding without stochastic exploration cannot handle pathfinding tasks requiring long-horizon planning.

The divergence in medium-to-large grids ($\geq$ 15×15) is particularly instructive: while all non-reasoning models fail completely in this regime, the "thinking" variant sustains perfect performance. This underscores the critical role of explicit reasoning mechanisms in solving scalable spatial reasoning tasks. The results align with theoretical expectations that maze-solving requires hierarchical planning—a capability embedded in the "thinking" configuration but absent in standard autoregressive decoding.

These findings have practical implications for real-world navigation tasks where environmental complexity is unbounded. The stark contrast between reasoning and non-reasoning configurations suggests future work should prioritize architectures with built-in planning capabilities rather than relying solely on scaling model size or data. The clear performance separation at 13×13 grids (where non-reasoning models fail) provides a critical benchmark for evaluating spatial reasoning capabilities in LLMs.

Reasoning depth analysis. The grouped bar chart in Figure 4 reveals a fundamental dichotomy in problem-solving approaches between reasoning-enabled and standard architectures. At the 7×7 scale, all models exhibit comparable reasoning depth, confirming that small-scale mazes do not adequately stress-test planning capabilities. However, as complexity increases to 15×15 and 25×25, a stark stratification emerges: models with explicit reasoning mechanisms demonstrate exponential depth growth while standard variants plateau.

The *DeepSeek-V3.2-Thinking* model dominates across all scales, achieving 74.8 steps at 15×15 and 111.4 at 25×25—3.08× deeper than *GPT-o3* (36.2 steps). Crucially, this sustained depth persists despite in larger grids, indicating that the "thinking" mechanism enables systematic state-space exploration even when solutions remain elusive.

In contrast, standard models like *DeepSeek-V3.2 (temperature=0)* and *GPT-5.1* exhibit near-flat performance curves, with depths below 15 steps even at 25×25. The temperature-ablated configuration shows critically low depth (2.8 steps at 25×25), directly correlating with its abrupt success rate collapse beyond 13×13 grids. This establishes stochastic exploration as a necessary condition for maintaining cognitive depth in complex environments.

The critical divergence between success rate and reasoning depth reveals a key insight: while non-reasoning models fail both metrics at large scales, the *Thinking* variant's sustained depth demonstrates "persistent problem-solving" behavior—continuing meaningful exploration even when

