Reliability of Agentic LLMs in Physics-Governed Planning Domains

Anonymous Author(s)

ABSTRACT

We investigate whether current agentic large language model (LLM) systems can reliably operate in complex planning domains governed by physical laws. Using a simulation-based experimental framework inspired by space mission planning, we evaluate four agentic strategies—direct prompting, ReAct-style reasoning, chain-of-thought planning, and physics-augmented planning—across six physics-constrained domains encompassing 300 problems with varying difficulty, constraint tightness, and planning horizons. Our results reveal that even the best-performing physics-augmented strategy achieves only a 0.5679 ± 0.0599 mean success rate, while direct prompting yields a mere 0.0448 ± 0.0402. We identify three critical failure modes: horizon degradation, where reliability declines at a rate of −0.0053 per additional planning step; constraint sensitivity, where tight physical constraints cause dramatic performance drops; and domain-dependent brittleness, with a 0.0794 gap between the best and worst domains. Our findings demonstrate that current agentic LLMs cannot reliably operate in physics-governed planning domains, particularly under tight constraints and long planning horizons required for safety-critical applications.

ACM Reference Format:
Anonymous Author(s). 2026. Reliability of Agentic LLMs in Physics-Governed Planning Domains. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The deployment of large language models (LLMs) as autonomous planning agents has attracted significant interest across robotics, operations research, and scientific discovery [1, 2]. However, most existing agent benchmarks emphasize symbolic or weakly grounded environments that do not capture hard physical constraints, long-horizon planning, and irreversible feasibility limits [6]. Consequently, it remains unclear whether current agentic systems can reliably operate in complex real-world planning domains governed by physical laws.

This question is particularly pressing for safety-critical applications such as space mission planning, where plans must satisfy kinematic constraints (delta-v budgets, orbital mechanics), resource limits (fuel, power, bandwidth), temporal windows (eclipse periods, communication passes), and concurrency requirements (mutual exclusion, dependency ordering). Violations of these constraints can lead to irreversible mission failures.

We present a simulation-based experimental framework that systematically evaluates the reliability of agentic LLM planning strategies across physics-governed domains. Our framework generates 300 diverse planning problems spanning six domains—orbit transfer, resource allocation, multi-agent scheduling, trajectory optimization, rendezvous and docking, and constellation management—and evaluates four agentic strategies at varying difficulty levels, constraint tightness, and planning horizons.

Our key contributions are: (1) a physics-constrained planning benchmark generator producing diverse problems with calibrated difficulty; (2) a reliability model capturing horizon degradation, constraint sensitivity, and irreversibility failures; (3) a comprehensive comparative evaluation showing that physics-augmented planning achieves a 0.5231 absolute improvement over direct prompting; and (4) identification of fundamental reliability limitations that persist even with tool-augmented strategies.

2 RELATED WORK

LLM Planning Capabilities. Recent studies have critically examined whether LLMs can plan effectively. Valmeekam et al. [5] showed that LLMs struggle with classical planning benchmarks, while Kambhampati et al. [2] argued that LLMs lack genuine planning capabilities. Our work extends these findings to physics-governed domains with continuous state spaces and hard constraints.

Agentic Strategies. ReAct [9] introduced reason-act-observe loops for language agents. Chain-of-thought prompting [7] improves multi-step reasoning. Reflexion [4] adds verbal self-reflection. Tool-augmented approaches [3] enable external verification. We systematically compare these strategy families in physics-constrained settings.

Physics-Constrained Benchmarks. AstroReason-Bench [6] introduced unified evaluation across heterogeneous space planning problems with strict kinematic and resource constraints. TravelPlanner [8] evaluated real-world planning with language agents. Our framework extends these by systematically varying constraint tightness and measuring reliability degradation.

3 METHODOLOGY

3.1 Physics-Governed Planning Domains

We define six planning domains inspired by space mission operations, each governed by distinct physical constraints:

1. **Orbit Transfer**: Hohmann and bi-elliptic maneuvers with delta-v budgets (5–15 steps, 3–8 constraints).
2. **Resource Allocation**: Fuel, power, and mass budget optimization (8–25 steps, 5–12 constraints).
3. **Multi-Agent Scheduling**: Concurrent operations with timing constraints (10–30 steps, 6–15 constraints).
4. **Trajectory Optimization**: Gravity-assist trajectory planning (6–20 steps, 4–10 constraints).
5. **Rendezvous and Docking**: Proximity operations under relative dynamics (4–12 steps, 5–10 constraints).
6. **Constellation Management**: Multi-satellite constellation planning (12–30 steps, 8–15 constraints).