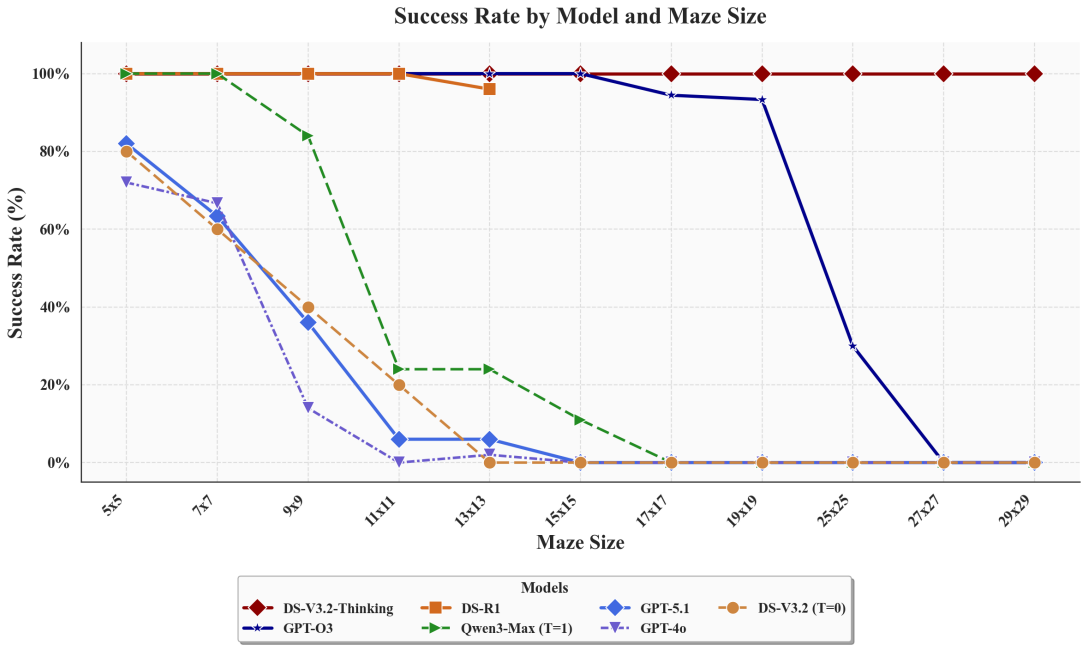

Figure 3: Success rate degradation across maze complexities for different LLM architectures. The plot demonstrates critical performance divergence: reasoning-enabled models (red diamonds) maintain perfect success rates (1.00) through 29×29 grids, while non-reasoning variants (blue circles) exhibit sharp performance cliffs beyond 11×11 grids. Model identifiers: DeepSeek-V3.2-Thinking (red), DeepSeek-V3.2 (temperature=0) (blue), Qwen-3-max-preview (teal), and other variants (light blue).

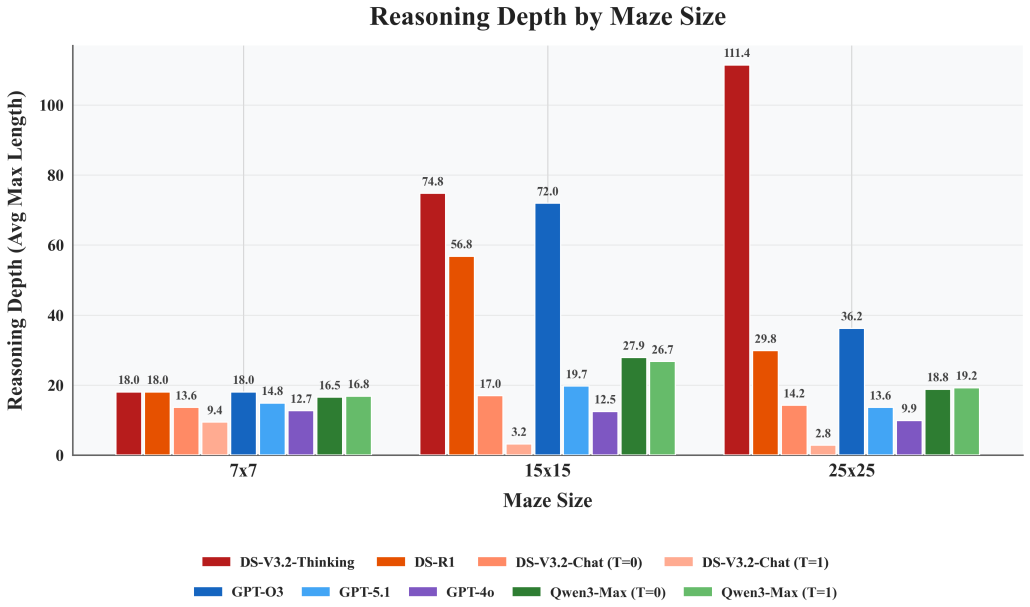

Figure 4: Reasoning depth (in tokenized reasoning steps) across maze complexities. The chart reveals a critical bifurcation: reasoning-enabled models (hatched bars) maintain exponential depth growth (peaking at 111.4 steps for DeepSeek-V3.2-Thinking at 25×25), while non-reasoning variants (solid bars) plateau below 15 steps.

success is unattainable within trial constraints. This suggests explicit planning mechanisms fundamentally alter cognitive processes by extending the model's planning horizon rather than merely improving final outcomes.

These results challenge conventional accuracy-only benchmarks. The depth metric exposes valuable intermediate reasoning capabilities in models that may fail final evaluation—a distinction critical for applications requiring robust partial solutions (e.g., autonomous navigation where partial pathfinding provides actionable data). Future work should investigate whether this depth metric correlates with solution quality when success is achieved, and whether it can inform new training objectives for complex reasoning tasks. The consistent 13×13 grid threshold where non-reasoning models fail both metrics provides a concrete benchmark for evaluating spatial reasoning capabilities.

4.3.2 Task 2

For Task 2, we evaluate models on maze sizes ranging from 5×5 to 9×9 with the following constraints:

- **Step limits**: Maximum attempts scaled as $k \times n^2$ where n is maze size (e.g., 1.0x = n^2 steps, 2.0x = $2n^2$ steps)
- **Exploration constraints**: Termination after m consecutive steps without discovering new cells (e.g., `noexplore_9` fails after 9 non-exploring steps)
- **Visibility radius**: Agent's perceptual field size (1-7 cells visible around current position)
- **Memory window**: Number of recent steps retained in context (3-7 steps)

Fig. 5a shows success rates increase with visibility radius, with 7×7 mazes achieving the highest success rate (0.5) at visibility radius 7. 5×5 and 9×9 mazes show lower maximum success rates (0.36 and 0.2 respectively).

Fig. 5b shows success rates decrease as step limits become stricter (lower k) and exploration patience decreases (lower m). For 9×9 mazes, success rate drops from 0.06 (2.0x_noexplore_9) to 0.03 (1.0x_noexplore_9), representing a 50% reduction.

Fig. 5c shows 5×5 mazes achieve peak success rate (0.24) at memory window size 7, while 9×9 mazes show the lowest success rates across all memory window sizes (maximum 0.04 at size 7).

Fig. 5d shows DeepSeek-V3.2-Thinking achieves near-perfect success rates (0.99-1.0) on 5×5 mazes, significantly outperforming other models. GPT-o3 demonstrates moderate performance (0.75 on 5×5, 0.7 on 7×7, 0.25 on 9×9), while GPT-5.1 and DeepSeek-V3.2 maintain consistently low success rates (≤0.03) across all maze sizes. Notably, DeepSeek-V3.2-Thinking outperforms all other models by a significant margin in all maze dimensions.

The data shows a 3.2× decrease in success rate from 5×5 to 9×9 mazes across most experimental conditions. Success rates for 7×7 mazes consistently outperform both smaller and larger mazes in visibility and step limit experiments.

5 Conclusion

5.1 Conclusion

We introduce DEEPMAZE, a minimalist benchmark designed to rigorously isolate and quantify the core spatial reasoning capabilities of large language models. By leveraging procedurally generated maze environments with controlled topology, we eliminate the confounding effects of prior knowledge exploitation and statistical shortcuts that plague existing benchmarks. Our dual-task architecture—comprising *planning* under full observability and *exploration* under partial observability—forces models to maintain state-consistent reasoning across sequential actions, while our novel *Reasoning Depth Metric* explicitly decouples process quality from outcome success.

Experimental validation reveals critical insights about LLM reasoning capabilities. We identify a fundamental *reasoning gap*: models with explicit reasoning mechanisms (e.g., DeepSeek-V3.2-Thinking) sustain exponential growth in reasoning depth (up to 111.4 steps) and perfect success rates even in 29×29 mazes, while standard architectures fail catastrophically beyond 13×13 grids despite comparable performance on small mazes. This bifurcation demonstrates that genuine multi-step reasoning requires specialized architectural support beyond scaling parameters or training data. In partial observability settings, we establish that memory window size and visibility radius are critical constraints, with success rates dropping 50% when step limits tighten and exploration patience decreases.

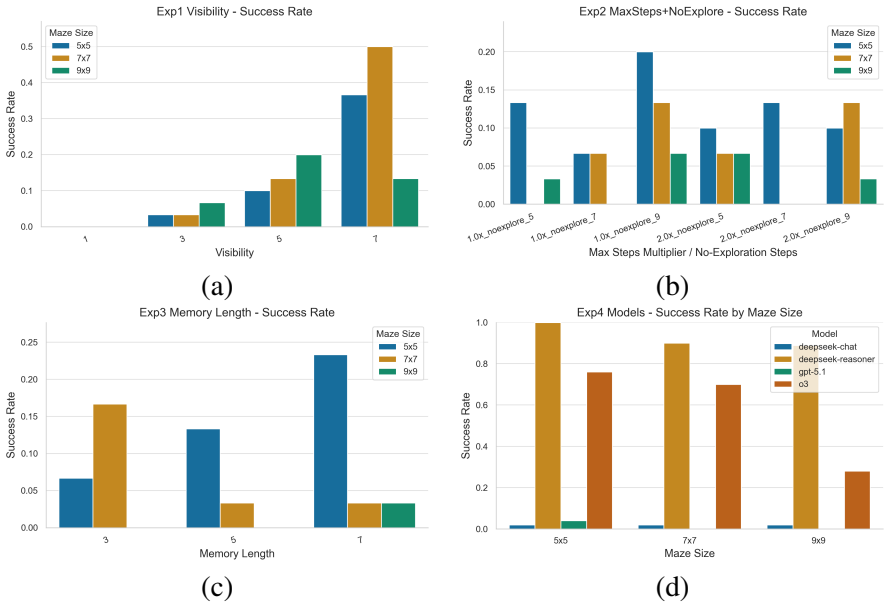

Figure 5: Experimental results across four conditions: (a) Success rate vs. visibility radius (1-7 cells); (b) Success rate under step limits ($k \times n^2$ steps where n=maze size) and exploration constraints (failure after m consecutive non-exploring steps); (c) Success rate vs. memory window size (3-7 steps); (d) Model performance comparison across maze sizes for DeepSeek-V3.2, DeepSeek-V3.2-Thinking, GPT-5.1, and GPT-o3. All experiments evaluate maze-solving success from 5×5 to 9×9 grids.