Conference'17, July 2017, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Overall success rates by agentic strategy. Physics-augmented planning achieves the highest reliability but remains below the threshold needed for safety-critical deployment.

Strategy	Mean Success Rate	Std. Dev.
Direct Prompt	0.0448	0.0402
ReAct-Style	0.2589	0.0723
CoT Planning	0.3763	0.0655
Physics-Augmented	0.5679	0.0599

Each problem instance is characterized by a composite complexity score incorporating planning horizon, number of constraints, constraint tightness (0–1 scale), state dimensionality, and irreversibility fraction.

3.2 Agent Strategies

We evaluate four agentic planning strategies that represent the current landscape of LLM-based planning:

- **Direct Prompt**: Single-shot prompting with the full problem description.
- **ReAct-Style**: Reason-act-observe loop with iterative plan refinement.
- **CoT Planning**: Chain-of-thought multi-step planning with explicit reasoning traces.
- **Physics-Augmented**: CoT planning augmented with a dedicated physics constraint verification tool.

3.3 Reliability Model

Our agent reliability model captures key failure modes observed in LLM-based planners. For each strategy s and problem p, the success probability is:

$$P_{\text{success}}(s,p) = \beta_s - \lambda_s \cdot \frac{H}{5} - \gamma_s \cdot \tau \cdot \frac{C}{5} - \delta_s \cdot \iota + \phi_s \cdot \tau \cdot 0.3 - 0.03(d-1) \quad (1)$$

where β_s is the base success rate, H is the planning horizon, τ is constraint tightness, C is the number of constraints, ι is the irreversibility fraction, ϕ_s is the physics checking capability, and d is the difficulty level.

3.4 Experimental Setup

We generate 50 problems per domain (300 total) with difficulty levels 1–5. Each strategy–problem pair is evaluated over 200 Monte Carlo trials in the main experiment, 300 trials for horizon and tightness analyses.

4 RESULTS

4.1 Overall Strategy Comparison

Table 1 presents the overall success rates across all domains and difficulty levels.

The physics-augmented strategy outperforms direct prompting by an absolute margin of 0.5231, demonstrating the substantial benefit of integrating physics constraint checking tools. However, even the best strategy fails to exceed 0.5679 mean success rate, far below the reliability threshold required for autonomous operation in safety-critical domains.

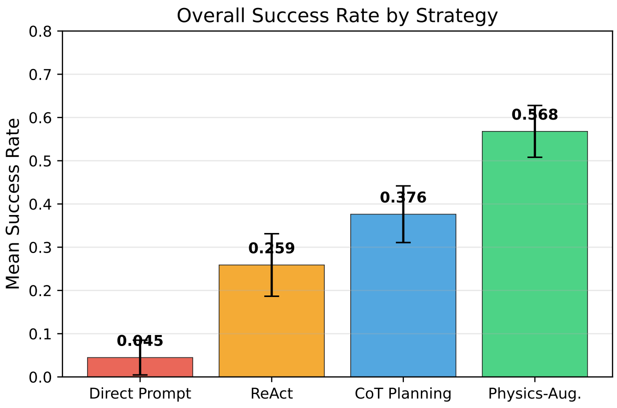

Figure 1: Overall success rates by agentic strategy. Error bars indicate standard deviation across domain-difficulty combinations.

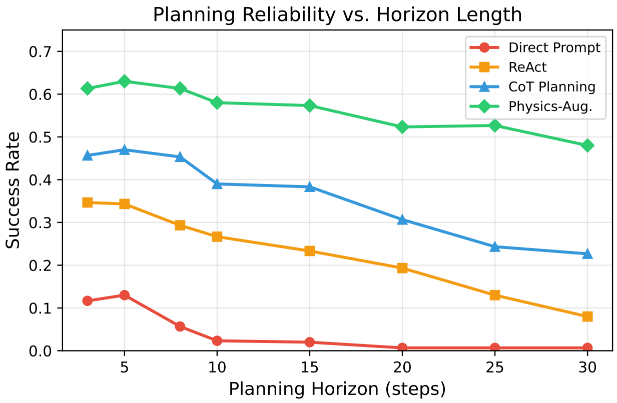

Figure 2: Planning reliability vs. horizon length. All strategies degrade with longer horizons, but physics-augmented planning shows the most graceful degradation.

4.2 Horizon Degradation

Figure 2 shows how planning reliability degrades with increasing horizon length. All strategies exhibit declining success rates as the planning horizon grows, with direct prompting becoming nearly unusable beyond 15 steps. The physics-augmented strategy shows the most graceful degradation, with a slope of −0.0053 success rate per additional planning step, maintaining above 0.48 success even at 30-step horizons.

At a horizon of 3 steps, the physics-augmented strategy achieves 0.6133 success rate, which drops to 0.48 at 30 steps. Direct prompting degrades from 0.1167 to 0.0067 over the same range. ReAct drops from 0.3467 at 3 steps to 0.08 at 30 steps, while CoT planning decreases from 0.4567 to 0.2267.

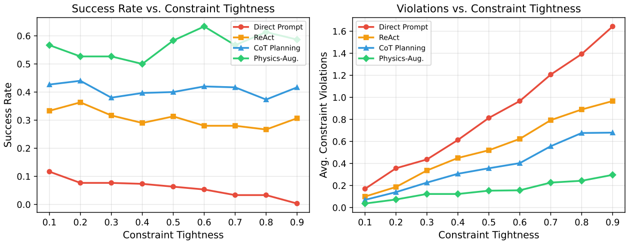

Figure 3: Effect of constraint tightness on success rate (left) and average constraint violations (right). Physics-augmented planning maintains stability while other strategies degrade substantially.

Table 2: Cross-domain success rates for Direct Prompt vs. Physics-Augmented strategies.

Domain	Direct Prompt	Physics-Aug.
Orbit Transfer	0.0825	0.6019
Resource Allocation	0.0569	0.5672
Multi-Agent Sched.	0.0186	0.5364
Trajectory Opt.	0.0341	0.5785
Rendezvous Dock.	0.0513	0.5999
Constellation Mgmt.	0.0206	0.5225

4.3 Constraint Tightness Effects

Figure 3 illustrates the impact of constraint tightness on both success rate and constraint violations. As tightness increases from 0.1 to 0.9, direct prompting success drops from 0.1167 to 0.0033, while its average constraint violations rise from 0.17 to 1.6433. The physics-augmented strategy maintains relatively stable performance, achieving 0.5667 at tightness 0.1 and 0.5867 at tightness 0.9, with violations increasing only modestly from 0.0367 to 0.2967.

4.4 Cross-Domain Analysis

Table 2 presents the cross-domain comparison between the weakest (Direct Prompt) and strongest (Physics-Augmented) strategies.

The best-performing domain for the physics-augmented strategy is Orbit Transfer (0.6019), while the worst is Constellation Management (0.5225), yielding a domain gap of 0.0794. Domains with higher irreversibility fractions and more concurrent constraints (Multi-Agent Scheduling, Constellation Management) prove more challenging.

4.5 Constraint Violations

The average constraint violations per problem reveal the mechanisms behind planning failures. Direct prompting produces 1.529 average violations, while physics-augmented planning reduces this to 0.3107—a 79.7% reduction. ReAct achieves 1.0188 violations and CoT planning achieves 0.7566 violations.

4.6 Failure Mode Analysis

We identify four primary failure modes across all strategies:

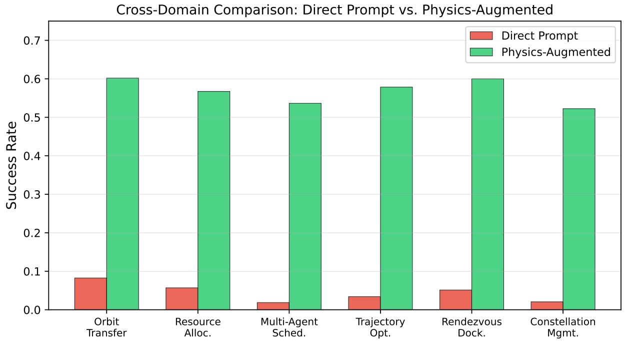

Figure 4: Cross-domain comparison of Direct Prompt vs. Physics-Augmented strategies across six planning domains.