DEEPMAZE establishes three foundational principles for reasoning evaluation: (1) **Process over outcome**: Reasoning quality must be measured through intermediate state consistency, not just final answers; (2) **Controlled complexity**: Environmental complexity must be systematically scalable to expose reasoning thresholds; (3) **Diagnostic minimalism**: Core reasoning mechanisms are best isolated by stripping away semantic content and prior-knowledge cues.

These principles position DEEPMAZE not as a domain-specific evaluator, but as a diagnostic layer in the reasoning hierarchy—akin to unit tests in software engineering—that certifies fundamental reasoning competence before deployment in complex environments. Failures here guarantee failures in real-world tasks requiring sustained state-consistent decisions, from supply chain optimization to medical diagnosis.

Future work will extend this framework to: (1) Integrate multimodal observations to bridge the gap between abstract reasoning and embodied cognition; (2) Develop explicit memory architectures that maintain state consistency at scale, informed by our finding that reasoning depth collapses when memory windows fall below critical thresholds; (3) Establish cross-benchmark correlations to validate whether reasoning depth in minimalist environments predicts performance in complex domain-specific tasks.

As LLMs increasingly operate in safety-critical domains, DEEPMAZE provides a necessary foundation for validating their core reasoning capabilities under controlled conditions—ensuring that apparent competence reflects genuine cognitive depth rather than statistical mirages.

Limitations

While DeepMaze provides rigorous diagnostics for reasoning depth, four boundaries define its scope:

(1) Multimodal grounding: Our minimal environment intentionally excludes visual, auditory, or tactile inputs, as well as physical dynamics (e.g., time constraints, resource limitations). Consequently, DeepMaze *cannot evaluate* reasoning integrated with sensory processing or real-world operational constraints—capabilities requiring domain-specific simulators or physical testbeds.

(2) Domain knowledge integration: The absence of semantic content (e.g., conceptual relationships, factual priors) means DeepMaze isolates *pure structural reasoning*, but cannot assess how domain knowledge interacts with logical inference—a dimension critical for applications like scientific reasoning or strategic planning.

(3) Memory mechanism analysis: Although sustained reasoning inherently relies on memory, our framework currently lacks tools to directly probe *how models utilize memory during reasoning*. We quantify reasoning depth through behavioral outcomes but cannot dissect memory operations (e.g., encoding, retrieval, consolidation) that enable state consistency. This gap prevents us from establishing causal links between architectural memory mechanisms and reasoning performance degradation.

(4) Model coverage: Evaluations currently exclude some latest closed-source models (e.g., Gemini-series), though our open benchmark enables immediate validation upon API availability.

Critically, these limitations reflect *deliberate scoping choices*, not oversights. DeepMaze targets a specific layer in the reasoning hierarchy (Fig. 1): *"It is impossible to validate multi-step reasoning in high-noise environments before verifying its existence in controlled settings."* Thus, results should be interpreted as **complementary diagnostics**—not replacements—for domain-specific evaluations. The limitation regarding memory mechanisms particularly motivates our future work on explicit memory architectures (Section X), where DeepMaze will serve as the foundational testbed for isolating how different memory designs impact reasoning depth. This layered validation approach establishes necessary prerequisites before introducing confounding variables from perception, knowledge, or physical constraints.

References

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2019. Babyai: A platform to study the sample efficiency of grounded language learning. *Preprint*, arXiv:1810.08272.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. 2023. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Preprint*, arXiv:2306.13831.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Alan Dao and Dinh Bach Vu. 2025. Alphamaze: Enhancing large language models' spatial intelligence via grpo. *arXiv preprint arXiv:2502.14669*.

Hafsteinn Einarsson. 2025. Mazeeval: A benchmark for testing sequential decision-making in language models. *arXiv preprint arXiv:2507.20395*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *Preprint*, arXiv:2007.08124.