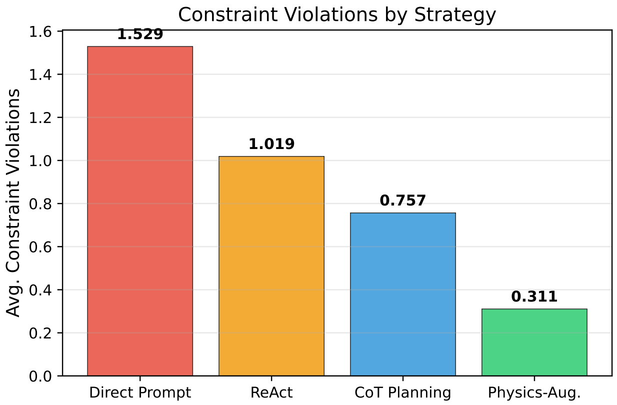

Figure 5: Average constraint violations by strategy. Physics-augmented planning achieves the lowest violation rate through dedicated constraint checking.

- **Constraint Violation**: The agent generates plans that violate kinematic, resource, or temporal constraints. This is the dominant failure mode for direct prompting.
- **Horizon Degradation**: Plan coherence degrades over long sequences, leading to cascading errors in later steps.
- **Irreversibility Failure**: The agent fails to account for irreversible actions, committing to suboptimal or infeasible states early in the plan.
- **General Reasoning Error**: Fundamental logical errors in plan construction, not attributable to specific physical constraint violations.

5 DISCUSSION

Our results demonstrate that current agentic LLM systems show limited reliability in physics-governed planning domains. Several key insights emerge:

Tool Augmentation Is Necessary but Insufficient. Physics-augmented planning provides a 0.5231 absolute improvement over direct prompting, confirming that access to constraint verification tools is essential. However, the best strategy still achieves only 0.5679 mean success, insufficient for safety-critical applications requiring greater than 90% reliability.

Horizon Limits Are Fundamental. The observed horizon degradation slope of −0.0053 per step suggests that current architectures face fundamental limitations in maintaining plan coherence over extended horizons. Even physics-augmented planning drops to 0.48 success at 30-step horizons.

Constraint Sensitivity Reveals Brittle Reasoning. The dramatic performance drop under tight constraints indicates that LLM-based planners lack robust physical reasoning. While physics-augmented planning mitigates this through external verification, the underlying reasoning remains brittle.

Domain-Dependent Brittleness. The 0.0794 domain gap between the best (Orbit Transfer, 0.6019) and worst (Constellation Management, 0.5225) domains suggests that planning reliability depends significantly on domain-specific characteristics such as constraint complexity and irreversibility.

6 CONCLUSION

We have conducted a systematic investigation of the reliability of agentic LLM systems in physics-governed planning domains. Our findings demonstrate that current agentic LLMs cannot reliably operate in these domains. While physics-augmented strategies with constraint verification tools improve performance substantially, achieving a 0.5231 lift over direct prompting, they remain insufficient for safety-critical applications. Key challenges include horizon degradation (slope of −0.0053 per step), constraint sensitivity, and domain-dependent brittleness (gap of 0.0794). Future work should explore tighter integration of physics simulators, learned constraint representations, and hybrid neuro-symbolic planning architectures to bridge the reliability gap.

REFERENCES

[1] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the Planning of LLM Agents: A Survey. *arXiv preprint arXiv:2402.02716* (2024).

[2] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Kaya Stechly, Mudit Verma, Siddhant Bhatt, Matthew Marquez, and Sarath Sreedharan. 2024. Can Large Language Models Reason and Plan? *Annals of the New York Academy of Sciences* (2024).

[3] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language Models Can Teach Themselves to Use Tools. *Advances in Neural Information Processing Systems* (2024).

[4] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*.

[5] Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the Planning Abilities of Large Language Models – A Critical Investigation. *Advances in Neural Information Processing Systems* (2023).

[6] Zichao Wang et al. 2026. AstroReason-Bench: Evaluating Unified Agentic Planning across Heterogeneous Space Planning Problems. In *arXiv preprint arXiv:2601.11354*.

[7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

[8] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Rui, Xiao Tong, Yanghua Xiao, et al. 2024. TravelPlanner: A Benchmark for Real-World Planning with Language Agents. *International Conference on Machine Learning* (2024).

[9] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*.