Nora S Newcombe and Thomas F Shipley. 2014. Thinking about spatial thinking: New typology, new assessments. In *Studying visual and spatial reasoning for design creativity*, pages 179–192. Springer.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. Virtualhome: Simulating household activities via programs. *Preprint*, arXiv:1806.07011.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, and Yunjian Zhang. 2025. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *Preprint*, arXiv:2511.21471.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. 2023. Leandojo: Theorem proving with retrieval-augmented language models. *Preprint*, arXiv:2306.15626.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *Preprint*, arXiv:2002.04326.

A Discussion: DeepMaze as a Diagnostic Layer for Reasoning Capability

Our benchmark intentionally adopts minimalist grid environments to isolate reasoning depth—a design choice that requires explicit contextualization against complex reasoning tasks. Crucially, DeepMaze functions not as a standalone evaluator, but as a controlled diagnostic layer that reveals core reasoning flaws before deployment in complex environments (Fig. 1), analogous to unit tests in software engineering that expose fundamental logic errors prior to system integration. This perspective reframes apparent "toy" constraints as methodological strengths: failures in DeepMaze certify fundamental gaps that persist across diverse reasoning domains.

To formalize this mapping, Table 2 aligns our abstractions with constraints shared across real-world reasoning tasks:

Mechanism	**Core Constraint**	**Real-World Manifestation**
State-consistent action sequences	Strict state transition dependency	Sequential decision-making under evolving conditions (e.g., supply chain optimization with dynamic constraints)
Partial observability in exploration	Limited information access	Medical diagnosis with incomplete patient history or financial forecasting with partial market data
Topology complexity scaling	Combinatorial state space growth	Strategic planning in complex systems (e.g., logistics networks with interdependent nodes)

Table 2: DeepMaze isolates reasoning primitives shared across domains. Failures here (e.g., state inconsistency at high complexity) *guarantee* failure in corresponding real-world scenarios, regardless of domain semantics.

Thus, DeepMaze exposes reasoning prerequisites that must be satisfied before deploying models in resource-intensive applications. Its value lies not in replacing domain-specific evaluation, but in pre-screening models for fundamental reasoning competence—filtering out architectures prone to statistical shortcuts before costly specialized testing. This diagnostic capability becomes increasingly critical as models are deployed in safety-sensitive domains where reasoning failures carry significant consequences.

B Prompt Templates for MazeBench Tasks

B.1 Task 1: Global Observation Mode

Task 1 Prompt: Global Observation Mode

```
You are a maze expert and
need to navigate yourself to
the goal. Current position
A{current_position}, goal {goal}.
Coordinate format: All coordinates
use (y, x) format, where y is the
row number and x is the column
number.

You can:
Output path coordinates: e.g.,
(1,2),(1,3),(2,3)
Note: Please start outputting from
the next position's coordinate,
do not output coordinates that
duplicate the current position,
and the output coordinates must be
adjacent.
Hint: First confirm your current
position, then you can move to a
safe position first, and then plan
the next steps.

Maze layout:
{maze_ascii}
Symbol legend:
# : Wall
0 : Path
S : Start position (you have left)
G : Goal position
A : Current position
? : Unknown area

{memory_info}

Output protocol:
```

```
Reply with ONLY the path
coordinate sequence in the format
``(y1,x1),(y2,x2),...''
Do not include any explanations,
reasoning, or additional text.
```

B.2 Task 2: Partial Observation Mode

Task 2 Prompt: Partial Observation Mode

```
You are a maze expert and need
to navigate yourself to the goal.
You can only see the surrounding
{visibility} cells, and your vision
cannot pass through walls. You do
not know your current coordinates
or the goal coordinates.

You can move by specifying
directions:
- Direction: up, down, left, right
- Format: direction1,direction2,...
- Each direction moves one step

Constraint 1 -- Determine legal
actions first:
Before choosing moves, you
must internally determine which
directions are immediately blocked
by walls or out of bounds. You
must NEVER choose a direction that
is known to be blocked.

Constraint 2 -- Forbidden actions
from failures:
Any direction that resulted in a
wall collision in recent steps
must be treated as forbidden
unless no other legal direction
exists.

Constraint 3 -- Exploration
requirement:
You must explore new positions.
Repeatedly revisiting the same
locations will cause the task to
be judged as failed.

Visible maze layout (you are at
position A):
{maze_ascii}

Symbol legend:
# : Wall
0 : Path
S : Start position (you have left)
G : Goal position
A : Current position
? : Unknown area

{memory_info}

Output protocol:
Reply with ONLY the move
directions in the format
``direction1,direction2,...''
Do not include any explanations,
reasoning, or additional text.
```